Becoming Good at AI for Good

Meghana Kshirsagar*
Microsoft AI for Good
USA

Caleb Robinson*
Microsoft AI for Good
USA

Siyu Yang*
Microsoft AI for Good
USA

Shahrzad Gholami
Microsoft AI for Good
USA

Ivan Klyuzhin
Microsoft AI for Good
USA

Sumit Mukherjee
Microsoft AI for Good
USA

Md Nasir
Microsoft AI for Good
USA

Anthony Ortiz
Microsoft AI for Good
USA

Felipe Oviedo
Microsoft AI for Good
USA

Darren Tanner
Microsoft AI for Good
USA

Anusua Trivedi
Microsoft AI for Good
USA

Yixi Xu
Microsoft AI for Good
USA

Ming Zhong
Microsoft AI for Good
USA

Bistra Dilkina
University of Southern California
USA

Rahul Dodhia
Microsoft AI for Good
USA

Juan M. Lavista Ferres
Microsoft AI for Good
USA

ABSTRACT

AI for good (AI4G) projects involve developing and applying artificial intelligence (AI) based solutions to further goals in areas such as sustainability, health, humanitarian aid, and social justice. Developing and deploying such solutions must be done in collaboration with partners who are experts in the domain in question and who already have experience in making progress towards such goals. Based on our experiences, we detail the different aspects of this type of collaboration broken down into four high-level categories: communication, data, modeling, and impact, and distill eleven takeaways to guide such projects in the future. We briefly describe two case studies to illustrate how some of these takeaways were applied in practice during our past collaborations.

CCS CONCEPTS

• **General and reference** → *Surveys and overviews*; • **Human-centered computing** → Collaborative and social computing; • **Social and professional topics** → Sustainability.

KEYWORDS

AI for good; collaboration; sustainability; case study

ACM Reference Format:
Meghana Kshirsagar, Caleb Robinson, Siyu Yang, Shahrzad Gholami, Ivan Klyuzhin, Sumit Mukherjee, Md Nasir, Anthony Ortiz, Felipe Oviedo, Darren Tanner, Anusua Trivedi, Yixi Xu, Ming Zhong, Bistra Dilkina, Rahul Dodhia, and Juan M. Lavista Ferres. 2021. Becoming Good at AI for Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462599

*Equal first author contribution.

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462599

1 INTRODUCTION

Advances in artificial intelligence (AI) and computing power have given rise to powerful AI tools ubiquitous in many people's personal and professional lives. These abilities are integrated into our phones and computers, and are driven mainly by businesses that have productized advances in AI at massive scales. Many of these tools are broadly available and provide some social benefit (e.g., search engines, navigation tools). However, the promise of AI to improve lives and protect vulnerable people and ecosystems has not yet reached its potential.

AI for Good (AI4G) is a movement within the larger field of AI that aims to develop and use AI methods to further progress towards goals in sustainability, health, humanitarian aid, and social justice, guided loosely by the UN Sustainable Development Goals (SDGs) and priorities within local communities. Excellent literature reviews on the topic are offered by [17, 54, 61, 70]. A key difference from commercial applications of AI is that those AI4G problems and successes are often not defined by market need, but rather by non-profits, social enterprises and governments seeking

to solve problems that have not found solutions in the private sector. For example, researchers in the field of computational sustainability [24] develop and apply methods to tackle problems such as wildlife conservation [19], bioacoustics [82], bird-migration tracking [59] and poverty detection [34]. In recent years, a number of pieces of criticism have been directed at the AI4G movement [10, 26, 40]. While these critiques raise important concerns such as the bias of models trained on limited data, shifting attention away from root causes of societal problems, and a paternalistic understanding of the affected community, the discussion largely centers on the difficulty of defining what is "good" in our societal context.

In this article we distill first-hand experiences from our research lab focused on AI4G projects spanning several application areas over two years. Cognizant of the complexity of problems in the AI4G domain and our expertise restricted to the technical side of AI (statistics, modeling, and engineering), we collaborate extensively with external *partner organizations* (PO) to define good outcomes for our projects, source and curate data, and realize real-world impact from our modeling solutions. These *AI4G projects* contribute to solving problems in two ways: we develop and apply AI techniques to accelerate previously manual tasks such as data processing to enable the PO to arrive at their solution faster, and we analyze and model collected data for additional insights. The collaborative and practical nature of such projects means that the *deliverables* are not just model weights, source code, and technical papers; they crucially involve working with these POs (whose technical capabilities and infrastructure vary greatly) to develop workable engineering solutions for deployment that respect resource constraints that POs face, as well as communicating and documenting our solutions – and their limitations – for the domain experts outside of computer science and engineering who use our models to impact society.

We highlight challenges that are more pronounced in AI4G projects compared to machine learning (ML) projects in the academic and corporate spheres, outline strategies we have learned for undertaking such projects, and reflect on difficulties we have faced measuring our impact. We break these down into four sections in the rest of the discussion: communication, data, modeling, and impact. Finally, we describe two case studies, i.e. *AI4G projects*, that exemplify these difficulties and how we approached them in a real-world setting.

2 COMMUNICATION

The relationship and interaction between data scientists and POs – who are the domain experts that define problems, curate data, and act on model outputs – is an important first topic. Domain experts sometimes have decades worth of experience working in a problem area. Communicating all of this accumulated knowledge to data scientists within a few days or weeks during project planning can be difficult, but data scientists must be willing and ready to incorporate this knowledge into their modeling approaches. While straightforward, accurate bi-directional communication at all stages of an AI4G project is essential for its success, we focus below on areas where data scientists may need to drive the conversation with the PO.

2.1 Setting realistic expectations from AI

It is often the case that POs have inflated expectations about the capabilities of modern AI-based techniques due to the hype surrounding the field[1] and its misrepresentation in the media [38]. In our experience, when initially proposing and scoping projects, some POs may believe that there are pre-existing AI tools that can be immediately adapted for a niche purpose with little to no training data (cf. [15]). However, our experience also shows that POs respond well to open, honest communication about AI's capabilities, or the need for (labeled) training data. Early conversations often involved directing POs' expectations away from a model that achieves all of their goals, which would require more training data than is currently available, to more targeted models for which appropriate and sufficient training data is available, and which will still bring them significantly closer to their goals. For certain use cases, an adequate and achievable approach is to use AI as a complementary tool to streamline and accelerate current workflows, rather than entirely supplanting them.

One such example we have worked on involved a PO who initially approached us about building a natural language processing (NLP) model to extract very nuanced author intent from decontextualized 1- or 2-sentence texts in a small corpus of unlabeled documents. After gaining a better understanding of the specific use case for the model, we worked with the PO to reframe the problem in terms of multi-label topic classification, and also worked with them to label the data. The result was a model that allowed the PO to incorporate an entirely new data source (text) into a broader initiative focused on quantifying human activities' impacts on environmental resources.

In addition to discussing the general limitations of AI-based methods with regards to what can be achieved, depending on the PO's domain and technical expertise, it may also be crucial to make them aware of more concrete issues encountered while building ML models. These include overfitting in small data regimes, model bias, generalization issues after deploying a model, adversarial attacks, data and model privacy concerns, limitations of interpretable models, etc. For example, lesion shape irregularity is one of the most critical features in the clinical diagnosis of melanoma [1]. On the other hand, recent studies have determined that convolutional neural networks are negatively-biased in capturing shape-related information from images [5, 21]. Clinical researchers who may wish to develop a convolutional model for melanoma detection are likely to be unaware of this finding, and effective communication of this knowledge may facilitate the development of alternative approaches. We believe that knowledge transfer should be a core component of an AI4G project.

In contrast to these positive cases, there may be circumstances in which POs need to be informed when their goals – even with proper reframing – may not be achievable with AI. For example, a model trained to detect fish species from underwater cameras may be highly accurate in identifying a few common species, but the researchers may like to detect very rare species if, for example, detection of the rare species is key to deciding whether some economic development opportunity is allowed. In these cases, model

[1]For example, a Gartner 2020 report on emerging technologies places AI at the peak in terms of inflated expectations [46].

creation may need to be postponed for time-consuming data collection and labeling. And in some cases it may be necessary to ask if the current state of machine learning is able to meaningfully help the PO in meeting its goals.

> **Takeaway *1.*** Educating POs about AI's limits and opportunities is a core part of an AI4G project. Potentially unrealistic expectations for AI can often be reframed into achievable goals that streamline the PO's workflows.

2.2 Project scoping and implementation

Collaborations between POs and technical experts on AI4G projects are typically short-term or phased, hence it is important to set priorities and expectations before beginning a project. For example, one project we have worked on involves counting herd animals (e.g. cattle) over large areas from satellite imagery to monitor the effectiveness of conservation policies. This is an expensive task to perform manually, but can potentially be automated with computer vision models. Instance segmentation of the animals is a straightforward approach, however it is risky due to the lack of labeled data or low spatial resolution of the imagery. Less straightforward modeling approaches, such as coarser animal density estimation models, can satisfy the same project goals, but may take more effort to define upfront. Such approaches may also be informed by domain knowledge – herd animals move in groups, resulting in dense crowds more easily identifiable in imagery than individual animals. The back-and-forth communication with the PO to understand the goal and explore *appropriate* modeling solutions is a crucial part of the project life cycle.

We summarize questions to explore when scoping a project with a PO as a guideline for future projects:

(1) Setting accurate goals: is the project oriented towards the prediction or estimation of quantities, or is it towards the visualization and representation of data? If the project is oriented towards prediction/estimation, then *what* will be done with the output from the model, *who* will use it (say, for decision-making), and *where* will it be deployed? If the project is oriented towards visualization/representation or to interpretability/inference, *what* format is expected, *who* will be consuming the end result, and – again – *where* will it be deployed?
(2) What are the resource constraints of the PO? The compute and storage resources available and the type of device where the solution will be deployed (cloud-based, battery-powered devices) should be taken into account to determine what modeling solutions are possible.
(3) What is the technical expertise available at the PO? This will determine the amount of support available for maintaining a deliverable beyond the duration of the project.
(4) Does the problem require developing novel machine learning techniques? This will inform the project timeline and the risk. It could present an opportunity to identify larger research challenges and bring them into the ML community at large.
(5) To what degree is data/model privacy a concern? The data available to the PO can often contain personally identifiable information about individuals. This would not only make the dataset not releasable for the sake of reproducibility, but also limit the public release of models since they may be subject to privacy attacks [62].
(6) Is precise information on the data well documented? The PO should convey the assumptions and the procedures underlying data collection. If the data has already been processed, the processing steps should be communicated because certain pre-processing steps can introduce bias in the data [20].
(7) How could the PO's domain knowledge be incorporated in modeling? During the project development stage, it may be necessary to not rely entirely on data-driven techniques (for instance, for feature learning), but also utilize PO's knowledge of domain-relevant metrics and features that have been identified as significant in their field. For example, in radiology, dozens of hand-crafted radiomic features have been previously found to be significant predictors of cancer survival and response to therapy [4].

> **Takeaway *2.*** To ensure we develop solutions that are practically useful, project scoping needs to be an ongoing dialogue with the PO.

3 DATA

Most AI4G projects start with the PO sharing a dataset they have previously collected and labeled, or pointing out public datasets that are relevant to the problem. While specific funding opportunities for creating datasets for ML applications and novel methods to take advantage of weakly labeled public data [73, 83] or generated data [8] exist, exploiting these is often not possible. This is in contrast to enterprise applications where ML teams are also responsible for collecting training data and have budget allocated for curating datasets with specific ML tasks in mind. In this section we reflect on recurring challenges related to data availability and quality in the context of AI4G projects; for an in-depth discussion of data issues present in high-stakes AI applications, refer to [57].

3.1 Adapting to previously collected datasets

Limitations in the data collection process often influence the degree to which an AI4G project can be successful, possibly more so than they affect the outcomes of commercial or academic AI projects, which have more control over data collection.

First, there is a discrepancy between the purpose of data collection by POs that have the goal of solving an application problem and data collection by groups that have the goal of creating a generalizable model. The POs need not necessarily care about the metadata associated with data points that they label for furthering their goals, while such metadata may be important for quantifying how a model trained on such data will generalize. Put differently, if data collection is conducted by a PO with a focus on the content, rather than technical specifications of the data, then this can cause problems in post-hoc modeling steps.

Extending the example of counting herd animals from the last section, the PO might label herd animals from satellite imagery at a variety of spatial resolutions. The imagery could be at a spatial resolution of 0.1, 0.3, or 0.5 m/pixel; as long as the PO can count the types of animals of interest over a given area and date, they

can meet their goals. On the other hand, a machine learning model that has been trained to identify herd animals from only 0.1 m/pixel imagery will likely not generalize to 0.5 m/pixel imagery as objects will be 5 times smaller in each dimension. The metadata associated with the labeled data is necessary to inform modeling decisions (e.g. augmenting the scale of satellite data and labels during model training will allow the model to generalize over a range of scales). This is in contrast to settings where data is collected specifically for the purpose of building effective models, where metadata would be considered explicitly at the data collection stage.

In addition to the completeness of metadata, quality and consistency of data collection, which do not affect experts' ability to discern the content but which pose additional generalization challenges to machine learning models, are a frequent issue. For example, in a study applying speech feature modeling to analyze mental health issues, the recording of conversations between military personnel with suicide risk and their therapists is used to predict their emotional bond [43]. Here, the consistent use of dedicated microphones for the two speakers in a controlled environment was found to improve modeling outcomes, however this aspect of data collection would not affect the manual analysis of the recordings.

Second, AI4G projects often involve sensitive data necessitating the adherence to strict ethical and legal restrictions that make using such data more difficult. For example, many agencies would like to train computer vision models to detect images of Child Sexual Abuse Material (CSAM), but databases such as the Child Abuse Image Database (CAID) maintained by the UK Government are not accessible to other organizations [11]. Other examples include applications with healthcare data – models that identify pulmonary features from chest x-ray imagery need to be trained with labeled chest x-rays, a data source that initially requires pairing patient records with their chest x-rays. These sensitive applications require specialized privacy-preserving modeling methods and a layer of complexity that other applications do not entail. POs may choose to share synthetic data generated using generative models instead of real data but these approaches also may be prone to membership inference attacks [30]. Recent work on privacy preserving ML has developed both defenses [41, 77] and ways to measure the privacy risks from releasing models [33, 37].

Third, the amount and quality of data available in some AI4G projects will be limited. This is not specific to AI4G projects, however is worth mentioning due to the frequency with which such projects come up. For example, projects that involve detecting poultry barns, solar farms, or other relatively uncommon features from satellite imagery require having a labeled dataset of such features. However, creating labeled data in these applications is expensive as it requires annotators to first find examples of the objects in question over large landscapes before labeling them appropriately. In our experience, these type of satellite image annotations will not be in a format that is immediately useable (e.g. point labels for a segmentation problem), or will be a biased sample (e.g. many labels from a specific area rather than a sample of labels from a broad area).

Finally, open datasets often have significant data curation issues. For example, a Kaggle dataset for predicting the outcome of pregnancies in India has been shown to be missing key data from the original survey, resulting in misleading predictions [67]. POs that want to help reduce infant mortality can be misled by such datasets or the promise of effective models from open competitions that use such data. Another issue is the lack of query infrastructure around public datasets which creates significant friction in projects that might use such data. A positive example is Google Earth Engine [25], which allows researchers to quickly query across its collection of public satellite imagery, visualize sample patches, and assess if the data is of sufficient quantity and resolution for the intended analysis.

> **Takeaway *3.*** Datasets in AI4G projects may not be immediately useful for creating models. When creating models with such data, it is important to understand the associated metadata, collection process, and any security or privacy concerns.

3.2 Dealing with subjective data annotation

The variables of interest in several socially important domains involving human perception and decision-making are ambiguous and ill-defined. Different annotators might interpret the definition of labels differently, leading to inconsistent and noisy labels. For example, the colloquial meaning of "depression" might be different from its meaning in a clinical context [79]. The data annotation process for an AI system aimed at diagnosing clinical depression should be cognizant of the difference. Creating a taxonomy of labels could minimize the annotator's subjectivity. This also requires transparency in the interpretation of the labels and clearly communicating its limitations during different stages of the project life cycle, including annotation, modelling and deployment to minimize the semantic ambiguity of labels. An in-depth discussion on ambiguous labels in the context of computational social sciences can be found in [14]. When subjectivity of the labels are inherent due to the variability in human perception, it should be a standard practice to employ multiple annotators for each example and judge the feasibility of the task by evaluating inter-annotator agreement. Several methods have been proposed to obtain estimates of the true ground truth labels from the noisy/subjective labels collected from multiple annotators [42, 48].

> **Takeaway *4.*** In several socially important domains, labels suffer from subjective annotation. Such situation should be identified upfront to avoid introducing inconsistencies in the modeling pipeline.

3.3 Creating training and test sets with the application scenario in mind

Poor choices of training, validation and test set splits can result in an estimated model performance that does not reflect actual performance when deployed (for other lessons learnt in evaluating model performance, see section 4.3). This is especially relevant in humanitarian aid and conservation applications where models are expected to generalize well spatially and/or temporally.

For instance, with marine mammal sound detection [81], while generating train/test splits, consideration should be given to different types of underwater and anthropogenic noises such as those from commercial ship generators, mining, aircraft, and seasonal

effects on ocean waves. As another example, the xBD dataset associated with the xView Challenge [28] is a large-scale public dataset designed to enable building damage assessment for humanitarian assistance and disaster recovery. It consists of data from 19 disasters from around the world between 2011 and 2019. However, scenes from the 19 disasters are present in all of the official train, test and hold-out splits, whereas it would be more useful to report performance on unseen locations as the next disaster will likely strike elsewhere. The same is true for the SpaceNet Challenge Series for building footprint extraction, where the default splits created by the challenge's utilities contain overlapping locations [69]. Subsequent studies have shown that performance drops drastically when applying a trained building damage classifier to an unseen location, even within the same region [29, 68]. As another example requiring spatial generalization, until very recently, all studies of animal species classification on camera trap images were split across sequences of images but not across locations. This results in precision and recall metrics of greater than 90% [44]. In reality, performance is much worse if we split the data by camera location, and even worse if we split by ecosystem [9, 58].

Many past studies in wildfire risk prediction, another problem important to both disaster relief and environmental conservation, assume certain random variables to be independent and identically distributed in the evaluation phase [12, 53, 56]. However, natural hazards like wildfires are stochastic events with spatio-temporal dimensions, and evaluating models of such events based on randomized training and test splits leads to information leakage, misleading organizations who operationalize such models [22].

In applying machine learning to medical imaging, Zech et al. [80] found that a dataset of chest x-rays curated for screening pneumonia cases could be used to train a model to accurately predict which hospital system the x-ray comes from, indicating that the pneumonia detection model developed from the dataset could have been aided by features not related to the medical condition. The training and test splits should be chosen to measure how well the model will work for unseen x-ray machines.

> **Takeaway *5.*** Carefully consider how to split the data into training and test sets so that the model's ability to generalize to unseen instances of input is measured.

4 MODELING

The models used in AI4G contexts usually involve a different set of requirements and constraints compared to general AI applications. First, AI4G models are developed in applied ML contexts. Most of the models are developed with domain-specific motivations and limitations in mind. In consequence, models developed for mainstream ML fields, such as NLP or computer vision, require cautious adaptation and deployment to the specific domain. Furthermore, the process of model development is motivated first by application requirements instead of pure novelty or state-of-the-art performance.

4.1 Incorporating domain expertise

Domain expertise from the PO can help in model development as POs often have decades of experience and accumulated knowledge in defining and solving related problems. Specifically, domain expertise is useful in: i) determining adequate features and data representations, ii) enforcing inductive bias and regularization in models, iii) choosing simplified parameterizations, iv) interpreting the learned models and outputs.

Domain expertise is invaluable for collecting features that are relevant to a problem. This is especially relevant in AI4G problems where there are few samples compared to the number of features or where informative features are mixed with noisy features. For instance, to predict the late effects of chemotherapy on cancer patients, we found that including all chemotherapy drugs gave us a higher predictive performance as compared to prior work. However, clinical researchers know that only certain drugs have been clinically linked to certain late effects in other prior analyses. This suggests that the additional confounding features were probably spurious contributors to predictive performance and hence should not be included in the model.

In addition to helping determine data representation, domain expertise can help improve performance by embedding specific knowledge, often in the form of inductive bias or regularization, in model design. For example, finding promising solar cell technologies is often a difficult and a time-consuming task. A solar cell consists of a stack of various semiconductor materials, where each layer performs a certain electrical and optical function and fabrication parameters are optimized for maximizing solar energy captured. A machine learning model can avoid the need for resource-intensive physical experiments and accelerate the parameter optimization step. Combining a supervised machine learning model with a physical model of solar cell operation calibrated by an expert allowed for a model regularization method based on physical principles [49] and an order of magnitude reduction in the time and resources required to create a solar cell [45, 49].

Further, AI4G projects often involve collaborations spanning multiple countries and POs in a different country are likely to face unique challenges that only local experts would be aware of. For example, work on the anti-poaching PAWS project [18] uses the local knowledge of park rangers to constrain predicted search patterns to areas that can be feasibly visited.

If an interpretable model is used, then domain experts may be able to use aggregated model predictions to draw larger conclusions about a problem. For example, in our collaboration studying the food security in low-resource communities in Malawi based on survey panel data of households, domain experts were able to use the outputs of interpretable models to recognize spatial and seasonal patterns associated with the food security status of the communities and villages. These community-level insights can help local governments manage their resources more efficiently across the communities and over time.

> **Takeaway *6.*** Endeavor to incorporate the PO's domain expertise in model development when possible through methods such as feature selection and engineering, model choice, and model regularization.

4.2 Model development with resource constraints

Since deployed models are maintained by the PO who often has less resources than enterprises focused on mainstream ML applications, resource constraints once the model is operationalized limit the choice of models. In addition to the financial cost of running sophisticated models on potentially large datasets, deploying to remote regions in battery-powered devices and carbon emission related environmental cost are also important considerations.

For example, Robinson et al. [51] trained a fully convolutional neural network (CNN) on over 55 terabytes of aerial imagery to create a high-resolution land cover map over the United States. Differences in seconds of running time per batch translate to hundreds of dollars in the cost of the final computation. Here, a larger, state-of-the-art model would incur a ~270% increase in the cost of the final computation for a fractional increase in performance metrics such as accuracy and intersection-over-union, and so a trade-off in favor of lowering the cost was made.

In wildlife conservation [74] and accessibility applications [75], models need to be deployed to edge or mobile devices of varying capacity. For instance, the Seeing AI[2] mobile app, which helps people with vision impairment or low vision to better understand their surroundings, uses deep learning architectures specifically designed for low resource settings.

State-of-the-art deep learning models have large carbon footprints from training and operation [63], which is a concern for AI4G projects in particular. Applications such as the one described in Lacoste et al. [35] allow the model developer to choose a data center location powered to a large extent by renewable energy sources and a cloud provider who offsets the remaining emissions.

> **Takeaway *7.*** Carefully consider a project's constraints during deployment in advance before settling on a modeling approach.

4.3 Evaluation and metrics

Model validation is a crucial part of any AI project. In AI4G projects, validation metrics will not only need to measure how well the model is performing in standard ways (e.g. accuracy, AUC-ROC, intersection over union), but how well the model is performing with respect to any domain-specific requirements.

For example, the common part of commuters (CPC) [36] is a domain-specific metric used in measuring how well predicted commuting flows align with ground truth data. This metric jointly considers all commuting flows together, as opposed to a common ML metric such as mean squared error that only considers pairwise errors between a single origin and destination. Reporting CPC reveals more about the overall structure of a predicted set of flows and is thus important to report in AI4G projects that consider commuter or migration flows [50].

Another example comes from cancer imaging, where specialized extensions of the receiver operating characteristic (ROC) analysis are used to evaluate the lesion detection performance by radiology readers: localization ROC (LROC) quantifies not only the correct binary diagnosis, but also takes into account the accuracy of lesion localization within an image. The free-response operating characteristic (FROC) extends the notion of LROC to the multiple-lesion detection task [6].

These and other domain-specific metrics can potentially be included in modeling as well as in evaluation. For example, if the domain-specific metric is differentiable with respect to the predicted quantity and computed on a per-sample basis, then it can be used in combination, or in place of, common loss functions when training models with gradient descent based methods. In the medical-imagery domain, such loss functions have been used to better capture domain-specific problem characteristics [65].

In other cases, a domain-specific metric is not necessary, however domain experts will care little about commonly reported ML metrics. For example, mean average precision (mAP) involves averaging the precision of a model at all possible recall values. This average will include, for example, the precision of the model at 1% recall which is not informative as such performance would never be acceptable. Precision@k, where k is the lowest tolerable recall, is a more appropriate metric. As we discuss in Section 2, arriving at this metric involves extended dialogue with domain experts.

Finally, the data collection process by the PO can be imperfect, therefore model evaluation based solely on such datasets might be insufficient. For example, in the case of the anti-poaching PAWS project, the dataset is collected by limited park rangers via foot patrolling over a vast area. As such, many regions in the protected areas are not thoroughly covered by the rangers every month and the dataset does not perfectly represent the area under study. Building a long-term collaboration with POs to deploy machine learning solutions for pilot tests before a full commitment can bring important insights about the performance of the trained model in the wild [23, 78].

> **Takeaway *8.*** Check if domain-specific metrics can be incorporated during training and validation of models and determine which ML metrics are relevant to solving the problem at hand.

4.4 Humans in the loop

In the industry, continued data collection and user-supplied labels allow models to be improved over time, the so-called "data flywheel" effect [16, 66]. In a similar manner, many scientific fields have come up with labeled datasets and models [13, 64, 71] that are continuously updated as labeling techniques (such as physical simulations or data acquisition methods) are improved. A common industry practice for improving model performance is to iterate on improving *datasets* instead of iterating on improving *models* [31]. In both cases, having humans continually in the loop – whether by labeling or tuning model behavior based on feedback from a deployed system – provide large benefits to the overall project outcomes.

In general, the one-off nature of AI4G projects preclude this common way of improving model performance. Accumulating expert-annotated labels, even those created with efficiency gains enabled by the first version of the model, lies outside of formal infrastructures and therefore model re-training is not done as often.

[2] https://www.microsoft.com/en-us/ai/seeing-ai

At the same time, most of the models used in AI4G projects do not enable complete automation. For example, the output of a medical diagnosis model will be interpreted by health professionals, and final decision may be made based on the patient's clinical history and the presence of secondary signs and symptoms that are not captured by the model [39]. Here, a human *must* be in the modeling loop evaluating every output of a model. More broadly, the output of *all* AI4G models will be inspected by domain experts in the PO and their feedback will constitute a form of weak supervision that must be included in the modeling process in order to produce a suitable deliverable. For example, POs that rely on highly accurate land cover data will often need to make manual corrections to modeled outputs, and, as such, will have a difficult time using land cover predictions with artifacts such as rounded corners on building that are not easily correctable in GIS software. This type of feedback is only apparent after one iteration of modelling and inspecting the results with the PO. Glacier monitoring is another scenario where incorporating humans in the loop have been shown valuable. Baraka et. al proposed a glacier mapping tool that uses semantic segmentation predictions as a starting point and allows domain experts for easy adjustments of those predictions for a faster glacier mapping system [7].

Thus, achieving a balance between having human feedback included in the modeling process and staying within-scope is a crucial part of finishing such AI4G projects. We have found active learning pipelines to be beneficial towards this end. With an active learning pipeline, participants at the PO can be engaged directly during the modeling process and will allow their feedback (in the form of labels) to be directly incorporated in the final deliverable.

We note that the active learning loop can also incorporate humans more tightly. For example, when humans are further allowed to choose the locations to label (versus being presented locations), and can observe the effect labeling those locations has on model output after a retraining period, they can more efficiently train land cover models [52]. Finally, as active learning methods more tightly couple dataset collection with model training, they show promise in reducing the total amount of manual effort required to produce a final product. For instance, active learning training of camera trap species identification models has been found to match state-of-the-art accuracy with orders of magnitude fewer annotated training samples [55].

> **Takeaway *9.*** AI4G projects require humans in the loop to some extent. Active learning pipelines can enable POs to engage with the modeling process directly during a project.

5 IMPACT

Lacking in the usual business indicators such as revenue and user engagement, one of the most difficult aspects of an AI4G project is measuring the degree to which it is successful, and weighing the success by the potential impact in advancing a PO's mission. In turn, an AI4G project's potential impact will not be realized without the PO or the broader communities adopting the technology. Indeed, there are many more press releases, blog posts, and promising published results than functioning AI systems actively "doing good" in the world. In this section, we attempt to understand why that is by exploring the three related issues of deployment, adoption and impact.

5.1 Uphill path to deployment and adoption

Unlike research developing novel techniques or theoretical understanding, AI4G projects necessitate the deployment of any developed models, and separately and often more difficult, the adoption of such technologies by the PO and related communities, for the effort to be meaningful [72]. This "last mile" problem can be especially challenging in AI4G projects since engineering is often not a focus for a PO or a research group.

In our experience, deployment entails three scenarios:

i) a one-time scoring of relevant input data to produce derived data for the PO's downstream analysis and publication,
ii) a real-time API exposing the model for applications such as anti-poaching and invasive species monitoring,
iii) a batch processing mechanism triggered automatically or by the user to process a large quantity of raw data for recurring analysis, such as while processing conflict videos from a region for weapon detection [2].

The first scenario requires the least engineering effort beyond model development, but may not result in lasting impact. Real-time APIs have recurring cost, in addition to requiring upkeep and integration with the client application consuming model output. Batch processing could take advantage of discount cloud compute at low-traffic times and parallelize model scoring. It is often necessary to guide the PO in understanding whether they require real-time always-on model deployment or if offline batch processing is sufficient. There is also a lot of enthusiasm for deploying ML models on edge devices so that the PO can avoid uploading the raw data for processing in low-connectivity regions. In this case, it is important to communicate any trade-offs compressing the model through techniques such as quantization may have on performance [60, 76]. To truly enable productivity gains from using AI tools and cloud infrastructure, the POs often need a much larger piece of software to orchestrate scoring raw data using the model and interact with the model outputs, of which Wildlife Insights is a notable example for accelerating wildlife surveys using computer vision models [3].

Adoption is the harder problem because, in many ways, it is outside of the control of the technical team. Identifying how ML metrics translate into time saved in the PO's workflows is paramount. Taking input from human-computer interaction experts may be helpful at this stage, as is thinking about how to integrate model output with the software used in downstream analysis. For example, being able to preview model outputs above a certain confidence threshold can help the domain expert to filter out input files that do not contain any subjects of interest; pre-populating the label field with the most common class may save many keystrokes during manual review [27]. We have also found that open-sourcing the model development code builds trust in the model, and the code repository with discussion boards can act as a hub for the community involved in transfusing AI into their domain.

Takeaway *10.* Maintaining deployed models requires long-term engineering resource commitments. Focusing on time saved instead of pure ML metrics helps organizations adopt the technology.

5.2 Measuring impact

Typical machine learning projects measure success in terms of model evaluation metrics (e.g. F1-score, ROC-AUC, etc.) (also see discussion in the last section) or business key performance indicator (KPIs) (e.g. click-through rate, daily active users, etc.). However, in the life cycle of AI4G projects, model evaluation metrics serve more as a basis of discussions with POs about a model's capabilities and limitations; the KPIs important to the POs are outcomes that can be several steps removed from the model outputs. It is important to learn about the PO's KPIs in the scoping phase of the project to inform the approach.

A challenge in creating *lasting* impact comes from the lack of a business model for these AI4G endeavors. We are finding ways to step out of a funding mindset and grow the technical capabilities of the PO so that they could be self-sufficient in subsequent data collection and re-training efforts. This is another place where two-way communication is important (see Section 2): the technical team often does not get to see the impact of their work in the field. Maintaining a relationship with the PO after the technical portion of the project is complete to get updates on how their workflows have been impacted is important in maintaining a long-term collaboration and acts as part of a larger feed-back loop for solving the application problem.

In our engagements with POs, we do not concern ourselves with defining what is a positive impact against the final problem the PO aims to tackle. We rely on the domain experts at the PO to determine what the intended eventual impact is and to what extent a project furthers the PO's mission. More proximal to data modeling, in collaborating with POs on AI4G projects we have found two ways for AI techniques to realize impact: finding structure and insights from large datasets, and making domain experts' workflows more efficient so that they may scale out their work. Therefore, it may require working with the PO to find ways to track the immediate impact of an AI4G model on their data analysis or workflow efficiency, in addition to impacts on the PO's end mission.

Takeaway *11.* Domain experts within POs should define mission-related impacts. When quantification of direct model impact is needed, work with the PO to identify opportunities to quantify both immediate (workflow or analysis enhancement) and farther-removed (mission-related) impacts of the AI4G project.

6 CASE STUDIES

6.1 NLP to map Syrian conflict

Problem:
The Carter Center (TCC) has been working on supporting a political solution to the wars in Syria[3]. Since 2012, TCC has initiated a conflict mapping project that analyzes an unprecedented volume of citizen-generated information about the conflict. Every week, TCC compiles a report using the information it receives from the Armed Conflict Location and Event Data (ACLED) Project[4] [47], which curates news stories and articles recording incidents related to the war in Syria. This report is read by various committees in the UN, foreign ministries and NGOs. Given the weekly timeline (Takeaway 7), collating the incoming data into a structure suitable for their analysis manually has been difficult given the increasing volume of reports. Automating this curation process would reduce the thousands of hours of work needed by professional analysts.
Solution:
Our work automated this process by classifying the information into several categories such as shelling, artillery fire, and aerial bombardment. We helped TCC build a high-precision, neural network-based natural language processing (NLP) model that reclassifies the input conflict events at the granularity desired by TCC (Takeaway 6). This improvement in data processing allowed TCC employees to then focus on subsequent analysis of the conflict events (Takeaway 9).

6.2 Mapping solar farms across India

Problem:
At the end of 2020, India was only 2% away from the target of 40% installed non-fossil fuel electricity capacity, one of its Paris Climate Agreement targets [32]. While it is encouraging to see renewable energy production systems, such as solar farms, being rapidly built, it is also important to locate such installations in a way that avoids encroaching on the habitats of endangered species and other ecological reserves. An international conservation NGO has been working with states in India to create a tool for identifying areas where solar and wind developments are less likely to cause ecological harm. However, information on where solar installations are located is only available for two states, and so we worked together with the NGO to use satellite imagery to try identify solar installations across all of India.
Solution:
Finding solar farms from satellite imagery is straight-forward to formulate as a semantic segmentation task, but the labels that were available for the project were both few and not in the format needed for this ML task: only 72 point labels of locations of solar farms in two states were available (Takeaway 3). To overcome this limitation, we first pre-trained a convolutional neural network to cluster pixels in the input satellite imagery by color (i.e. in an unsupervised manner). We then used an interactive training application to quickly fine-tune the network to segment the classes of interest and used this fine-tuned model to obtain weak segmentation labels for the entirety of the study area. These weak labels make it possible to train a supervised semantic segmentation network that was capable of accurately detecting solar farms. Solar farms found by this supervised model were validated by analysts at the NGO. In total we were able to find and validate 1422 solar installations across India. The human-in-the-loop process we used was a crucial component in both training and evaluation, enabling ML models yield reliable results given the small amounts of labels available initially (Takeaway 9). Given the large area of interest, we must also rely

[3] https://www.cartercenter.org/countries/syria.html

[4] https://www.acleddata.com/data/

on free satellite imagery, which is lower in resolution than commercial imagery; we accept this constraint to ensure our solution remains practically useful in the long term as the NGO updates the map each year (Takeaway 10). To reduce the number of false positive identifications, we incorporated OpenStreetMap data to remove areas of roads, snow and water bodies, a post-processing step informed by expertise in geospatial analysis (Takeaway 6).

7 CONCLUSION

Our work presents a broad overview of the considerations necessary while working on AI4G problems and the challenges encountered therein. We observe that the most useful AI4G projects result from working closely with specific stakeholders and understanding their operations and needs; attention to the particular characteristics of the problem while developing ML models, metrics and evaluation; a deep understanding of ethics and fairness concerns; a commitment to sound scientific and engineering practices and a transfer of technology that empowers the beneficiaries to understand and learn from the solution, and hopefully, adapt it with their changing needs. To support our observations we present several examples from our own experiences and relevant literature and summarize the learned lessons in takeaways. We hope that our endeavor helps researchers who are passionate about social good causes by bridging the gap between ML methodologies and their potential for relevant impact. However, we note that there are many problems and questions still outstanding, and that we are continually learning and growing our own repertoire in tackling challenging issues and working with POs. Becoming good at AI4G is a process that we are actively engaged in, and we hope others join us and learn with us.

ACKNOWLEDGMENTS

We would like to thank Karthikeyan Ramamurthy for helpful feedback on an initial draft. We would also like to thank the many partner organizations who have collaborated with us over the past several years.

REFERENCES

[1] Naheed R Abbasi, Helen M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Iman Osman, Alfred W Kopf, and David Polsky. 2004. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *JAMA* 292, 22 (2004), 2771–2776.

[2] Raja Abdulrahim. 2021. AI Emerges as Crucial Tool for Groups Seeking Justice for Syria War Crimes. *The Wall Street Journal* (2021). https://www.wsj.com/articles/ai-emerges-as-crucial-tool-for-groups-seeking-justice-for-syria-war-crimes-11613228401

[3] Jorge A Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y Zhao, Walter Jetz, et al. 2020. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation* 47, 1 (2020), 1–6.

[4] Michele Avanzo, Joseph Stancanello, and Issam El Naqa. 2017. Beyond imaging: the promise of radiomics. *Physica Medica* 38 (2017), 122–139.

[5] N. Baker, H. Lu, Gennady Erlikhman, and P. Kellman. 2020. Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research* 172 (2020), 46–61.

[6] Andriy I Bandos, Howard E Rockette, and David Gur. 2013. Subject-centered free-response ROC (FROC) analysis. *Medical physics* 40, 5 (2013), 051706.

[7] Shimaa Baraka, Benjamin Akera, Bibek Aryal, Tenzing Sherpa, Finu Shresta, Anthony Ortiz, Kris Sankaran, Juan Lavista Ferres, Mir Matin, and Yoshua Bengio. 2020. Machine Learning for Glacier Monitoring in the Hindu Kush Himalaya. *arXiv preprint arXiv:2012.05013* (2020).

[8] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. 2020. Synthetic examples improve generalization for rare classes. In *The IEEE Winter Conference on Applications of Computer Vision*. 863–873.

[9] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in Terra Incognita. In *Proceedings of the European Conference on Computer Vision*. Munich, Germany.

[10] Bettina Berendt. 2019. AI for the Common Good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics* 10, 1 (2019), 44–65.

[11] Roderic Broadhurst. 2020. Child sex abuse images and exploitation materials. In *The Human Factor of Cybercrime*, Rutger Leukfeldt and Thomas J. Holt (Eds.). Routledge, 310–336.

[12] Mauro Castelli, Leonardo Vanneschi, and Aleš Popovič. 2015. Predicting burned areas of forest fires: an artificial intelligence approach. *Fire ecology* 11, 1 (2015), 106–118.

[13] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. 2020. The Open Catalyst 2020 (OC20) Dataset and Community Challenges. *arXiv preprint arXiv:2010.09990* (2020).

[14] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–20.

[15] Michael Chui, James Manyika, and Mehdi Miremadi. 2018. What AI can and can't do (yet) for your business. *McKinsey Quarterly* 1 (2018), 97–108.

[16] Jim Collins. 2019. *Turning the flywheel: a monograph to accompany good to great.* Random House.

[17] Josh Cowls, Thomas King, Mariarosaria Taddeo, and Luciano Floridi. 2019. Designing AI for social good: Seven essential factors. *SSRN 3388669* (2019).

[18] Fei Fang, Thanh Hong Nguyen, Rob Pickles, Wai Y Lam, Gopalasamy R Clements, Bo An, Amandeep Singh, Milind Tambe, Andrew Lemieux, et al. 2016. Deploying PAWS: Field Optimization of the Protection Assistant for Wildlife Security.. In *AAAI*, Vol. 16. 3966–3973.

[19] Fei Fang, Milind Tambe, Bistra Dilkina, and Andrew J Plumptre. 2019. *Artificial intelligence and conservation.* Cambridge University Press.

[20] Salvador García, Julián Luengo, and Francisco Herrera. 2015. *Data preprocessing in data mining.* Springer.

[21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

[22] Shahrzad Gholami, Narendran Kodandapani, Jane Wang, and Juan M. Lavista Ferres. 2021. Where there's Smoke, there's Fire: Wildfire Risk Predictive Modeling via Historical Climate Data. In *Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*.

[23] Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, Andrew J Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, Joshua Mabonga, et al. 2018. Adversary Models Account for Imperfect Crime Data: Forecasting and Planning against Real-world Poachers. In *AAMAS*. 823–831.

[24] Carla Gomes, Thomas Dietterich, Christopher Barrett, Jon Conrad, Bistra Dilkina, Stefano Ermon, Fei Fang, Andrew Farnsworth, Alan Fern, Xiaoli Fern, et al. 2019. Computational sustainability: Computing for a better world and a sustainable future. *Commun. ACM* 62, 9 (2019), 56–65.

[25] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* 202 (2017), 18–27.

[26] Ben Green. 2019. "Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*.

[27] Saul Greenberg. 2020. *Automated Image Recognition for Wildlife Camera Traps: Making it Work for You.* Technical Report. Science.

[28] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[29] Hanxiang Hao, Sriram Baireddy, Emily R Bartusiak, Latisha Konz, Kevin LaTourette, Michael Gribbons, Moses Chan, Mary L Comer, and Edward J Delp. 2020. An Attention-Based System for Damage Assessment Using Satellite Imagery. *arXiv preprint arXiv:2004.06643* (2020).

[30] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies* 2019, 1 (2019), 133–152.

[31] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Anjali Jaiswal and Madhura Joshi. 2020. Climate Action: All Eyes on India. https://www.nrdc.org/experts/anjali-jaiswal/climate-action-all-eyes-india

[33] Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. 2020. Revisiting Membership Inference Under Realistic Assumptions. *arXiv preprint*

arXiv:2005.10881 (2020).

[34] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[35] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700* (2019).

[36] Maxime Lenormand, Sylvie Huet, Floriana Gargiulo, and Guillaume Deffuant. 2012. A Universal Model of Commuting Networks. *PLoS ONE* 7, 10 (2012).

[37] Xiyang Liu, Yixi Xu, Sumit Mukherjee, and Juan Lavista Ferres. 2020. MACE: A Flexible Framework for Membership Privacy Estimation in Generative Models. *arXiv preprint arXiv:2009.05683* (2020).

[38] Gary Marcus. 2019. An Epidemic of AI Misinformation. *The Gradient* (2019). https://thegradient.pub/an-epidemic-of-ai-misinformation/

[39] D Douglas Miller and Eric W Brown. 2018. Artificial intelligence in medical practice: the question to the answer? *The American journal of medicine* 131, 2 (2018), 129–133.

[40] Jared Moore. 2019. AI for not bad. *Frontiers in Big Data* 2 (2019), 32.

[41] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, and Juan Lavista Ferres. 2019. Protecting GANs against privacy attacks by preventing overfitting. *arXiv preprint arXiv:2001.00071* (2019).

[42] Md Nasir, Brian Baucom, Panayiotis Georgiou, and Shrikanth Narayanan. 2015. Redundancy analysis of behavioral coding for couples therapy and improved estimation of behavior from noisy annotations. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1886–1890.

[43] Md Nasir, Brian R Baucom, Craig J Bryan, Shrikanth S Narayanan, and Panayiotis G Georgiou. 2017. Complexity in Speech and its Relation to Emotional Bond in Therapist-Patient Interactions During Suicide Risk Assessment Interviews.. In *INTERSPEECH*. 3296–3300.

[44] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* (2018).

[45] Felipe Oviedo, Zekun Ren, Xue Hansong, Siyu Isaac Parker Tian, Kaicheng Zhang, Mariya Layurova, Thomas Heumueller, Ning Li, Erik Birgersson, Shijing Sun, et al. 2020. Bridging the gap between photovoltaics R&D and manufacturing with data-driven optimization. *arXiv preprint arXiv:2004.13599* (2020).

[46] Kasey Panetta. 2020. 5 Trends Drive the Gartner Hype Cycle for Emerging Technologies, 2020. https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/

[47] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED: an armed conflict location and event dataset: special data feature. *Journal of peace research* 47, 5 (2010), 651–660.

[48] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, 4 (2010).

[49] Zekun Ren, Felipe Oviedo, Maung Thway, Siyu IP Tian, Yue Wang, Hansong Xue, Jose Dario Perea, Mariya Layurova, Thomas Heumueller, Erik Birgersson, et al. 2020. Embedding physics domain knowledge into a Bayesian network enables layer-by-layer process innovation for photovoltaics. *npj Computational Materials* 6, 1 (2020), 1–9.

[50] Caleb Robinson and Bistra Dilkina. 2018. A machine learning approach to modeling human migration. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–8.

[51] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. 2019. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12726–12735.

[52] Caleb Robinson, Anthony Ortiz, Kolya Malkin, Blake Elias, Andi Peng, Dan Morris, Bistra Dilkina, and Nebojsa Jojic. 2019. Human-Machine Collaboration for Fast Land Cover Mapping. *arXiv preprint arXiv:1906.04176* (2019).

[53] Marcos Rodrigues and Juan de la Riva. 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software* 57 (2014), 192–201.

[54] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2019. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433* (2019).

[55] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. 2019. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution* (2019).

[56] Youssef Safi and Abdelaziz Bouroumi. 2013. Prediction of forest fires using artificial neural networks. *Applied Mathematical Sciences* 7, 6 (2013), 271–286.

[57] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI.

[58] Stefan Schneider, Saul Greenberg, Graham W Taylor, and Stefan C Kremer. 2020. Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and evolution* 10, 7 (2020), 3503–3517.

[59] Daniel Sheldon, Andrew Farnsworth, Jed W Irvine, Benjamin Van Doren, Kevin F Webb, Thomas G Dietterich, and Steve Kelling. 2013. Approximate Bayesian inference for reconstructing velocities of migrating birds from weather radar. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[60] Tao Sheng, Chen Feng, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Mickey Aleksic. 2018. A quantization-friendly separable convolution for mobilenets. In *2018 1st Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2)*. IEEE, 14–18.

[61] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818* (2020).

[62] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[63] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[64] Shijing Sun, Armi Tiihonen, Felipe Oviedo, Zhe Liu, Janak Thapa, Noor Titan Putri Hartono, Anuj Goyal, Clio Batali, Alex Encinas, Jason Yoo, et al. 2021. A Physical Data Fusion Approach to Optimize Compositional Stability of Halide Perovskites. *Matter* (2021).

[65] Saeid Asgari Taghanaki, Yefeng Zheng, S Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. 2019. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics* 75 (2019), 24–33.

[66] Erik Trautman. 2018. The Virtuous Cycle of AI Products. https://www.eriktrautman.com/posts/the-virtuous-cycle-of-ai-products

[67] Anusua Trivedi, Sumit Mukherjee, Edmund Tse, Anne Ewing, and Juan Lavista Ferres. 2019. Risks of Using Non-verified Open Data: A case study on using Machine Learning techniques for predicting Pregnancy Outcomes in India. *arXiv preprint arXiv:1910.02136* (2019).

[68] Tinka Valentijn, Jacopo Margutti, Marc van den Homberg, and Jorma Laaksonen. 2020. Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment. *Remote Sensing* 12, 17 (2020), 2839.

[69] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. 2018. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232* (2018).

[70] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 1–10.

[71] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. 2019. The immune epitope database (IEDB): 2018 update. *Nucleic acids research* 47, D1 (2019), D339–D343.

[72] Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656* (2012).

[73] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B Lobell. 2020. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing* 12, 2 (2020), 207.

[74] Ben G Weinstein. 2018. A computer vision for animal ecology. *Journal of Animal Ecology* 87, 3 (2018), 533–545.

[75] Christine T Wolf. 2020. Democratizing AI? experience and accessibility in the age of artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 26, 4 (2020), 12–15.

[76] Carole-Jean Wu, David Brooks, Kevin Chen, Douglas Chen, Sy Choudhury, Marat Dukhan, Kim Hazelwood, Eldad Isaac, Yangqing Jia, Bill Jia, et al. 2019. Machine learning at facebook: Understanding inference at the edge. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 331–344.

[77] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* (2018).

[78] Lily Xu, Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Rohit Singh, Mustapha Nsubuga, Joshua Mabonga, Margaret Driciru, et al. 2020. Stay Ahead of Poachers: Illegal Wildlife Poaching Prediction and Patrol Planning Under Uncertainty with Field Test Evaluations (Short Version). In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1898–1901.

[79] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 1191–1198.

[80] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* 15, 11 (2018), e1002683.

[81] Ming Zhong, Manuel Castellote, Rahul Dodhia, Juan Lavista Ferres, Mandy Keogh, and Arial Brewer. 2020. Beluga whale acoustic signal classification using deep learning neural network models. *The Journal of the Acoustical Society of America* 147, 3 (2020), 1834–1841.

[82] Ming Zhong, Jack LeBien, Marconi Campos-Cerqueira, Rahul Dodhia, Juan Lavista Ferres, Julian P Velev, and T Mitchell Aide. 2020. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Applied Acoustics* 166 (2020), 107375.

[83] Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (2018), 44–53.