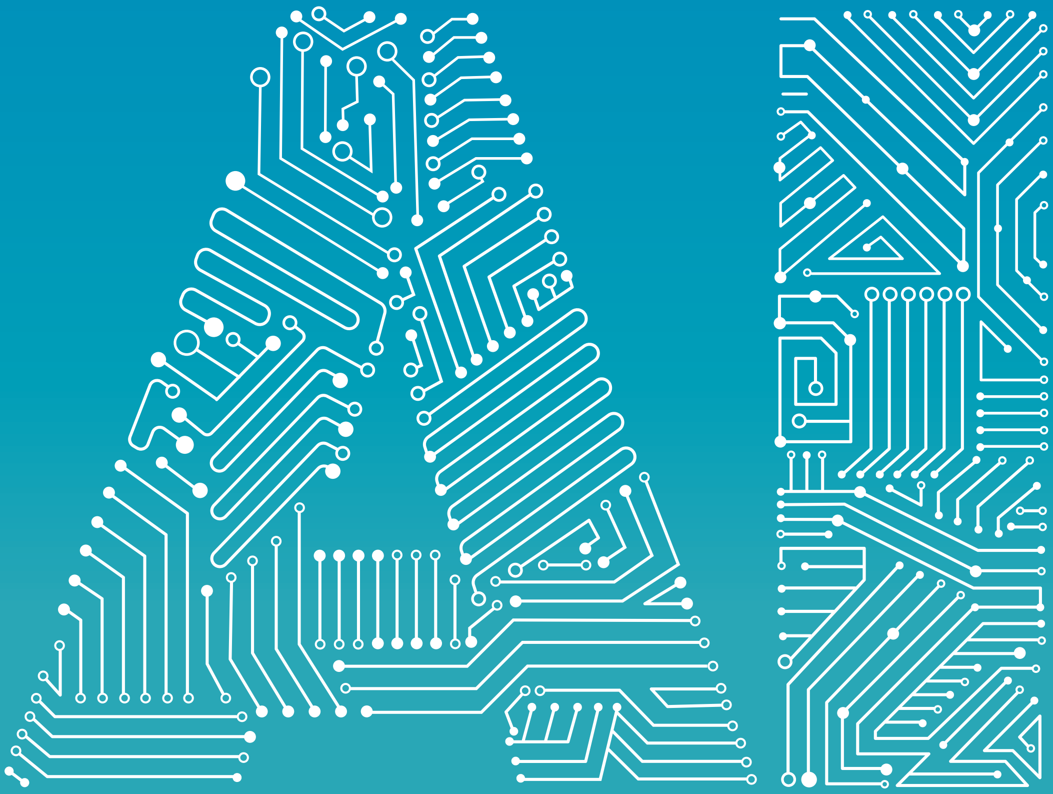


AI IN SOCIAL PROTECTION – EXPLORING OPPORTUNITIES AND MITIGATING RISKS

giz Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

ADB

Imprint

Author:
Tim Ohlenburg

Coordination:
Christina Betting, Dominique Leska

Editing:
Susan Sellars

Published by:
Deutsche Gesellschaft für Internationale Zusammenarbeit
(GIZ) GmbH

Registered offices:
Bonn and Eschborn

Address:
Friedrich-Ebert-Allee 32 + 36
53113 Bonn / Germany
T +49 228 44 60 - 0
F +49 228 44 60 -17 66

Dag-Hammerskjöld-Weg 1-5
65760 Eschborn / Germany
T +49 6196 79 - 0
F +49 6186 79 - 0

E info@giz.de
I www.giz.de

Design/Layout:
Bettina Riedel / Frankfurt

Copyright Cover:
istock.com

URL links:
Responsibility for the content of external websites linked in this publication always lies with their respective publishers. GIZ expressly dissociates itself from such content.

Responsible for the content:
Ralf Radermacher (GIZ)

Bonn, April 2020

Acknowledgements

The author would like to thank Ahmet Fatih Ortakaya, Alex Beyers, Alicia Spengler, Amy Paul, Craig Jolley, Dominique Leska, Harry Moroz, Joscha Albert, Kelvin Hui, McNeil Mhango, Michael Rothe, Minas Sifakis, Nicole Celikkesen, Pablo Suarez, Paul Dos Santos, Paul Niehaus, Phillippe Leite, Ralf Radermacher, Raul Ruggia-Frick, Raza Rizvi, Saurav Bhattarai, Seok Yong Yoon, Shachee Doshi, Simba Runyowa, Sri Wening Handayani, Thomas Abell, Uwe Wahser, Wendy Walker, Valentina Barca and Won Whang for their advice and support for this report.

Foreword

Artificial intelligence – AI – is the buzzword of the day, but not one we should dismiss as a fad. Too profound is the impact on our interactions at the personal, professional, social and political levels to believe that it is nothing but hype. The parallel surge in digital data, algorithms and computing power has created a wave of innovation that shows no sign of ebbing. Naturally, such a momentous development is debated heatedly and public perceptions veer between hope and despair. On the one hand, great strides in productivity, services for the previously excluded and entirely new offerings are held out as great promises. On the other hand, there are concerns that AI poses a risk to numerous professions, threatens privacy and can make people increasingly dependent on automated processes that they find opaque and unaccountable.

The public debate about AI itself will partly determine the shape of things to come and it is our duty as political advisors to make balanced contributions that help to steer us in a socially-beneficial direction. AI is one manifestation of the trend towards digitisation that offers exciting prospects for development cooperation. Mobile phones were the first instance of digital technology touching the lives of the poor on a large scale. The spread of mobile money in East Africa soon showed that communication networks could be leveraged to provide financial services, offering financial inclusion to those not served by traditional service providers. We are currently witnessing the proliferation of smartphones, which may well increase the speed, scale and scope for leapfrogging in ways we can‘t even imagine yet.

As the lives of people all over the world are becoming digitised, public sector services have accelerated in their transformation into the digital age. Social protection is an important field for this transformation because it touches on fundamental areas of public service provision and supports progress towards several of the UN Sustaninable Development Goals (SDGs):

- Taken together, the general aim of social protection programmes is to eliminate poverty, the first SDG.

- Food subsidy programmes for the poor combat hunger, SDG 2.

- Health insurance programmes promote good health, SDG 3.

- Education-supporting programmes such as cash transfers conditional on school attendance facilitate good education, SDG 4.

Social protection is an area of growing policy concern as many low and middle income countries shift their focus from combating extreme poverty towards a more equitable distribution of the spoils of economic growth, which supports SDG 10, the reduction of inequality. Despite the breadth of programmes that fall under the social protection umbrella, their implementation typically follows a common scheme. Many activities are the same, be it in social insurance or cash transfers, and it is at the activity level that technology can support the digital transformation of these programmes. In particular, this report seeks to identify activities that can be enhanced by AI solutions in the common activity pillars of outreach and registration, enrolment, and monitoring and management. There is great diversity in provision, and here, too, this report offers perspective on productive and innovative uses of AI for more effective social protection programmes and, ultimately, accelerated progress towards key SDGs.

To sum up, we believe that digitisation will have a profound impact and that AI is a prime example of the transformative technologies in our sights. Of course, the technology is still in its infancy and this report will hardly be the last word on the topic. Instead, its purpose is to explore how AI will affect social protection as a key area of policy interest. For one, we hope to present a perspective on the potential benefits that AI can bring to the table across social protection programmes. No less importantly, we wish to highlight the pitfalls and risks that could undermine the potential for positive change. Our intention is to promote discussion and encourage careful and responsible experimentation among you, the practitioners, who will need to turn this vision into a reality.

Dr Axel Klaphake
Director of Division
Economic and Social Development, Digitalisation
GIZ

Woochong Um
Director General
Sustainable Development and
Climate Change Department
Asian Development Bank (ADB)

Table of Contents

1.
Setting the Scene

Artificial intelligence (AI) is shaping up to be the transformative technology of our time. Concurrent advances in computational power, algorithms and digitisation have led to an explosion of applications that some call the Fourth Industrial Revolution. As with previous technological waves, its impact will not be limited to industry, but is set to enter the public sector, with it already making inroads into our homes and pockets.

A number of **recent achievements** illustrate the wideranging nature of AI. A superhuman ability to classify images was the forerunner of a wave of recent advances driven by deep learning methods. Since then, in the strategy domain, computers have learnt to beat humans in Go, the most complex of traditional human games, have defeated champions of modern strategy games such as StarCraft, and are beating top class players at poker, surely the pinnacle of human cunning. In natural language processing, a recent model developed by OpenAI (OpenAI, 2019) was able to generate such realistic text that its designers consider it dangerous in the wrong hands amid concerns about online disinformation campaigns and fraud. In our personal and professional lives, we have long benefited from AI technology. Spam filters have protected us from scams since the mass adoption of email, we carry phones that can respond to commands, and we may soon be able to converse with our toothbrush.

So far, advances have been driven by the tech sector, in which the big players are investing billions in research, development and the deployment of the latest tools. In view of the economic importance of the field, governments are looking to catch up and several countries have formulated AI strategies to foster competitive national innovation systems, including Germany (Bundesministerium für Bildung und Forschung, 2018). However, the adoption of **AI in the public sector** remains limited. The reasons for the relatively slow uptake include ethical and legal concerns, as well as scepticism about whether computer-driven systems are appropriate in the sphere of public policy and administration. The development cooperation sector is an example of a policy field in which exploratory efforts are under way, but where practical applications remain few and far between. Nonetheless, the promise of productive applications for poverty alleviation is great and various bodies, including GIZ and USAID, have formed dedicated teams to turn it into reality.

Knowing the practical realities, the experts on these teams will be among the first to sound **a note of caution** amid the widespread excitement about an impending AI revolution. Disappointment often follows initial euphoria and market research company Gartner (Gartner, 2018) already reckoned in 2018 that deep learning had reached the peak of the hype cycle. A brief look back at the surge in biotech stocks or the dotcom boom suggest that expectations of an imminent revolution may be premature. Changes are likely to be far reaching, given the current momentum and broad applicability, but our world is unlikely to be overturned overnight. In particular, the lives of the poor, who lack basic amenities and depend on social protection programmes for their livelihoods, can only be changed so much by sophisticated deep learning systems living in the cloud.

What are the special characteristics of AI technology and how does it relate to applied work in development cooperation? In what ways can AI contribute positively and what risks must be managed? Which use cases showcase effective applications and what wider conclusions can be drawn from them? Finally, what concrete, actionable advice can we give to those funding or collaborating with AI-based projects? As a promising technology, **our approach** to the topic is cautious optimism. Our expectation is of evolution rather than revolution, for the most part. As we set out below, we believe that mitigating the negative effects of AI is as important as realising its benefits, particularly in a field like social protection that is focused on supporting the vulnerable and excluded lest it create a new dimension of exclusion.

This report presents a framework for thinking about how AI can make a positive difference in social protection. We begin our discussion with an introduction to key concepts that set the scope for the discussion, in particular presenting AI systems as a data science pipeline. Our listing of benefits is mirrored by a discussion of the risk and we end the introductory chapter with a note on AI systems in low resource environments. The second chapter links AI to social protection, looking at typical tasks along the delivery chain that most programmes share from a practical perspective and giving tentative suggestions for each area. To illustrate actual efforts, the next chapter reviews some use cases, which are meant to bring the previous ideas to life as well as share cautionary tales from the coalface of implementation. The final chapter concludes with suggestions for practitioners on the evaluation and assessment of AI initiatives in their field.

What we talk about when we talk about AI

The term ‘artificial intelligence’ means different things to different people (Russell & Norvig, 2016), just as ‘intelligence’ differs in its use depending on the context. We begin this section by describing how we interpret AI for the purpose of this report and how it relates to a few relevant concepts. To start with, let us outline what we think of as AI, and then clarify what we see as not AI. In his futurist tour de force, *The Age of Intelligent Machines*, Ray Kurzweil (1990) proposes an instructive definition of AI:

“The art of creating machines that perform functions that require intelligence when performed by people”.

The core of this definition is the ability to act in a way that resembles human intelligence, a practical perspective we find most useful for the task at hand. The allusion to human intelligence suggests an ability to deal with new situations, to handle previously unseen problems. Endowing a machine with this ability means that we must give it a mechanism to generalise from a specific case to find a more general underlying relationship. In effect, it must be able to learn how to relate the determining features of a situation to the performance of a function.

Machine learning

Machine learning is the key technology that underlies the recent wave of AI innovations. Its core discipline of supervised learning is concerned with the task of teaching a machine to learn from examples. Neural networks, the models behind deep learning, have been tremendously successful because they can learn any relationship that exists in a dataset between the situation determinants (call them inputs) and the appropriate response (or output). The capacity to learn sophisticated input-output mappings, which pick up broad relationships, subtle effects and complicated patterns, is the secret sauce of supervised learning methods which allow machines to perform tasks in human-like or even superhuman fashion. Apart from the flexible mapping of inputs to outputs of supervised learning, the other widely used machine learning discipline is called unsupervised learning. This set of methods finds order in unstructured information, allowing the user to identify patterns and distil masses of data into a more manageable format.[1]

Machine learning is a computational discipline with theoretical roots in statistics, probability theory, mathematics and algorithmics. A key concern of the statistical learning theory that forms its basis is how many examples an algorithm needs to be able to learn a relationship and how much computational time this learning will consume. These questions are far from abstract, as the practical implementation requires sufficient data and computational resources. The bottom line is that machine learning systems require data and computation to achieve human-like functionality in complex tasks. Before moving on to related concepts, it is worth pointing out a commonly misinterpreted aspect of machine learning. A common assumption is that machine learning systems improve as more data is collected: as they learn, much like people, they should also keep learning as they keep performing a task, as people do. Although specific machine learning methods exist that can do this – online learning and reinforcement learning represent the major branches – regular supervised learning methods such as neural networks are not set up this way. Instead, they are built in a training stage and deployed in a prediction stage, where the latter doesn’t feed back into the former unless the model is re-trained. This illustrates a limitation of the intelligence analogy for deep learning. More practically, it also highlights that the machine learning systems need to be re-tested and adjusted as conditions evolve if they are to be effective over the life of a government programme.

[1] Reinforcement learning, the third pillar of machine learning, is concerned with learning from interactions with the environment. Although impressive in simulations, it is the least practically developed of the three. The separation is somewhat arbitrary today in any case as modern systems often draw on more than one pillar. For example, DeepMind’s stable of AlphaGo, AlphaZero and AlphaStar use both reinforcement and supervised learning methods to beat humans at their game. For a broad overview of machine learning and its use in development cooperation we recommend *Reflecting the past, shaping the future: Making AI work for international development* (USAID, 2018).

Data science

The deployment of a machine learning-based system requires skills and resources in data management, statistical modelling and coding, which all fall under the umbrella of data science. The University of California at Berkeley is a pioneer in data science teaching and one of its course descriptions offers a useful graphical summary. According to Figure 1, data science can be described as a sequence of tasks.

Figure 1 illustrates the following processes:

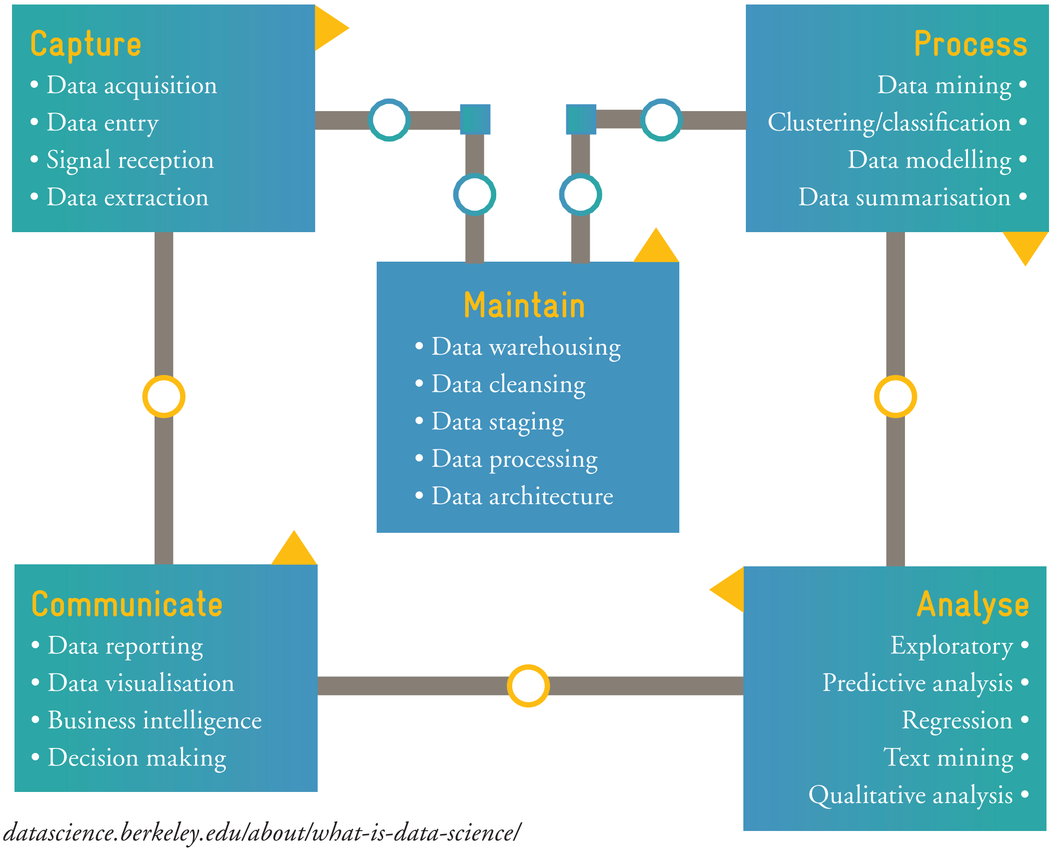
- **Capture:** The collection (and entry, if not done automatically) of data, as well as the possible conversion of raw data into informative signals, is the first step in the data science life cycle. Data collection can take many forms and may involve surveys, remote sensing, or the acquisition of administrative data from other sources, among other things.

- **Maintenance:** Once captured, data needs to be pre-processed before it is usable for modelling, available as and when required for processing, and warehoused to secure it. Pre-processing typically involves data cleaning, transformation and feature extraction and often accounts for a significant proportion of data scientists' time.

- **Processing:** The foundational task of data analysis is to understand its basic structure and relationships. The elimination of inessential data and the selection and implementation of a suitable model that represents the essence of the data-generating process are the most important skills of the data scientist. For an AI system, the data scientist would build a machine learning model at this stage.

- **Analysis:** Data processing and model preparation lead up to the analysis that finally yields information

Figure 1. The data science life cycle



Capture
- Data acquisition
- Data entry
- Signal reception
- Data extraction

Process
- Data mining
- Clustering/classification
- Data modelling
- Data summarisation

Maintain
- Data warehousing
- Data cleansing
- Data staging
- Data processing
- Data architecture

Communicate
- Data reporting
- Data visualisation
- Business intelligence
- Decision making

Analyse
- Exploratory
- Predictive analysis
- Regression
- Text mining
- Qualitative analysis

Source: datascience.berkeley.edu/about/what-is-data-science/

from the captured digital data. This most interesting step is the culmination of previous efforts. During this step, AI systems draw on the machine learning model's predictions.

- **Communication:** For information to become knowledge, it needs to be contextualised and communicated. Data science thus includes tools such as data visualisation for communicating insights and reporting actionable insights.

What these steps illustrate is that processing and analysis are only two of five steps in the data science process and AI mainly features in the processing and analysis steps. Machine learning methods can form the computational heart of what we term a pipeline that pumps data after its capture through a decision-making system that would have previously required a human. The key point is that AI requires a supporting system, and also that it can be deployed in many situations, as long as such a system can be construct-ed. We will return to this point later, arguing that AI deployment means finding space for a data science pipeline within a wider delivery system.

Big data

We will now briefly digress to talk about 'big data', a term often used in relation to AI. For practical purposes big data refers to collections of data with either many dimensions (like a spreadsheet with very many columns) or lots of data points (a long database with very many entries), and often both. A typical government programme database, such as a social registry, does not constitute big data. Even though it may contain millions of user entries in an SQL database and may be too large for a standard computer, a regular server architecture loaded with appropriate programmes should be able to perform most relevant tasks, including the running of analytics and machine learning methods (except for the registries of a few of the most populous nations).

Big data needs to be stored in an NoSQL database and has to be analysed via specialised systems for distributed computation, such as Hadoop, as the information is too vast for one computer to manage. Examples of big data are the on-site actions of Facebook users, the masses of high frequency transactions on a digital financial market or the images uploaded to Instagram. These data volumes require specific methods and skills, representing a distinct field that draws on machine learning and uses specialised methods, but is not synonymous with it. Although public systems can hold extensive data, such requirements may only arise for digital platforms with frequent user interactions or systems in populous nations. However, that is actually preferable as it allows a more flexible data science pipeline without specialised tools and staff. We can say that big data needs AI, but AI doesn't need big data.

Expert systems & AGI

In the early history of AI, the main development strand was to build logic-based systems that reason in specific ways and to encode human knowledge in decision rules. Although highly effective in a range of scenarios, the approach essentially codifies responses to particular situations. When the number of situations that the system is likely to encounter is large, then it is impossible to decide a suitable action for each one. Due to the lack of the generalisation property of machine learning systems, expert systems are unable to deal with situations that fall outside pre-specified parameters. This fails the intelligence criterion of being able to act as a human would.

Expert systems have an important place in information and communication technology (ICT) systems. The encoding of insights about the processes that the data represents is a key part of system development. However, when a desired system behaviour can be defined in code then there is no need for a machine learning component that mimics learning and approximates the human insight. Building an AI system for a task that can be handled well with a rule-based definition devised by humans is simply unnecessary. Most of the core functionality of a government ICT system falls into this category. Although system administrators still need to work with subject matter experts, there is no need to involve learning algorithms.

If expert systems are the past of AI, efforts to build the future are focused on the achievement of 'artificial general intelligence', or AGI. For instance, Google's DeepMind subsidiary has the pursuit of AGI as its foundational objective and its game-playing machines are only a stepping stone. The aim is to build machines that solve increasingly complex problems until they can reason, converse and act like intelligent beings. Despite impressive topical advances such as image recognition and speech generation, computers still lack any conception of the physical nature of the objects they name or the significance of the text they generate. The achievement of true reasoning ability and abstract thinking, which AGI efforts strive towards, is still a long way off. For better or worse, we must make do with developing systems that are very good at a narrow range of tasks and embed these skills in a carefully crafted ICT pipeline.

We can summarise that expert systems with encoded human knowledge can learn what to do in a particular situation. Machine learning methods trained on data can learn a specific rule, allowing them to handle any situation that fits the rule. In future, AI may be able to determine the type of rule that it needs to apply to a situation, inferring from the rules for other situations what the appropriate approach should be. However, this remains a distant prospect.

Opportunities

Machine learning based AI systems can offer major benefits, when the conditions are right. Computer systems have advantages over humans when they can perform a similar task and labour can be replaced or augmented with an AI pipeline. We propose six such advantages and discuss each briefly.

Superhuman capability	There are many examples of tasks in which AI systems outperform humans by a significant margin in terms of accuracy. A classic problem is to classify the content of images from a large collection called ImageNet. In this benchmark task, humans recognise around 95%[2] of images correctly, whereas the best current machine learning classifier achieves 98%[3]. Although such success is unrealistic for all tasks, a higher performance level may still make a compelling argument for an automated solution.
Speed & productivity	The speed of execution is a major benefit for large-scale government systems, as the machine learning stage of the pipeline generally takes a split second per case. Speed is indicative of productivity and one system can process many cases quickly.
Scalability	Scaling up or down a manual administrative process requires adjustments in headcount. Hiring requires costly training and the acquisition and equipping of physical facilities can also be a major cost factor. The benefits are greater the more the process is digital, as storage and processing capacity can be adjusted easily in the cloud and capacity utilisation kept high.
Reliability	AI systems perform steadily, albeit within the confines of their statistical accuracy. They don't get tired or have days off, which may be of considerable benefit to service quality.
Flexibility	Models can be adjusted or exchanged with relative ease as better options are found. If new data is used, the entire pipeline needs to be adjusted, but similar changes may be necessary if the task was to be executed by humans. Training for human operators is not necessary and operational practices don't need to be changed.
Unit cost	Once an AI system is in place, the cost of a single action or decision is practically zero. Of course, this calculation assumes that the running costs are spread over many cases. The scale at which AI systems become economical depends on the data, hardware, software and operational aspects of the programme. Whereas a single data scientist may be able to implement a system that can draw on existing data structures and is entirely digital in nature, a nationwide roll-out of a new image recognition system is an example of the other end of the resource intensity scale. A detailed cost-benefit analysis (see Suggestion 9 in the final chapter) that estimates expenses along the project life cycle and also quantifies the gains is advisable ahead of major investments.

[2] Russakovsky *et al.* (2015) provide an historical background on the task and https://paperswithcode.com/sota/image-classification-on-imagenet lists the current state of the art.

[3] https://paperswithcode.com/sota/image-classification-on-imagenet

Risks

The promise of AI is great, but its risks are daunting. When applied to government programmes – which should further the public interest – there is a particular need to manage risk and mitigate adverse effects[4]. The legal, moral and ethical issues around the use of AI in public administration are a complex topic that is far from fully explored. We can only scratch the surface of such issues here and suggest reference to *Reflecting the past, shaping the future: Making AI work for international development* (USAID, 2018) as a starting point for a deeper exploration of the various issues raised by AI methods in public policy and international cooperation.

Data protection	Personal data on livelihoods and health status is some of the most sensitive information that exists and its disclosure can have major consequences for the data subjects. The careful curation and provision of data in AI systems can offer full and easy access to that personal data to anyone with access to the system. The risk is multiplied where AI systems compile information from several sources. Where AI systems result in greater data compilation, their creators are obliged to use data protection as a design criterion.
Data poverty	The support of secure, sustainable livelihoods is the aim of welfare programmes. Although data-driven aspects of programme delivery can lower the cost and increase the quality of provision, they rely on potential beneficiaries being represented in the data used for identification and targeting. Individuals may be excluded because they have no access to the data-generating technology that the AI system relies on, such as a mobile phone. Access to digital services tends to rise with income, so the poorest are most likely to be data poor as well.
Bias & discrimination	The replacement and augmentation of human judgement with a machine learned process can lead to greater reliability and consistency of application. However, it can also perpetuate historical biases and institutionalise systematic discrimination (e.g. Zou & Schiebinger, 2018). Major challenges for AI revolve around biased data, the direct and indirect use of prohibited variables and unequal accuracy. When a learning algorithm is presented with data that reflects historical discrimination, it will learn to imitate the biased patterns of the past. Computers cannot distinguish between ethical and unethical decisions and, although the data can be adjusted for known biases, data scientists need to be aware of historical biases before they can be addressed. A second (and often related) issue is that the model may find that certain variables which should not be used as a basis for decision-making are useful statistically and will be incorporated into the model if provided to the machine. An example could be the ethnic origin of a potential programme beneficiary. Unless the programme is specifically targeted at ethnic groups, in most situations and societies an ethnicity-based decision is unlawful and inappropriate. Excluding ethnic origin as a direct variable is certainly called for, but may not be sufficient

[4] A GIZ report on automated decision-making in financial services (Dix *et al.*, 2019) covers risks and mitigation from a regulatory perspective.

Bias & discrimination	as it may correlated with other, permissible variables. For instance, a combination of other variables can serve as a proxy for ethnicity and have the same effective result as including ethnicity as a variable. The boundary between legitimate and problematic predictors can be hard to set, requires careful analysis for each specific context and remains an open research area. Another issue is that some groups may be under-represented in the data, possibly due to data poverty or simply because they are in the minority. The characteristics of minorities are likely to differ in ways that have a bearing on government programmes. Machine learning algorithms are usually trained for overall accuracy and may not put much weight on minorities, leading to relatively more incorrect outcomes for such groups due to their different characteristics. Mitigating approaches such as hierarchical models or multi-task criteria exist, but they can only be used when the AI system is audited and adjusted for such effects, requiring awareness of the risk to begin with.
Transparency & interpretability	Unlike more traditional statistical methods that use linear models which humans can reason about, AI systems often rely on non-linear methods that are hard or impossible to understand in detail. The lack of direct interpretability can undermine trust and impede the adoption of such methods. On the other hand, the patterns recognition power of AI often relies on complex models and represents a trade-off between model interpretability and predictive accuracy. A simple call to restrict methods to interpretable ones ignores the benefits that 'black box' models may provide for beneficiaries. A case-by-case consideration of appropriate model choice is called for and tentative model audit measures are suggested in the final section.
Overfitting & spurious correlation	Although modern statistical methods are powerful tools, they are far from fail-safe and need to be handled with care and expertise. The chimera of machine learning methods is the issue of 'overfitting'. This describes a situation where an algorithm extracts patterns from a dataset that are particular to the data, but which don't hold true in general. Machine learning methods will identify patterns and it is up to the data scientist to constrain them to generally valid patterns rather than arbitrary and specific ones. With sufficient data, methods such as cross-validation can, and should, be deployed to build systems that function beyond the training data. The second major issue in statistical misspecification is spurious correlation. This occurs when the available data suggests a relationship between variables, but this is either coincidental or due to the presence of another, unmeasured variable. Most machine learning methods work on correlations and are not able to distinguish between that and causation. They also do not need to, as long as the

	relationship continues to hold. If there is an unmeasured cause and its relationship to the measured data changes, then the predictive model will be thrown off. A combination of domain expertise and modelling skills can help to avoid spurious correlations, but it cannot replace the continual comparison of predictions with actual results to identify model misspecification.
Deception & fraud	Fraud and deception in the digital sphere require technical expertise, but they can also be easier to achieve when there is no human oversight to check abuses. Deception may be by programme beneficiaries, for instance, through the provision of inaccurate data that leads to unwarranted receipt of benefits. Fraud may also be committed by programme administrators ranging from frontline staff with client contacts to back office staff such as database administrators. The insertion of ghost beneficiaries is a common problem in government programmes and here, too, a lack of human oversight can facilitate misbehaviour. Finally, beneficiaries and staff might collude in altering data in order to share the ill-gotten gains. Although a detailed consideration of this topic is beyond the scope of this report, we will touch on countermeasures for digital fraud later in the report.
Set-up cost	If it requires additional data collection, processing and storage, the introduction of an AI solution for a government programme can be costly to design and build. It is difficult to judge up front whether the ultimate benefits warrant the investment, as initial estimates and final results can differ widely. It should also be borne in mind that large ICT projects are fraught with execution risk, experience frequent cost overruns and have a considerable chance of outright failure. This report advocates for a gradual approach and integration with existing systems, rather than introducing a large, single purpose system.

Deployment in low and middle income countries

The construction of AI systems requires significant resources and some of these may be scarce in low and middle income countries (LMICs). We have reviewed the range of skills required from a data science perspective, but further inputs are required to deploy and maintain an AI system. As an integrated ICT solution, the failure of any one component can lead to an overall breakdown. Given the need for reliable performance in government services, in this section we briefly highlight critical factors of production that may be in short supply and mention areas of special opportunity.

Data constraints	Digitisation has been the main driver of data availability, be it at the level of businesses, households or the public sector. At the risk of undue generalisation, public sector systems in LMICs are not as widely digitised. This poses an obvious problem as data is the primary input for AI systems. Beyond programme-specific information, complementary datasets for AI systems may be unavailable. An example is limited data for low-resource languages. An example of unsuitable data is that image recognition software struggles with people of colour (Grother, Ngan & Hanaoka, 2019)[5].
Infrastructure	AI systems are commonly deployed online via cloud infrastructure and require a steady high-speed Internet connection. Also, when run on a server, outages in communications infrastructure and intermittent electricity supply can be a problem in some countries. Back-up infrastructure may be necessary to guarantee consistent uptime, which raises running costs.
Human resources	The construction and maintenance of AI systems requires an extensive skill set. While data capture and maintenance skills may be relatively easily available, demand for data scientists, machine learning engineers and development operations engineers – who are necessary to build and maintain AI systems – exceeds supply in many countries. Even if aspects of the development process can be handled remotely, the medium-term maintenance and further development of AI systems requires local staff with the appropriate skill set. Those skills may be out of reach for public sector bodies with rigid pay scales and, even where the skills are available and affordable, the in-demand nature of the personnel poses retention challenges that ought to be considered in operational planning. A combination of training for existing staff and the hiring of subject experts is likely to be required.
Regulatory environment	The governance of data privacy is an evolving policy topic the world over. The European Union's General Data Protection Regulation (GDPR) probably represents the most ambitious and far-reaching effort to date, but its comprehensive practical implementation remains elusive. Regulatory frameworks are progressing in many countries – for example, Ghana's is well-established and Uganda passed a new

[5] This US government report shows the problem with commercial face recognition systems https://www.nist.gov/sites/default/files/documents/2019/07/03/frvt_report_2019_07_03.pdf

	law in 2019 – but cannot be taken for granted everywhere. In the absence of a clearly defined and enforced framework, the duty of care of AI system operators demands adherence to alternative standards. Although public bodies are often excluded from such laws, responsible practice calls for equivalent rules and procedures.
Mobile money, data & services	Although digitisation is probably strongly correlated with income levels, one area where LMICs are leading in the digital sphere is the use of mobile money. East African nations are the pioneers of the technology and offer an exciting glimpse of the future in terms of digital connectivity. Mobile phone usage has spread rapidly in LMICs and smartphones are becoming widespread even in the least affluent corners of the world. Beyond telecommunication, they represent a connection to the digital world of social networks, media, education, entertainment and digital services. In combination with mobile money, which already offers the opportunity to provide financial services to the previously excluded, these devices offer a great scope for leapfrogging in delivery systems, as they offer direct access to potential beneficiaries, digital data and a more extensive financial link to financial services.
Biometrics	An AI-based technology that is of particular relevance to population-facing services in LMICs is biometric identification. The provision of a legal identity for all is a Sustainable Development Goal for 2030 and the World Bank's ID4D Initiative highlights the wide-ranging uses and importance of enabling citizens to verify their identity. Public-facing programmes typically involve a registration function and a unique identifier based on biometrics can be assigned to potential beneficiaries (see Asian Development Bank, 2016 for examples from Asia and the Pacific such as India's Aadhar with over 1.2 billion registrants). Facial recognition, in which a computer vision system spots faces, maps them onto a 2D or 3D representation and then measures distances between characteristic points, is the most widely-used approach. Voice, fingerprint, skin texture and behavioural features are further options that can be combined for greater accuracy and security. Social registries, in which beneficiary records for several programmes and key personal information are linked, are a complementary approach that can leverage the power of unique identification and ID verification for beneficiary convenience and programme effectiveness. Biometric identities may be recorded for the registry or be linked to a national identity scheme and either approach can help combat fraud and corruption (World Bank, 2018). However, Barca and Makin (2018) point out that numerous risks need to be managed and that benefits only fully accrue if the system is carefully designed and implemented. In addition, Sepulveda Carmona (2018) analyses the privacy and data protection risks associated with biometrics in social protection programmes and proposes minimum requirements for safeguarding personal information. As touched on in the following section, particular risks (e.g. accuracy differences across ethnic groups and electronic fraud) also arise from the AI technology itself, underlining that modern technology is no panacea.

2. AI in social protection

The field of social protection covers a diverse group of programmes, ranging from cash transfers for refugees via mobile money to health insurance systems for entire populations and training schemes for specific demographic groups such as youth. Most countries, if not all, have some kind of social protection programme. So far, we have defined AI in fairly general terms and discussed its application to government programmes loosely. This section narrows the focus and explores its potential application to social protection programmes.

Let us start this discussion with a few words about the purpose, nature and beneficiaries of social protection programmes to give context to the ensuing discussion. The section draws heavily on the World Bank's extensive *Sourcebook on the foundations of social protection delivery systems* (Lindert *et al.*, forthcoming). It is worth quoting the opening paragraph as a concise and insightful **description of social protection:**

> *Most countries possess social protection systems that seek to build equity, opportunity, and resilience for their people. They provide support through a wide range of benefits and services that redistribute income to reduce poverty and inequality, support investments in human capital, and help insure against shocks and various risks, including poverty, loss of earnings from old age, economic crisis, natural disaster, climate change, conflict and forced displacement. Those interventions are typically grouped into three "pillars" of social protection: social assistance (non-contributory benefits and social services), social insurance (contribution-based benefits), and labor (both contributory and non-contributory benefits, as well as employment services).*

The **beneficiaries** of social protection programmes include various groups. Due to their age, the young and elderly are vulnerable. In terms of wealth, the income and asset-poor require special protection due to their lower ability to withstand economic shocks. The unemployed and underemployed have already experienced a disruption to their livelihood, a situation that also holds true in different forms for refugees. Disabled people may be unable to make a decent living without support and numerous barriers frequently prevent persons with disabilities form participating equally in society. Finally, certain social, ethnic or religious groups may be marginalised in economic life and require the assistance of social protection programmes. Despite the remarkable range of programmes, we will find much in common between them.

In the remainder of **this chapter** we proceed with an administrative look at social protection programmes that shares some similarity with the data science pipeline of AI systems, providing a joint frame of reference for the two. We will consider the common components of most social protection programmes and link them to potential AI applications that could enhance their functioning. The aim is to identify situations in which an AI data science pipeline can be plugged into the social protection delivery chain to augment current or planned operations. Note that we take the design of social protection programmes as given and focus on implementation. AI systems can likely fulfil useful roles in planning as well, a topic we leave for future investigation.

A **disclaimer** is in order before we launch into a discussion of the social protection delivery chain and attendant AI potential. So far, practical applications of AI in social protection programmes are limited, yielding little practical evidence to work with. Like any predictions of future technology, the potential applications presented here are speculative and unlikely to turn out quite as imagined. Rather than a confident assessment of things to come, the potential uses outlined here are intended to catalyse further discussion and narrow down the many possible directions to a few suitable ones.

The social protection delivery chain

At first sight, the many facets of social protection appear linked more by their social purpose than by similarities in operating procedures. From the perspective of fitting in AI applications, it might, therefore, appear that there would be relatively little overlap between the various programmes. However, closer observation reveals quite the opposite: there is surprising uniformity in the operational aspects of social protection programmes.

In the previous section, we made the case that AI systems can also be seen as data science pipelines that contain several complementary stages which are similar for all implementations. The AI aspect is represented by machine learning models that perform specific tasks. When framing social protection programmes as a series of steps, we can identify those that are amendable to automation through **integration of an AI data science pipeline** in the delivery chain. The key idea is that AI becomes a small component of the regular system, rather than a separate functionality. In view of the multitude of AI tasks that can now be handled automatically, a considerable share of social protection programme activities may qualify for AI augmentation.

The social protection **delivery chain** consists of four broad stages, each made up of a series of steps:

- **Assessment:** The first stage is to reach out to the population of intended beneficiaries, identify and register the candidates and assess their needs.

- **Enrolment:** The selection of intended beneficiaries from the candidate pool and the assignment of the appropriate benefit or service level is the next step, completed by notification and on-boarding.

- **Provision:** The diversity of programmes – transfers, insurance, labour market interventions – is such that single steps are specific to the particular programme. We can only summarise this diverse set as the ‘disbursement of benefits or provision of services’.

- **Monitoring & management:** In view of their social purpose, social protection programmes need to acknowledge and redress client grievances to safeguard service quality. Secondly, programme conditionality needs to be monitored. The administration also needs to comply with legal and performance standards, which need to be monitored internally. Finally, the exit of current beneficiaries has to be handled.

Figure 2 illustrates the delivery chain described in the World Bank Sourcebook (Lindert *et al.*, forthcoming). The rest of this section will review data sources for social protection programmes and the components of the delivery chain more closely. We describe the purpose of the steps, lay out their core activities and finally sketch AI solutions that may support some of the activities.

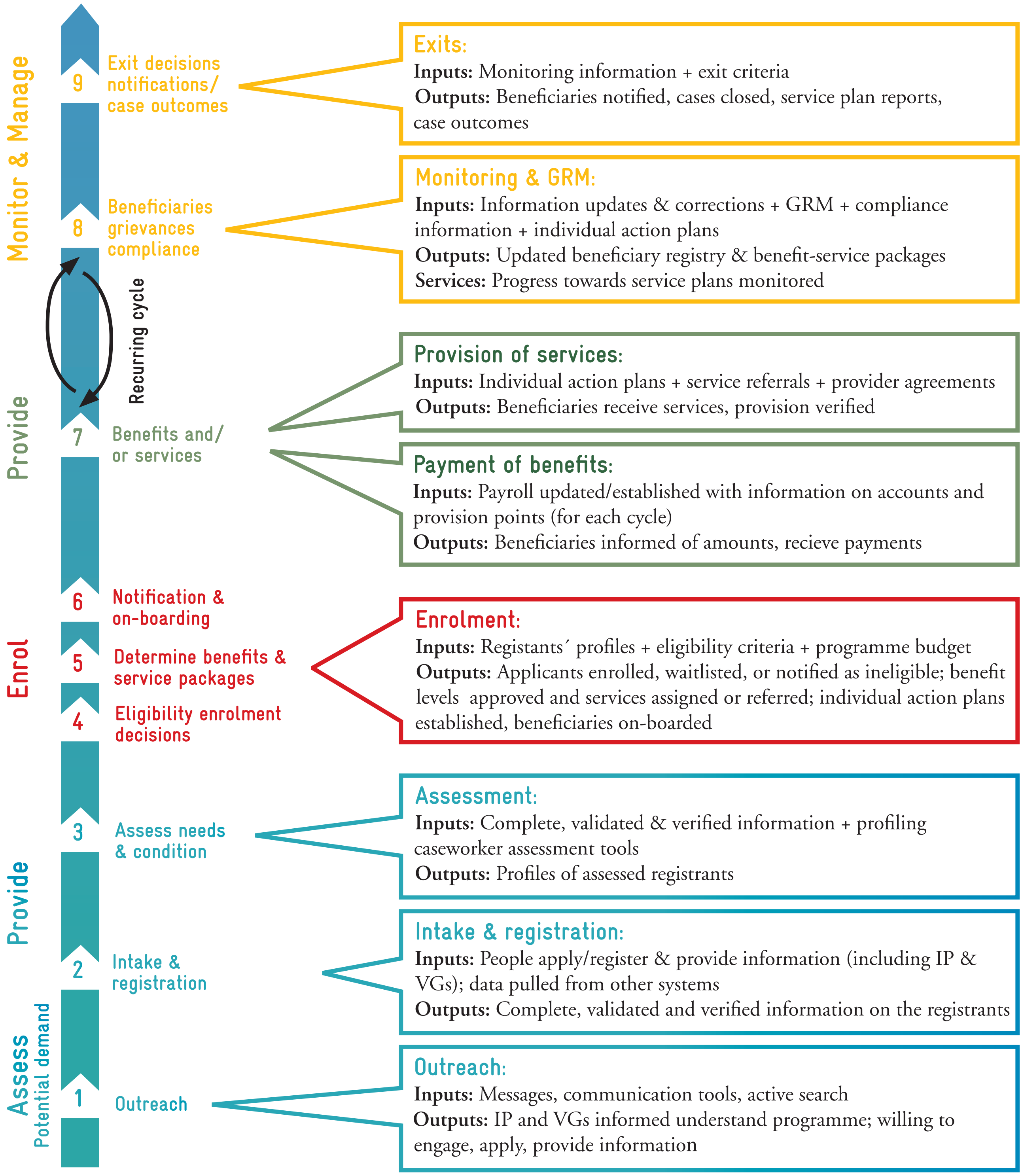
Data for AI in social protection can come from a variety of sources. The traditional collection of beneficiary data for specific programmes is a useful basis, but the combination of several sources can support decision quality with AI methods that thrive on a variety of information. Social registries (Leite *et al.*, 2017) that pool registration and assessment data from various programmes such as national ID systems, civil registries or the tax administration offer particularly relevant variables and represent a gateway to the digital future of social protection. Use cases and references for this report highlight data from domains other than the public sector, such as mobile phone records and geographical or climatic records collected by satellites.

The broader conception of adaptive social protection (Jorgensen & Siegel, 2019), which includes responses to conflict or humanitarian and climatic disasters and market variability, expands the scope of relevant information sources further. The example of Indonesia (Pahlevi, 2019) shows that social protection programmes are starting to collect non-traditional information, such as images of dwellings and GPS coordinates, opening the door to analytical and predictive methods from the geography and computer vision communities.

Finally, GIZ (Chirchir & Barca, forthcoming) provides a highly relevant consideration of social protection information system design and also addresses biometric data, an especially sensitive case of personal data use in programme operation.

A final note on combined sources is that the old phrase of ‘garbage in, garbage out’, which applies to all statistical and computational methods, is also applicable to automated decision making in social protection programmes. As accuracy is essential for the achievement of fair outcomes, only well maintained and verified sources provide a useful basis for decisions.

Figure 2. Social protection delivery chain


Monitor & Manage
9
Exit decisions notifications/ case outcomes
8
Beneficiaries grievances compliance
Recurring cycle
Provide
7
Benefits and/ or services
6
Notification & on-boarding
Enrol
5
Determine benefits & service packages
4
Eligibility enrolment decisions
3
Assess needs & condition
Provide
2
Intake & registration
Assess
Potential demand
1
Outreach

Exits:
Inputs: Monitoring information + exit criteria
Outputs: Beneficiaries notified, cases closed, service plan reports, case outcomes

Monitoring & GRM:
Inputs: Information updates & corrections + GRM + compliance information + individual action plans
Outputs: Updated beneficiary registry & benefit-service packages
Services: Progress towards service plans monitored

Provision of services:
Inputs: Individual action plans + service referrals + provider agreements
Outputs: Beneficiaries receive services, provision verified

Payment of benefits:
Inputs: Payroll updated/established with information on accounts and provision points (for each cycle)
Outputs: Beneficiaries informed of amounts, recieve payments

Enrolment:
Inputs: Registants´ profiles + eligibility criteria + programme budget
Outputs: Applicants enrolled, waitlisted, or notified as ineligible; benefit levels approved and services assigned or referred; individual action plans established, beneficiaries on-boarded

Assessment:
Inputs: Complete, validated & verified information + profiling caseworker assessment tools
Outputs: Profiles of assessed registrants

Intake & registration:
Inputs: People apply/register & provide information (including IP & VGs); data pulled from other systems
Outputs: Complete, validated and verified information on the registrants

Outreach:
Inputs: Messages, communication tools, active search
Outputs: IP and VGs informed understand programme; willing to engage, apply, provide information

Note: IP = intended population; VG = vulnerable groups; GRM = grievance redress mechanism Source: Lindert *et al.* (forthcoming)

Assessment & registration

The first stage of the delivery chain is to assess the target population, which consists of three steps: outreach, intake and registration, and assessment.

Outreach

Typically, participation in social protection programmes is voluntary and there is no master register of the intended population. The first activity, therefore, is to reach all potentially eligible persons and households. This involves information programmes that can be delivered through direct contact, community interaction, intermediaries or the media. The aim is to inform the intended population about the intervention, eligibility criteria, registration modality, and payment and complaints procedures and to encourage them to enrol.

Intake & registration

The next step is to register the population, either newly aware thanks to the outreach activity or actively contacted with previously held contact information or specifically compiled lists. The intake process can similarly take place in person in various locations, via community organisations, local government, intermediaries such as NGOs or communication platforms including telephone and the Internet. Registration then involves the collection of information about individuals and households, which needs to be validated and verified. The information may be complemented by existing sources such as government records and relevant databases from other programmes.

Assessment

Once the necessary information has been collected, validated and verified, profiles of the potential beneficiaries are prepared in the assessment stage. The nature of the assessment is closely tied to the features of the intended population. In demographic targeting, simple personal information can be sufficient. Economic programmes are often concerned with assessing approximate income or wealth levels to allow for means testing. For labour market programmes, the unemployed may be classed according to their skill and employability profile to allow for suitably targeted interventions. Programmes for persons with disabilities are another example in which interventions can be highly person specific and consequently require careful assessment.

Potential AI applications

- Outreach involves the targeting of potential beneficiaries with programme and registration information. To concentrate resources, targeting should be as precise as possible. **Computational advertising and social network analysis** offer targeting methods that can achieve this objective. Implementation requires data access, which could consist of social network data – almost one in six Africans were on Facebook in late 2018 according to Forbes Magazine (2018) – or mobile phone data. Social network activity is highly predictive of personal characteristics (Kosinski, Stillwell & Graepel, 2013), potentially including those suitable for programme targeting. Phone records could provide a viable alternative, as they have strong predictive power for socio-economic variables (Blumenstock, Cadamuro & On, 2015), as well as yielding social network information via communications data. Targeting via social networks would probably be premature given user numbers among potential beneficiaries, but penetration is increasing quickly. As with phone data, which already covers much of the population, data access is in private sector hands and either needs to be paid or legislated for. Data poverty poses a serious inclusion challenge in either case. In addition, the aggregation of information from several sources increases data protection risks.

- Intake & registration is the primary step for data collection. Human error and miscommunication often result in lack of data quality, a major practical issue with a big impact on targeting accuracy. AI can help to reduce errors and provisionally fill in missing information. **Outlier detection** can be used to identify unusual data points and either trigger a human audit or highlight locations at which such entries are made too frequently. For example, it might spot an error where a person has few assets, but a high recorded income. **Predictive methods for data imputation** (Bertsimas, Pawlowski & Zhuo, 2018) can fill in missing information with estimated values, allowing the beneficiary record to be used in other data-driven tasks. For instance, a labour market programme might require information on educational attainment and employment history. On the basis of past employment, it can then estimate the likely level of education.

- Assessment involves the categorisation of registrants into brackets suitable for the programme. **Clustering** can help to identify beneficiary groups that can be helpful for planning the programme roll-out. For example, the clustering of registrants for a labour market programme may reveal a considerable group of elderly work seekers in a region that warrants special adjustment of the offering there. Any available data can be used for clustering, but the choice of variables affects cluster composition. Simple algorithms such as k-means or Gaussian mixture models can be used. Domain experts need to judge whether the clusters identified are meaningful from a programme perspective and not just for statistical curiosity.

- Client communication can in the foreseeable future be facilitated by **chatbots** and **natural language processing** (NLP) systems. This would include outreach and registration for which automated systems could support functions such as initial contact and the collection of basic information. For now, providers already offer phone communication services, but these are the equivalent of an expert system with a set communication path. Corpora for low-resource languages (Tsetkov, 2017) that can provide translation services in countries with multiple languages and more flexible communication protocols that adjust to the client are not yet viable, but are on the horizon. Text-to-speech and speech-to-text technology could support illiterate users, a key consideration for ensuring universal programme access. As an example of recent developments in smartphone or tablet-embedded AI tools, NLP methods can be leveraged on low-end devices even where online connectivity is costly or missing. NLP systems are applicable to most stages of the delivery chain and can be expected to become a major application of AI in social protection given their capacity to facilitate communication with beneficiaries. Donor support for the collection of open access training data can be an important catalyst for the realisation of such potential benefits.

Enrolment

Once the eligible population have been selected from the registrants on the basis of inscription profiles, the programme beneficiaries can be enrolled. This section looks at the enrolment process.

Eligibility & enrolment decisions

The comparison of registrants' profiles with the programme eligibility criteria allows for eligibility decisions to be made. Economic assessments are often combined with demographic or other programme specific ones to produce eligibility rules. Resource constraints mean that these criteria may need to be adjusted to a sustainable level once empirical data from the population has been collected. The final enrolment decision may be based on clear rules, staff discretion or a mixture of the two, and might involve placement on a waiting list when there are capacity constraints.

Benefit/service level determination

In programmes in which different benefits or service levels are available, the next step is to select the appropriate one. As with the general enrolment decision, the setting of benefits or service levels may need to be adjusted in line with resource constraints in terms of programme capacity, both financial and administrative. The level is set according to clear rules according to beneficiary profiles. Especially in social and labour market programmes in which a categorisation via registration information is difficult, discretion in setting benefits or service levels is common.

Notification & on-boarding

The final step of the enrolment process is to inform registrants of the decision made in their case. Details of a complaints procedure may be made available to those deemed ineligible. For the eligible population, an onboarding process is required to ensure correct understanding of the programme process and any compliance requirement. Personal contact, possibly in group sessions, remote communications such as phone calls or printed materials are possible information sources. Any outstanding information needed on the programme administration side can be collected at this stage.

Potential AI applications

- Eligibility and benefit level decisions such as proxy means testing (PMT) are some of the most obvious and compelling uses of AI methods in social protection, albeit fraught with serious challenges[6]. The assignment of scores for eligibility and classes for benefit levels is a quintessential machine learning task. In particular, supervised learning methods are well suited to copying decision-making patterns from training data. The data could come from a previous manual system or have been expressly collected for the training task, with the latter preferable because it suggests higher data quality. In either case, the data needs

[6] A PMT is far from the only means test available and certainly not always most appropriate method. The forthcoming World Bank *Source book on the foundations of social protection delivery systems* (Lindert *et al.*, forthcoming) contains alternative approaches such as verified and hybrid means tests.

to be representative and current. The key question for PMT or other automated targeting methodologies is whether the outcomes are similarly fair and accurate as those arrived at through other methods – a high bar to clear.

- Current eligibility decision protocols can be augmented, rather than replaced, by AI targeting systems. Predictive methods that provide a **measure of uncertainty** about the decisions, such as Gaussian Processes, are suitable for a hybrid system in which the AI classifies obvious cases and passes less clear-cut decisions on for review. There is a risk of a false sense of assurance if the model is inaccurate. Regular review of the uncertainty estimates as well as sample cases is advisable to verify that the pre-selection works as expected.

- Continual assessment can be implemented with little cost when conducting assessment and eligibility decisions through an existing, automated data science pipeline. An update in personal characteristics could then trigger a **data-driven re-assessment**, making the programme more responsive to changes in circumstances. A potential weakness would be a situation where the prediction vacillates – think of a borderline case near the eligibility threshold of a cash transfer programme – and a minimum programme review period would likely be necessary to avoid frequent reassignment. As experience is gathered on an integrated programme approach, such as those along the productive economic inclusion lines discussed by Jorgensen and Siegel (2019), it may become possible to use predictive methods to recommend intervention packages for individuals in line with a graduation model approach.

- Demand forecasting can facilitate internal planning and increase the responsiveness of a programme to changing circumstances, allowing administrators to implement adjustments in a proactive rather than reactive fashion. Such demand forecasting can draw on **sequential prediction** methods such as recurrent neural networks, for example linking current eligiblity levels for administrative regions to wider socio-economic trends and forecasting them in line with macro-economic expectations and other forward-looking information.

Provision

The next stage is service or benefit provision. In view of the wide variety of social protection programme types, the discussion here is limited to the three major types of programmes: social assistance, social insurance and labour market programmes. Labour market programmes stand as an example of service provision.

Social assistance

Social assistance programmes principally revolve around the disbursement of funds to beneficiaries in line with an assessed benefit level. Information about financial data such as bank accounts, mobile money numbers or physical coordinates for cash disbursement are essential here. The financial system that initiates the financial transfers needs to check continued eligibility and payment status based on a reconciliation of past transactions. In cases of fees or subsidies paid directly to third parties, such as school fees or service providers, their information needs to be managed separately and linked with the beneficiaries. The monitoring of disbursement conditionality is discussed in the next stage.

Social insurance

Insurance schemes bolster early action in the face of an adverse event such as a natural disaster and speed up recovery to restore livelihoods in case a predefined outcome occurs. Claims submission and review are key aspects of any such system. Submission may take place at the insurer's premises, by post or via digital means such as an app. The task of verifying claims is often the most labour intensive and critical task in the provision function. Index insurance, an innovative approach that pays out benefits based on a pre-determined index or loss of assets and investments resulting from weather and catastrophic events (Global Index Insurance Facility, 2019), is a related task for which we give an example in the use cases in the next chapter. Many of the considerations important for conditional transfers also apply to social insurance.

Labour market programmes

Apart from unemployment benefit-type programmes, labour market programmes tend to provide idiosyncratic services, often in cooperation with third party providers in education and the private sector. The key aspects monitored here are whether the service has been provided as per the programme objectives and whether the beneficiary participated. As economic and personal conditions change, beneficiary and provider information needs to be kept current.

Potential AI applications

- In social assistance, for cases in which **eligibility estimation** is defined on the basis of data, PMT represents one approach to programme targeting. Potential proxies emerge as more data on potential beneficiaries becomes available through mobile phones, social registries and an expansion of public systems in tax administration and other data-generating areas. Data diversity implies the opportunity to proxy income through various avenues, allowing for its application to a broader audience and more accurate predictions for beneficiaries for whom multiple sources are available. However, systems that are fully or largely automated remain a distant prospect, despite recent advances, such as the use of mobile phone data for income estimation (Blumenstock *et al.*, 2015). Various machine learning models are potentially applicable and relatively simple and interpretable approaches such as logistic and regularised regression or decision trees may be preferable to black box approaches such as neural networks, given the need to audit the assignment mechanism carefully. Data protection, including which data the state can legitimately use to determine eligibility, is just one of the many questions in this promising, but still immature, field. The GiveDirectly use case outlined in chapter 3 offers a cautionary tale regarding the use of novel data sources, and current research such as Steele *et al.* (2017) suggests that targeting methods with novel data are not yet accurate enough for use at the household level.

- In social insurance, **shock prediction** is a promising avenue for AI applications. The aim is to act early to cushion the shock rather than to let it run its course and pick up the pieces after. The data and methods will depend on the particular application, which can include economic shocks, natural disasters, labour market dislocations or agricultural cycles. For example, the Red Cross use case describes how a hydrological prediction system helps to prepare a flood response before the waters break their banks. The story of the boy who cried wolf illustrates the need for fair accuracy. In addition, the setting of an appropriate false positive versus false negative balance is a key application-specific consideration.

- Another key social insurance application is **claims management,** where the aim is to make fast and accurate decisions on payouts. As with PMT, a promising approach may be to separate cases into clear, automated decision and uncertain ones for human review. Medical insurance claims are an application in which large numbers of cases need to be handled accurately and are often pre-assessed with a rule-based system. Given the large number of variables and complex treatment patterns, AI promises potential advances in accuracy for a wider range of claims, potentially reducing administrative efforts significantly. As the Pula Advisors use case from the field of (private) agricultural insurance illustrates, funds can even be disbursed automatically without the need for a claims submission. As with all critical social protection decisions, livelihoods are at stake and both beneficiaries and programme staff need confidence in the accuracy and fairness of automated systems.

- Social insurance programmes may benefit from **prediction for prevention.** In view of historical patterns, AI systems may be able to determine individuals likely to suffer from shocks that could be avoided with action at the personal or geographical level. For example, medical insurance data could be used to predict the occurrence of chronic diseases on the basis of personal characteristics and medical history, forming the basis for a programme of preventive intervention. An app-based advice programme along these lines is implemented in Republic of Korea (Joint Learning Network, 2018). Although it relies on specifically collected medical data and an expert system for giving advice, the general idea could be carried over to an AI application and might also apply to areas such as re-training for those in vulnerable employment. Of course, shock prediction in terms of probability and severity can never be 100% accurate and empirical testing over some time would be needed to verify the utility of an automated predictive system.

- Labour market programmes could benefit from **AI-assisted job matching.** As the Harambee use case points out, a good job match goes beyond suitable skills to include economic factors.

- Another labour market programme use of AI lies in the **natural language processing of job postings.** Since recruitment ads have migrated online, the scanning of job ads to extract information on the current state of the market is a promising approach to complement job seeker advice based on more accurate, but also more slowly disseminated, survey and administrative data. Similarly, automated processing of employee resumes can provide insights on skill requirements in different industries.

Monitoring & management

In line with the execution of the programme itself, interactions with beneficiaries need to be monitored and managed. Beyond monitoring the administrative aspects of the programme, there are the wider monitoring tasks for assessment of programme performance and also of AI performance within the system. The former is beyond the scope of this report, but we comment on the latter in the final section as an important element of system management. The key steps are discussed here.

Complaints

Service quality and equity demand a robust and responsive process for the handling of grievances, complaints and errors. Available and accessible communication channels, such as customer service agents, a hotline or an online redress channel, are advisable to ensure fair and equal treatment. Sources of complaints can include incorrect profile information, eligibility decisions, assigned service levels and service quality. A separate appeals process is necessary in view of the importance of social protection programmes for beneficiaries' personal lives. Structured communication with potential and current beneficiaries forms part of the grievance handling process.

Compliance monitoring

There are two aspects to compliance monitoring. The first is the compliance of beneficiaries with any programme conditionality. Some programmes with conditional cash transfers may, for example, require attendance at work, trainings or health-related actions. Information about compliance usually needs to be collected by the third party that delivers the interaction with which compliance is required, such as an education or health facility. Direct data entry rather than via intermediaries could be preferable for a fair process. The monitoring of conditionality can also be interpreted to include monitoring of continued eligibility for the programme in general or the currently assigned service level.

The compliance of programme staff is the second aspect that needs monitoring. Adherence to programme rules and regulations needs to be monitored with a view to ensuring service quality and guarding against fraud and corruption. Collusion with registrants to enter deceptive information and discretion in eligibility and service level determination are critical areas that need to be monitored.

Exit decisions & notifications

Beneficiaries can become ineligible for a programme due to a change in their circumstances or as a sanction for failing to meet programme conditionality. They have to be marked as such in the system and be informed of the decision and any appeals process.

Potential AI applications

- Participation monitoring in conditional programmes is a key ingredient for programme effectiveness and equity. The objective is to ensure that all participation is duly recorded to safeguard disbursement and that non-participation is marked for appropriate sanctioning. The most common compliance requirement is a person's physical presence at a work or education site. **Image recognition with biometric data** can provide automated monitoring, for instance, through cameras at central locations or at a registration counter. Deep learning pipelines with convolutional neural networks are especially effective (see Balaban, 2015), but careful accuracy checks are called for. Differences in accuracy rates according to skin complexion are a cause of concern. In addition, many societies are not comfortable with ‘Big Brother’ type surveillance systems.

- Complaints monitoring can be supported by analytical AI methods. The aim would be to find underlying reasons for complaints to ensure service quality and fair provision. As highlighted in the Harambee use case below, **decision tree analysis** of complaints in relation to user and service information can be used to identify factors that are associated with complaint probability. That analysis can pave the way for a review of factors that may not otherwise have been evident. As the relationships are correlations, expert knowledge is required to identify causal factors, whether from the data itself or through deeper investigation of the underlying reasons for the observed data.

- Fraud detection is an important objective in the fight against deception and corruption. **Outlier detection** can identify unusual disbursement patterns that suggest manipulation and a cluster of such outliers may be a good indicator of malfeasance. For the monitoring of decisions about eligibility and benefit level, a **staff-level predictive model** that compares the aggregate profile of staff decisions with what would be expected on average given the staff’s catchment population can reveal systematic deviations from correct practice.

- Exit prediction at the beneficiary or aggregated level allows administrative adjustments to be made in time for anticipated changes in benefit or service volumes. Like other **demand forecasts with time series methods**, it relies on aggregate information on geographic area, economic sector etc. that can be linked with external socio-economic predictions. Such a system could also be extended to the beneficiary level to allow early notification and thereby prepare beneficiaries for a likely impending end to transfers or services. As with all medium-range predictions, telling the future is an uncertain business, especially in LMICs prone to economic volatility. Once again, careful testing and robustness checks are needed if the system is to enjoy the confidence of staff and bring meaningful benefits.

3. Use cases

This section provides concrete examples of the use of AI systems and related methods in applications that are relevant to social protection. In each case, we first give an overview of the organisation and its purpose followed by a brief description of the particular tasks undertaken and method and data used. Finally, we consider the implications of the case for the wider field of social protection.

Harambee Youth Employment Accelerator

The organisation

The Harambee Youth Employment Accelerator is a South African organisation that solves youth unemployment through partnerships. It is a not-for-profit social enterprise with experience building solutions and innovations that can address the global youth unemployment challenge at scale. It works with business, government, young people and others who are committed to results. Harambee has provided 125,000 jobs and work experience placements to its network of over 600,000 youth. It employs programmes that place candidates directly with private sector employers, as well as programmes that support young people to find their own employment, both in the private and informal sectors of the economy.

Tasks, method and data

Harambee used machine learning methods for two tasks. The first was to **improve the conversion rate** of candidates invited to attend introductory workshops at a Harambee centre. Often candidates who had only completed a limited profile were invited to these workshops, however, as a data-driven organisation it tracks interactions with each candidate in the network. To optimise the response to its invitations, Harambee used historic data to predict the chances of a candidate turning up to a workshop following an invitation on the basis of a set of time-deltas and the home-to-workshop-location distance. A random forest predictive model revealed that the candidate's physical location, the time between the invitation and the workshop date, as well as the time between the last engagement with Harambee and the invitation were the most crucial factors. A three-day lead time was found to be optimal and the recentness of engagement insight was built into a new spatial invitation policy that has seen the conversion rate soar from around 2% to 25%.

The second task is job matching and goes to the core of the Harambee model. When a company approaches it with a number of jobs to be filled, the big question is who in the large pool of potential candidates has the best chances of success. Filling openings is a task with two components, in that a job must be a good match for both the business and the job seeker. A key consideration for jobseekers is the cost of transport, which can absorb a large proportion of their salary and is a major predictor of drop out. Formerly, Harambee used straight line distances to approximate travel costs, but most people travel to work via an informal taxi system without regulated routes or standardised fares for which actual costs don't correspond to straight lines. To gain better travel cost estimates, Harambee surveyed 15,000 candidates over 3 months for their actual travel time, cost and distance. This data was used to construct a model capable of predicting the cost per kilometre for taxi fares across the country with confidence bands that align to the spatial sampling of the work seekers in the survey. Representing the South African dwelling framework as 77,000 cartographic partitions holding 500 households each, they constructed an origin-destination matrix using a routing algorithm traversing the actual road network optimised to emulate the informal taxi network of South Africa. The origin-destination matrix is deployed on scalable cloud infrastructure, which allows for **optimised real-time matching** of candidates to opportunities as well as real-time employer demand planning simulations at a national scale with a fine degree of spatial accuracy. This methodology aims to maximise candidate take-home pay, particularly in low earning opportunities.

Challenges and mitigation

Although Harambee collects a lot of information and could build an AI-based automated system on that basis, so far it has opted for a more limited use of the technology. In the first example of improving the conversion rate of workshop invitations, the analysts used a machine learning method to extract highly predictive features in a way that would have been difficult with traditional statistical analysis using linear methods. However, putting a premium on transparency and interpretability, it then built a heuristic system that only draws on the most predictive and well understood factors of the model. As such, machine learning was used as an analytics tool, rather than as a direct decision mechanism. One factor that has constrained the development of fully automated systems in Harambee is the availability of skilled staff. Although the South African labour market can supply people with the skills for model development, the whole life cycle of an AI project is more hard graft than statistical wizardry. Recruiting technicians willing and able to build familiarity with the data and process it into a useful resource for AI systems has been difficult – a somewhat ironic constraint given their expertise, but one that underlines the multifaceted nature of building well-functioning AI systems.

Relevance for social protection programmes

Outreach and registration are key aspects of social protection programmes. Data-driven optimisation of the contact modalities with potential beneficiaries can substantially reduce administrative costs by raising capacity utilisation at contact points and raising the effective coverage of programmes via increased uptake. Similarly, the development of a data-driven algorithm that takes detailed local data into account can also result in higher programme uptake and increased wages net of costs for beneficiaries, amplifying the effectiveness of the programme. This use case illustrates that it is not necessary to deploy a fully automated system from the start. Instead, the gradual adoption of methods to support analytics and enhance particular service aspects can build experience with, and confidence in, the technology as a stepping stone towards a fully integrated AI system. Finally, if suitable staff are a bottleneck even for a specialist organisation with a data-driven culture, a gradual approach may be advisable in settings where skills come at a premium.

Cash transfers at Give Directly

The organisation

GiveDirectly is a charity that gives unconditional cash transfers to very poor households, originally in Uganda, Kenya, Rwanda, the Democratic Republic of the Congo, Morocco, Malawi and Liberia. Its name comes from the idea that people living in poverty know best how to improve their own lives and that the most effective way of supporting them is to provide means in the least costly and constrained way possible – in other words, give them directly what they need. GiveDirectly only operates in parts of East Africa because mobile money reaches the most remote communities there and provides a low-cost and relatively secure way of making transfers. Their most common model is for eligible households to receive around USD 1,000 over several instalments, representing roughly a year's income for an average family.

Tasks, method and data

The most critical aspect of cash transfer programmes is the selection of beneficiaries. Simply put, those who are eligible should be registered and those who are not eligible should not. For most cash transfer programmes, household income is the eligibility criterion and GiveDirectly aims to support very poor households. When the characteristics that determine eligibility are not directly measurable, proxy means testing is often deployed. Household assets are a good proxy of income and a dwelling typically represents a household's single most valuable asset. GiveDirectly used satellite imagery to identify households that live in houses with a thatched roof, as this gives a strong clue as to the value of the house. Richer households tend to live in brick houses with corrugated tin or tiled roofs. Although algorithms exist to perform this task[7], GiveDirectly uses crowd-

[7] For example, see Helber *et al.* (2019) for cutting edge research in identifying informal settlements.

sourced labels of satellite images. Such labelled data is then used to train a predictive machine learning model, but in this case the automation step turned out to be an unnecessary overhead given the relatively limited scale of the task and the factors discussed below.

Challenges and mitigation

Proxy means testing is an inexact science that usually leads to fairly high rates of incorrect eligibility decisions. On the other hand, traditional surveys and other methods also lead to wrong decisions and, ultimately, it is a cost-benefit question. GiveDirectly had a low tolerance for wrong decisions, forcing them to use a number of other approaches in addition. Over time, they found that satellite image-based identification gave an excessive false positive rate with few true positives and required various other verification mechanisms. After several years of experimentation, it became clear that the cost of verification outweighed the savings that remote sensing provided in the first place. Although promising in principle, the AI method of automated beneficiary identification did not pass an extended cost-benefit test and was abandoned in favour of a more traditional targeting approach. GiveDirectly now selects entire villages according to publicly available data, side-stepping the estimation of differential income levels between neighbours.

Relevance for social protection programmes

Remote sensing data is a primary candidate for proxy means testing and GiveDirectly's work represents an early attempt in this direction. Although it did not use machine learning methods, it did produce the data that would have been fed into an AI system. In this case, it did not pass a cost-benefit test as it was too unreliable an indicator and did not yield enough useful information to serve as a viable proxy. This result is especially striking because GiveDirectly was able to select its coverage area, yet was still unable to use the data effectively. Government programmes generally have to operate over the whole territory of a country, or at least entire regions, and, therefore, require robust indicators that either work across the board or that can be supplemented effectively by mitigation of exclusionary and misclassification effects to prevent at least some of the population from suffering from systematic exclusion.

Agricultural insurance at Pula Advisors

The organisation

Pula Advisors is a Kenya-based agricultural insurance specialist. The company aims to provide protection against weather, pest and disease shocks to farmers in LMICs, with a special focus on smallholders. Pula partners with the private and public sector, bundling its insurance with agricultural inputs like seeds, fertiliser and credit products. They currently cover 12 cash and food crops, work in 11 countries and enrolled 812,000 farmers in 2018. The two main insurance products are short-term rainfall insurance (if it doesn't rain for three weeks after planting, farmers get a bag of replacement seed) and yield index insurance. The latter provides farmers with a payout if the yield of their crop in their local area falls below a predetermined threshold. Both insurance products do not require estimation of a farmer's actual germination success (for rainfall insurance) or actual harvest yield (for area index insurance), as these are estimated for geographic areas. Farmers register their participation at local agro-dealers and are contacted by mobile phone in case of payout, as well as to receive tailored advice on agricultural practices.

Tasks, method and data

Rainfall data is collected by satellites that estimate the amount of precipitation for relatively small areas to which farmers can be matched. A more challenging task is to estimate the yield for a crop in a given area for the yield index insurance. Administrative zones such as districts do not align well with the agro-economic zones in which farmers face similar growing conditions. A key task in providing the insurance product is to split the coverage area into zones that are big enough to allow for economical yield estimation (a number of fields are harvested to measure yield in each zone, which is expensive) and small enough to give an accurate view of growing conditions. To determine these zones, Pula uses an unsupervised machine learning clustering algorithm that splits a map into clusters with similar characteristics, in particular historical rainfall patterns. Yields within these zones are strongly correlated, allowing the fair measure of agronomic conditions and triggering appropriate payouts to farmers who face drought or other yield shocks.

Challenges and mitigation

In a programme in Nigeria in partnership with CGAP (CGAP, 2018), Pula tried to estimate farm-level yields with the help of remote sensing data such as rainfall, soil light reflection and temperature from various sources. If the predictions were accurate enough, it would allow Pula to insure farmers individually rather than by area, providing tailor-made protection against adverse growing conditions and thus protecting them more effectively from income shocks. Unfortunately, at most, 20% of the variation in crop yields could be explained by predictive models using the remote sensing data. What the analysis did show was that the variation in average yields across local government districts could be predicted with 60–80% accuracy. This level of predictive power is still not accurate enough for actuarial purposes, but it prompted Pula to invest more effort in area-specific yield estimation, ultimately coming up with the current system of agro-economic zones outlined above. Another issue was that although most farmers have small plots and highly variable yield, the data available was mainly for larger farms that typically have higher and more stable yields. To balance its evidence base, Pula had to send out teams to collect 2,500 further yield measurements, with a bias towards small farmers.

Relevance for social protection programmes

The bulk of the rural population in LMICs depends on farming for their livelihoods. Often, lack of savings and welfare nets, climate shocks and pests represent the biggest risks to the wellbeing of vulnerable communities. Crop insurance, therefore, represents a protection mechanism that can help to protect the economically vulnerable from hardship and hunger. As social protection is increasingly coordinated with wider risk management programmes along adaptive social protection lines (Jorgensen & Siegel, 2019), the provision of index insurance can form a key part of response systems. Pula's experience also holds lessons for programmes experimenting with remote sensing for social protection programmes. The lack of predictive power of coarse grained remotely measured indicators underlines the need for the careful testing of systems to verify the extent to which individual outcomes can really be predicted this way. Nonetheless, as with area index insurance, great value may lie in monitoring conditions for larger groups or areas via remote sensing, e.g. to tailor service levels. Finally, this example of the need for detailed data collection at the household level to counteract unrepresentative data shows once more that a data-driven approach relies not only on data quality, but also on representativeness.

Forecast-based financing at the Red Cross in Togo

The organisation

The International Committee of the Red Cross and Red Crescent is a global humanitarian aid network. It consists of 191 national chapters that assist people at risk from natural or manmade disasters, conflict, ill health and social problems. The main natural risks in Togo are weather related, namely drought and flooding. The latter triggers the most acute need for humanitarian action and is the subject of this use case.

Tasks, method and data

In humanitarian aid, even foreseeable disasters typically occur before funding is made available for their mitigation. The funding delay causes delays in humanitarian action, which can lead to tragedy. Forecast-based financing is a humanitarian aid modality that provides pre-allocated funding for foreseeable disasters. The availability of funding allows for response preparation ahead of a disaster. When a disaster is expected, it is then possible to take early action to mitigate its impact. Of course, the effectiveness of forecast-based financing depends on the accuracy of disaster prediction.

The Mono river runs through much of the length of Togo and a large population is regularly affected by the seasonal flooding of its plains. The Nangbeto dam is a binational dam of Togo and Benin on the border of the two countries. Its inflow and outflow are key predictors of downstream flooding. Other important measures for flood prediction include rainfall estimates from upstream communities provided via SMS. On the basis of these inputs, the FUNES model has been developed, a self-learning algorithm that predicts downstream river flow for the current and following two days. Suarez and Mendler de Suarez (forthcoming) describe the context, methods and intervention. Previously, the Red Cross would take action once flooding occurred, but in combination with forecast-based financing they are now able to use the model's forecast to start relief operations such as provision of water purification tablets and waterproof bags when a flood is expected.

Challenges and mitigation

Being based on limited data and a mixture of methods, the predictive power of the FUNES model was limited. In fact, the staff coordinating the system noticed that they could predict rising waters from a few data sources before the predictive model rang the alarm. In other words, humans proved better forecasters than this particular algorithmic implementation. At least two interesting lessons can be drawn from this experience. The first is that AI systems need to be developed and updated as the evidence base and understanding of the system that they model grows. If human intuition can extract more information from many disparate data points than an algorithm, then the algorithm can probably be optimised. The second point is that the construction of the data-driven system led to a culture change and the evolution of operating procedures. The Red Cross in Togo now sees pre-disaster prevention as a key task and the technological system not only supports implementation, but also catalysed an evolution in organisational perspective.

Relevance for social protection programmes

Although forecast-based financing with predictive AI methods may be seen as a humanitarian aid modality, recent social management approaches underline the importance of a comprehensive approach to livelihood protection. There is similarity between social insurance and disaster mitigation in that both arise from a shock to individuals. The prediction of such shocks before they occur can help beneficiaries to build resilience and preparedness, reducing the adverse impact. Non-communicable diseases in the context of health insurance are another example. At the programme level, predictions of the overall level of demand may be useful in managing administrative processes and resources. Finally, the Togolese Red Cross embraced the opportunity for progressive organisational change that a data-driven, empirically-based approach to programme management offered.

4. Suggestions for practitioners

Few social protection practitioners are familiar with the technical details of computer systems, machine learning, computational statistics or digital databases. Even though an increasing part of projects and programmes in the sector will involve aspects such as digitisation spreads, not everyone wants or needs to be a techie. What practitioners do need to do is to pay special attention to the pitfalls that AI systems might hide. This section presents actionable suggestions that those planning or evaluating an AI-based project or programme can work through. Not every suggestion will be relevant to every project, but when the applicable ones are taken into account, the chances of the promise of AI being realised will be better.

Audit data protection

Verify that data is stored and shared responsibly.

The issue

Data is the basis of AI systems, but apart from serving as an input to a process, personal information collected for social protection programmes is also of the highest sensitivity. Leakage of someone's details on income, assets, or health and employment status can have serious social, economic and even security consequences for the person and their family. The secure storage of such data is an essential duty of any organisation working with it. Where data is shared with a third party, such as a service provider to which part of the AI is deployed, data protection needs to extend to the third party and protocols for responsible data sharing should form part of standards.

Mitigation strategies

Data security starts with the internal procedures of organisations storing information. Limiting access to those individuals who require specific data items lowers the risk of data loss, theft and manipulation. Data access policies should specify who has access to each item on the database. Aspects of the policy should include password-protected storage, the segregation of users into groups with varying rights, data access and modification logs, and clear staff guidelines that detail both operational procedures and expectations around responsible data use. Other dimensions of secure storage include such diverse areas as password policy, software maintenance, network administration, cryptography, device management, and cloud service restrictions (Dix et al, 2018). Financial Conduct Authority (2008) is an example of regulatory guidance on appropriate practice.

Practical implementation

Given the fast pace of change and national differences in this sphere, detailed recommendations that hold across countries are unrealistic. Instead we recommend the production of a country-specific checklist that verifies data protection in line with current national standards. Where such standards have not been defined or where they are insufficiently precise, alternative ones of a suitable level should be selected and applied instead. Although initial verification is essential, the duty to safeguard security standards persists as long as the information is stored and verification of continued alignment should be made part of regular programme activities.

Insist on openness, transparency and sustainability

Secrecy, complexity and exclusivity breed mistakes and are largely unnecessary.

The issue

AI systems are a combination of statistical methods and software design, and AI experts often have a statistical or mathematical background rather than one in software development. Their ad hoc code can be difficult to understand and maintain, regardless of whether it currently works for the problem at hand. Knowing this, they may be unwilling to share the code. The field's emphasis on superior performance and the skeletons often buried in practical implementation code can further reduce the willingness to share it openly. Last, but not least, clear and concise systems are the result of hard work. As in writing text, only careful editing or 'refactoring' leads to clear code; it is faster and easier to write many complex lines than a few clearly structured and transparent ones. Such complexity hides mistakes from the software developer herself and also from those working with the code in future. The production of well-structured code requires significant time and effort, making it costly in the short term, although it is necessary for long-term success.

Mitigation strategies

An important insight in software engineering is that code is read much more often than it is written. Developers change, others build new components and it is, therefore, important to have clear, repeatable and well-documented code. Given their importance for many people's lives, AI systems in social protection need to be transparent so that they can be tested, audited and updated on a regular basis. Although model details often should not be shared publicly to avoid gaming behaviour, fraud and deception, developers should invest time and effort in meeting standards of open, replicable data science for internal users. A related point is that the sustainability of the AI system depends not only on initial design, but also on continuous adaptation. The future availability of human resources and deployment infrastructure are other conditions for sustainability.

Practical implementation

Workflow tools such as Kedro and Read the Docs[8] can help to build a reproducible data science pipeline. Quality control methods such as code review by trusted programmers, ideally with experience in similar efforts, can help to identify issues early before they affect the operation of the social protection programme. At the time of writing, machine learning expertise is still in relatively short supply outside of Silicone Valley in relation to the scale of demand, but local capacity in LMICs is growing fast. In the likely case that an AI system design is fully or partly designed by experts unlikely to be involved long term, a strategy to transition to local staff and the requisite capacity development can become part of the programme plan. Similarly, the deployment infrastructure for system recalibration and daily operation should be designed with a view to local implementation.

Kick the tyres and check the engine

AI runs on data – call it the tyres – and its machine learning is the system's engine.

The issue

Machine learning methods are powerful pattern recognition tools that will find a relationship when let loose unchecked – even if the relationship is spurious or unreliable, holding no predictive power beyond the dataset that they were trained with. Such 'overfitting' is perhaps the biggest risk to originate from the statistical underbelly of modern AI systems. Much could be said about model design etc., but we must limit ourselves to a few tangible subjects. The adequacy of the data on which the model is trained should also be verified to ensure that operational systems are built on a solid basis. Despite the 'intelligence' in AI, such systems are incapable of drawing conclusions that they have not seen direct evidence for. To give AI the chance of performing a certain task, we must give it all the information required, as it cannot draw conclusions beyond that.

Mitigation strategies and implementation

The best way to test the statistical effectiveness of AI is to compare it to human judgement, the process that it is meant to replace. We take up this point again below

[8] A combined illustration of Read the Docs and Kedro can be found at kedro.readthedocs.io/en/latest/.

when discussing discrimination, as it is worth saying twice: The best way to test the predictions of an AI system is to collect fresh data – say applicant profiles or insurance claims – and to note down the appropriate decision for each case. Apart from borderline cases where disagreement may well happen, the human and automatic decisions should mostly match. A related point is that no amount of technical wizardry can extract insight from data that is not contained in it. AI cannot do magic, so a close look and sense check is called for in relation to whether the predictive capability of a system is credible in view of the data information content. The data must be representative of the population, so that the relationships capture all relevant groups and valid predictions are possible for any reasonably likely case that arises in a social protection programme. Finally, it could be a good idea to request a description from the model designer of how overfitting was avoided and what typical failure cases in model training look like.

Take an ethical test drive

Automated systems don't know right from wrong, but we can test if they act as if they do.

The issue

The learning algorithms that power AI systems replicate the patterns that they find in the training data given to them. If the record of past decisions includes discriminatory patterns then these patterns will be learnt and replicated. For example, assume that an algorithm determines eligibility for a school feeding programme and that there is a small ethnic group in a single district that was discriminated against in the past due to cultural prejudice. Even if the ethnic identity is not captured in the data, but membership is associated with measured attributes such as profession or household location, a machine learning-based automation of the process of determining eligibility may internalise the past bias against this group and discriminate against them in future decisions.

Mitigation strategies

Some machine learning methods provide relatively transparent decision factors, but the more powerful ones that are able to capture complex patterns and achieve strong performance on complicated tasks are inscrutable to the human mind. One strategy is to insist on methods that yield interpretable decisions, but this may rule out many useful applications. An alternative is to focus on the decisions that an algorithm makes rather than its internal process. The decision of an algorithm should make sense and be free from bias and discrimination. An elementary way[9] of verifying this is to produce synthetic inputs that vary only in dimensions that should not be the basis of a decision, such as ethnicity or gender. It is evidence of bias or discrimination if the output is different for otherwise identical inputs. Although it is not possible to cover all eventualities, a dataset of such test cases can be produced to verify that the decision-making process is reasonably sound and fair.

Practical implementation

The best way of procuring such data is to collect it in person in a similar way to how it would be collected for a social protection programme, including examples all identifiable groups vulnerable to discrimination. Care needs to be taken to select test cases, households or beneficiaries that are not present in the data used to train the algorithm. On the basis of such real data, further artificial data points can be created that are the same except for characteristics that should not be taken into account. Eligibility decisions should then be made by trusted experts and compared to the algorithm outputs. Systematic differences point to a problem that needs to be addressed. Finally, the predictive accuracy of machine learning components in the AI system should be similar across groups to ensure all groups can enjoy the benefit of AI.

[9] Our understanding of algorithmic fairness is evolving quickly and current, situation-specific expert guidance should be sought if discrimination is a meaningful risk, as is often the case.

Pilot first

Practical experience beats up-front statistical testing, so run a pilot to make sure it all works.

The issue

Most statistical methods exploit correlations between variables and the machine learning algorithms that underlie most AI systems are no exception. The structure of actual causation may differ significantly from that implicitly inferred by the model, which is problematic if the causation and correlation drift apart over time or due to programme intervention. It may also be that the data collected to train the system is not representative of the population or that it was biased or otherwise flawed. Again, no amount of work with flawed data that is unrepresentative of the real world setting can produce a robust AI system.

Mitigation strategies

The only way to determine whether an AI system that performs well on paper will be able to replicate its success in real life is to test it. There are so many ways in which the data could be unrepresentative and so many ways in which the world could have changed since it was collected that a practical application is the only dependable way to expose issues. Similarly, the high-powered statistical tools of AI may have been used incorrectly or have fallen victim to artefacts in the data and such mishaps are best confronted by hard facts.

Practical implementation

Before rolling out an extensive programme with automated processes, each of them ought to be tested in a small-scale trial. The results should be compared with the predictions and any meaningful discrepancy has to be addressed if it is to function at scale. Small-scale trials are advisable before large-scale rollouts. Even if a trial goes well, this may be down to chance and further trials should be undertaken in different settings. The experience gained should not just validate the theory, but can also give important clues to the appropriate management of the eventual scale-up. The cost and time requirements of extensive testing may be high, but the costs of large-scale failure would probably be much higher.

Mitigate exclusion

Cater to the marginalised to stop data poverty from compounding exclusion.

The issue

Data-driven processes may work well for those who are included in the system, but cannot function for those whose data is not captured. People who were never registered in existing systems or who are not registered when new technology-based systems are implemented are likely to be among the most marginalised members of society in economic and social terms. Geographical distance or travel time to the nearest urban centre and lack of access to media or social networks that disseminate information on economic opportunities such as social programmes are both more likely to constrain those on lower incomes and those who generally have less access to social services and programmes. In other words, some of the most deserving of social protection programmes are also most likely to be excluded from access to them due to a lack of awareness or ability to register. Economic poverty is closely related to data poverty. One hypothetical example is poverty estimation via mobile phone data. Although research has shown remarkable accuracy in this endeavour, only households that own and use a phone are captured this way, while the poorest who do not own a phone are systematically excluded.

Mitigation strategies

The movement towards universal coverage for those eligible for social programmes implies a need for universal data collection for 'data-based' programmes. Although complete coverage is unrealistic, programmes can follow a mix of strategies to mitigate the effects of digital exclusion. The first is to run additional data collection efforts in marginalised communities to overcome hurdles in the regular system. Such an effort could include efforts to subsidise data collection methods, for instance, by supporting the use of mobile phones, to keep with the above example. However, the effectiveness of extending data coverage via additional outreach may be limited and a parallel, non-data-driven system of provision can be required to reach the data poor.

Practical implementation

Distributional effects deserve special consideration in the design of social protection programmes. The identification of groups that will not be covered comprehensively by a data-driven system deserves centre stage in programme design, rather than being an afterthought. There is probably no general rule as to how to best achieve inclusion, but the issue and its cost and organisational implications need to be recognised as an integral part of programme design. The same principles of running pilot programmes, data protection (see above) or auditing complaints (see below) apply, but need to be adjusted to the specific circumstances of the digitally excluded.

Encourage independent audits

Third parties can protect automated systems from becoming tools for repression or exclusion.

The issue

Automated decision-making removes human contact, lessens scope for common sense discretionary interventions and makes it harder to gather information or challenge decisions. The catchphrase 'computer says no' exemplifies the apparent arbitrariness of faceless decision-making systems and obscures the fact that these decisions are ultimately due to human choices. The technical nature of IT system design erects further barriers that may keep both beneficiaries and programme staff from challenging unfair decisions. The underlying reason for unfairness may include inadequate data, or algorithms, or issues of implementation. Importantly, they may also reflect purposeful adjustment of the system to act in a certain way, but hidden behind the supposed objectivity of AI. Take as an example a public health insurance system that is under financial pressure. Managers may be under pressure to reduce costs by rejecting more claims and could target opposition-voting areas by biasing algorithms to take such geographical data into account. With the editing of a few lines of code, automated systems can become tools of accidental or planned discrimination and repression. Eubanks (2018) documents how algorithmic decisions can be instrumentalised politically and result in inequitable outcomes in the USA, an experience that holds stark lessons for the rest of the world.

Mitigation strategies

Transparency and independent review may be the best approaches to guarding against systemic errors and abuse. Two components can help to expose such issues. The first is the regular and systematic use of data analytics to spot unfair treatment. The other is the review and analysis of customer complaints and appeals, as these may reveal problems as they arise and grow more apparent in line with the issue's severity.

Practical implementation

Internal control systems are essential, but independent review by a third party may also be appropriate and effective in providing impartial monitoring. Academic or public interest institutes may be good candidates to conduct regular reviews of service levels, rejections, complaints and appeals. Both overall analyses by beneficiary characteristics and random auditing of individual cases may be necessary to guarantee an overall level of fairness and to promote the system's inclusive rather than repressive and exclusionary potential.

Develop with the developers

Close involvement of software and statistical model developers at each stage is a key to success.

The issue

In many organisations, the IT and statistics departments are supporting functions separate from the service delivery functions. Although there may well be some overlap, a traditional split is probably the norm even though all of these stakeholder are important to implement an AI project. In the design and implementation of the project, split organisational units often result in somewhat independent work streams that run in parallel, rather than with close integration in planning and execution. When the streams intersect, the difference in perspectives on the task and the different interpretations of specific goals can cause misunderstandings, delays and even conflicts that ultimately hamper progress.

Mitigation strategies

Different domains speak different languages, but coworkers need to understand each other if they are to cooperate effectively. Convergence on a shared understanding of the task at hand is best achieved through early and sustained interaction that is built into the project structure. The objective of the project can be mapped onto data collection and analysis and its automated implementation onto the organisation's IT system. The attempt to map early will likely highlight issues that lead to the redesign of some aspects and holds the key for a more integrated approach with achievable milestones.

Practical implementation

Data, statistics, IT and domain experts can conceptualise the project together. Each group can consider what porting the idea to their field means in practice and integrate the results with those of the others. Only once a joint plan has been agreed should the project proceed to the next stage, as it goes from planning to implementation to monitoring and evaluation. Although laborious at first, a joint approach is most likely to ultimately run smoother and faster, and to deliver better results.

Is it worth it?

Consider costs versus benefits and risks versus rewards.

The issue

The saying goes that economists know the price of everything, but the value of nothing. We might similarly expect AI experts to get carried away with the applications of their technology, even if they are not always best placed to judge the economic suitability of their work. This report has discussed a significant number of issues that AI programmes in social protection must face, such as the custody needs of highly-sensitive data, strict accuracy standards due to the criticality of decisions or potential exclusion and attendant needs for mitigating action. In view of these hurdles, one key question is whether the benefits outweigh the costs. It may be that a traditional approach is more cost-effective and that the improvement that AI offers is simply not worth the transition to a new way of doing things. In a similar vein, the risks may outweigh the opportunities. For instance, the collection of sensitive information about vulnerable groups might expose them to discrimination or persecution if the data falls into the wrong hands and a small gain in programme efficiency may not justify the benefits to the programme.

Mitigation strategies

The costs and benefits of a programme are always uncertain in the planning stage, but it should be possible to estimate them as part of the overall project plan. Cost-benefit analysis is a well-developed field, albeit one that requires significant expertise to be done properly. Nonetheless, a basic use of the framework should be feasible without disproportionate effort. One useful source is Dhaliwal *et al.* (2012), which provides a framework for considering intervention cost effectiveness based on empirical impact estimates. The use of data-driven methods offers an opportunity to produce detailed estimates at the beneficiary level, as information about characteristics should be available in digital form. The weighing of risk and opportunity will often be a more subjective exercise for which quantitative measurement standards and conventions are unavailable, but it is no less important and consequently the analysis should be clearly documented.

Practical implementation

Pilot studies are the most reliable and direct way of collecting information on both costs and benefits. They should provide a fair idea of the intervention's effectiveness, perhaps in direct comparison to the previous approach. The system development necessary for piloting should already give a good indication of data collection and system deployment costs. Importantly, a wide-ranging pilot that covers various groups may yield information on the need for the mitigation of exclusion, which is a potentially costly aspect of a universal coverage social protection programme with an AI component. Pilots should also be instructive for highlighting risks and opportunities. Information can be supplemented with qualitative data collection in the form of structured interviews that explore potential side effects. Special emphasis can be given to vulnerable populations and the solicitation of previously unconsidered risks and opportunities from interview partners. Beneficiaries, programme administrators and service providers, potentially excluded populations, domain and IT experts, data protection specialists, local and central governments, politicians and relevant business interests may all have important insights to offer in weighing risks and rewards.

Leverage AI for change

The digital foundations of AI can propel wider reform.

The issue

Machine learning and AI represent the cutting edge of scientific and industrial development. Their implementation in social protection programmes is to be welcomed as far as they meet their cost-benefit and risk-reward tests, but in many cases the use of highly sophisticated methods may be overkill. More basic efforts such as the development of systematic data collection, the integration of existing databases, the digitisation of paper-based administrative processes, the implementation of rule-based or expert systems that work without much statistical background or the production of relatively simple data analytics to identify areas for improvement might yield much higher benefits. In essence, there is a risk that the pursuit of high-flying objectives distracts from such easier, cheaper and more effective interventions. At the same time, the promise of major progress with new technology, potential enthusiasm for AI among stakeholders, and the availability of financial support are valuable resources that can be turned into better programmes.

Mitigation strategies

The foregoing brief list of possible areas for improvement includes many that are actually preconditions for the installation of an AI system. For example, we have seen that comprehensive and reliable data that covers the bulk of the eligible population is the basic input for AI systems. Similarly, basic analytics are necessary background research during the design stage of an AI system. In that light, the implementation of an AI system requires a lot of investment in more basic supporting functions that are of benefit in themselves. It may well be that the AI aspect is 'the cherry on top' of a wider push for digitisation and system reform. The operational, institutional and cultural changes involved in a significant upgrade of delivery processes can pose a major hurdle to reform. Uniting the stakeholders involved in this effort around the goal of installing a first rate AI system may ease the process of change and even lead to wider progress beyond the narrow scope of the system in question.

Practical implementation

Organisational change and systems modernisation are invariably context-specific. One implementation strategy for AI proposals is to go from the top down. Where an AI proposal could support an aspect of a social protection programme, but lacks systems support, planning the construction of the necessary supporting infrastructure can become an integral part of the AI project. The other is bottom up in that there may be a recognised need to improve processes and systems. In that case, an AI application that could be built on top of the reformed system could be identified and serve as the ultimate focal point of the upgrading process.

References

Asian Development Bank, *Identify for development in Asia and the Pacific,* 2016, https://www.adb.org/sites/default/files/publication/211556/identity-development-asia-pacific.pdf (accessed 28 October 2019).

Balaban, Stephen, 'Deep learning and face recognition: The state of the art', *Biometric and Surveillance Technology for Human and Activity Identification XII,* Vol. 9457, International Society for Optics and Photonics, 2015, https://arxiv.org/pdf/1902.03524.pdf (accessed 15 September 2019).

Barca, Valentina and Makin, Paul, *Integrating digital identity into social protection – an analysis of potential benefits and risks,* 7 May 2018, unpublished.

Bertsimas, Dimitris, Pawlowski, Colin and Zhuo, Ying Daisy, 'From predictive methods to missing data imputation: an optimization approach', *The Journal of Machine Learning Research,* 2018, 18(1): 7133–7171, http://www.jmlr.org/papers/volume18/17-073/17-073.pdf (accessed 12 December 2019).

Blumenstock, Joshua, Cadamuro, Gabriel and On, Robert, 'Predicting poverty and wealth from mobile phone metadata', *Science,* 2015, 350(6264): 1073–1076, http://jblumenstock.com/files/papers/jblumenstock_2015_science.pdf (accessed 15 September 2019).

Bundesministerium für Bildung und Forschung, *Strategie Künstliche Intelligenz der Bundesregierung,* 2018, www.ki-strategie-deutschland.de/home.html?file=files/downloads/Nationale_KI-Strategie.pdf (accessed 15 September 2019).

CGAP, 'Using satellite data to scale smallholder agricultural insurance', *Consultative Group to Assist the Poor (CGAP) Brief,* August 2018, www.cgap.org/sites/default/files/researches/documents/Brief-Using-Satellite-Data-Smallholder-Agricultural-Insurance-Aug-2018.pdf (accessed 15 September 2019).

Chirchir, Richard and Barca, Valentina, *Building an integrated and digital social protection information system,* GIZ Technical Paper, January 2020.

Dhaliwal, Iqbal, Duflo, Esther, Glennerster, Rachel and Tulloch, Caitlin, 'Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education', *Education Policy in Developing Countries,* 2012, pp. 285–338.

Dix *et al.*, *Responsible use of personal data and automated decision-making in financial services,* GIZ Paper, 2018.

Eubanks, Virginia, *Automating inequality: How high-tech tools profile, police, and punish the poor,* St. Martin's Press, 2018.

Financial Conduct Authority, *Data security in financial services,* London, UK, April 2008, www.fca.org.uk/publication/archive/fsa-data-security.pdf (accessed 15 September 2019).

Forbes Magazine, *Almost all of Facebook's 139 million users in Africa are on mobile,* 18 December 2018, www.forbes.com/sites/tobyshapshak/2018/12/18/almost-all-of-facebooks-139m-users-in-africa-are-on-mobile/ (accessed 15 September 2019).

Gartner, *5 Trends emerge in the Gartner hype cycle for emerging technologies,* 2018, 16 August 2018, www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/ (accessed 15 September 2019).

Global Index Insurance Facility, *What is Index Insurance?,* 2020, https://www.indexinsuranceforum.org/faq/what-index-insurance (accessed 26 February 2020).

Grother, Patrick, Ngan, Mei and Hanaoka, Kayee, *Ongoing face recognition vendor test (FVRT)*, National Institute of Standards and Technology, US Department of Commerce, 5 July 2019, www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing (accessed 15 September 2019).

Helber, Patrick, Gram-Hansen, Bradley, Varatharajan, Indhu, Azam, Faiza, Coca-Castro, Alejandro, Kopackova, Veronika and Bilinski, Piotr, *Mapping informal settlements in developing countries with multi-resolution, multispectral data,* Paper presented at the ICLR AI for Social Good Workshop 2019, http://aiforsocialgood.github.io/iclr2019/accepted/track1/pdfs/49_aisg_iclr2019.pdf (accessed 11 December 2019).

Joint Learning Network, *Using health data to improve universal health coverage: Three case studies,* 2018, www.jointlearningnetwork.org/resources/using-health-data-case-studies (accessed 15 September 2019).

Jorgensen, Steen Lau and Siegel, Paul B., *Social protection in an era of increasing uncertainty and disruption,* World Bank, May 2019.

Kosinski, Michal, Stillwell, David and Graepel, Thore, 'Private traits and attributes are predictable from digital records of human behavior', *Proceedings of the National Academy of Sciences,* 2013, 110(15): 5802–5805, https://www.pnas.org/content/pnas/110/15/5802.full.pdf (accessed 11 December 2019).

Kurzweil, Ray, *The age of intelligent machines,* MIT Press, Cambridge MA, 1990.

Leite, Phillippe, Karippacheril, Tina George, Sun, Changqing, Jones, Theresa and Lindert, Kathy, *Social registries for social assistance and beyond: A guidance note and assessment tool,* World Bank, 2017, https://openknowledge.worldbank.org/bitstream/handle/10986/28284/117971-REVISED-PUBLIC-Discussion-paper-1704.pdf?sequence=1 (accessed 10 December 2019).

Lindert, Kathy, Karippacheril, Tina George, Rodriguez Caillava, Ines and Nishikawa Chavez, Kenichi, *Sourcebook on the foundations of social protection delivery systems,* World Bank, forthcoming.

OpenAI, *Better language models and their implications*, 14 February 2019, https://openai.com/blog/better-language-models/ (accessed 15 September 2019).

Pahlevi, Said Mirza, *Indonesia's unified database,* Presentation at the ADB Social Protection Week 2019, unpublished.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean and Huang, Zhiheng, *et al.,* 'ImageNet large scale visual recognition challenge', *International Journal of Computer Vision,* 2015, 115(3): 211–252.

Russell, Stuart J. and Norvig, Peter, *Artificial intelligence: A modern approach,* Pearson Education Limited, Malaysia, 2016, http://aima.cs.berkeley.edu/ (accessed 15 December 2019).

Sepulveda Carmona, Magdalena, *Is biometric technology in social protection programmes illegal or arbitrary? An analysis of privacy and data protection,* ESS Working Paper 59, International Labour Office, Geneva, 2018.

Steele, Jessica E., Sundsøy, Pål Roe, Pezzulo, Carla, Alegana, Victor A., Bird, Tomas J., Blumenstock, Joshua and Bjelland, Johannes, *et al.*, 'Mapping poverty using mobile phone and satellite data', *Journal of The Royal Society Interface 14,* No. 127, 2017.

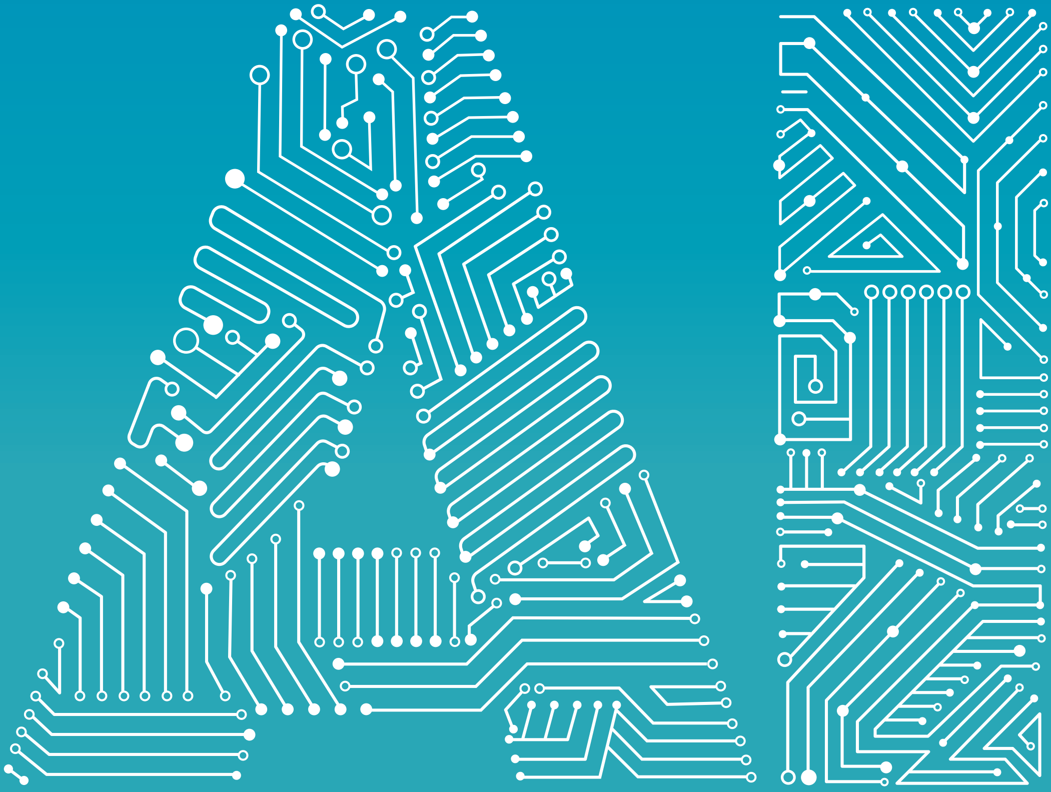
Suarez, Pablo and Mendler de Suarez, Janot, *Linking machine learning and rapid financing for flood preparedness: Red Cross innovations to manage changing climate risks in Togo,* forthcoming.

Tsetkov, Yulia, *Opportunities and challenges in working with low-resource languages,* Carnegie Mellon University, 2017, https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf (accessed 15 September 2019).

USAID, *Reflecting the past, shaping the future: Making AI work for international development*, September 2018, www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf (accessed 15 September 2019).

World Bank, *Public sector savings and revenue from identification systems: Opportunities and constraints,* 2018, http://documents.worldbank.org/curated/en/745871522848339938/pdf/Public-Sector-Savings-and-Revenue-from-Identification-Systems-Opportunities-and-Constraints.pdf (accessed 28 October 2019).

Zou, James and Schiebinger, Londa, 'AI can be sexist and racist – it's time to make it fair', *Nature* 559 (2018), 324–326, https://www.nature.com/articles/d41586-018-05707-8 (accessed 15 September 2019).



Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn and Eschborn

Friedrich-Ebert-Allee 32 + 36
53113 Bonn, Deutschland / Germany
T +49 228 44 60-0
F +49 228 44 60-17 66

Dag-Hammarskjöld-Weg 1 - 5
65760 Eschborn, Germany
T +49 61 96 79-0
F +49 61 96 79-11 15

E info@giz.de
I www.giz.de