Responsible AI

FOR DISASTER RISK MANAGEMENT

Working Group Summary

Deltares

GFDRR
Global Facility for Disaster Reduction and Recovery

WORLD BANK GROUP

The Global Facility for Disaster Reduction and Recovery (GFDRR) is a global partnership that helps developing countries better understand and reduce their vulnerabilities to natural hazards and adapt to climate change. Working with over 400 local, national, regional, and international partners, GFDRR provides grant financing, technical assistance, training and knowledge sharing activities to mainstream disaster and climate risk management in policies and strategies. Managed by the World Bank, GFDRR is supported by 34 countries and 9 international organizations.

www.gfdrr.org

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
CDR	Call Detail Record
DRM	Disaster Risk Management
FCV	Fragility, Conflict and Violence
GFDRR	Global Facility for Disaster Reduction and Recovery
HOT	Humanitarian OpenstreetMap Team
ICT	Information and Communication Technologies
ML	Machine Learning
RS	Remote Sensing
SAR	Synthetic Aperture Radar
TRA	Technology Readiness Assessment

Acknowledgements

This publication was a collaborative effort of numerous individuals, organized and convened by staff of the Global Facility for Disaster Reduction and Recovery, Deltares, and the University of Toronto.

The authoring team consisted of Robert Soden (University of Toronto & GFDRR), Dennis Wagenaar (Deltares), and Annegien Tijssen (Deltares). We received additional support from Dave Luo, Caroline Gevaeert, Marc van den Homberg, Grace Doherty, Manveer Kalirai, Vivien Deparday.

Participants and guest speakers in the Responsible AI for Disaster Risk Management Working Group included: Sarah Antos, Simone Balog-Way, Derrick Bonafilia, Erin Coughlan de Perez, Rob Emanuele, Sheldon Fernandez, Catalina Jaime, Marc van den Homberg, Caitlin Howarth, Heather Leson, Emily Miller, Martha Morrissey, Jonathan Nuttall, Alex Pompe, Joao Porto de Albuquerque, Tyler Radford, Mark Wronkiewicz, and Zhuang-Fang Yi.

The team expresses thanks to Julie Dana who as GFDRR Practice Manager encouraged us to bring an ethics lens to this issue. Additional feedback on early drafts of the report was received from Louis Conway, Erin Coughlan de Perez, Allan Hallsworth, Christian Klose, David Lallemant, Emily Miller, Diana Morales, Jonathan Nuttall, and Shaun Williams.

The report was finalized after a peer review process chaired by Niels Holm-Nielsen (Practice Manager, GFDRR) with inputs from (Data Scientist, World Bank), Vivien Deparday (Disaster Risk Management Specialist, World Bank), Pierre Chrzanowski (Disaster Risk Management Specialist, GFDRR), and Nicolas Longépé (Disaster Risk Management and AI focal point, European Space Agency). Nick Jones acted as Task Team Leader for GFDRR.

Design and illustrations by Estudio Relativo.

CHAPTER

01. Introduction

This document is intended to help practitioners and project managers working in disaster risk ensure that the deployment of artificial intelligence (AI), and machine learning (ML) in particular, is done in a manner that is both effective and responsible. The content of this report was produced as part of a 6-month interdisciplinary collaboration between experts from intergovernmental organizations, non-profits, academia, and the private sector. While we do not claim to offer the last word on this important topic, we are publishing the results of this collaboration in order to generate further discussion and help advance efforts towards better understanding the role of these technologies in pursuit of a world that is safer, more equitable, and more sustainable. It is our hope that—as a product produced through intensive consultation with the community for whom it is written—this document will inform and improve the important work carried about by data scientists, risk modellers, and other technical experts working in disaster risk management (DRM).

Many members of our community are working to explore opportunities offered by machine learning technologies and expand the range of applications for which they are used.[1] While we welcome the potential of these tools, we also need to pay close attention to the significant risks that unconsidered deployment of these tools may create in extremely complex contexts in which DRM is undertaken, across the whole cycle of preparedness, response, recovery and mitigation. Important questions are currently being raised by academics, journalists, and the public to questions of the ethics and bias of AI systems across a variety of domains including facial recognition, weapons systems, search, and criminal justice. *Despite the significant potential for negative impacts of these tools and methodologies in disaster risk management, our community has not given these issues as much attention as other domains.*

Some specific risks of improper use of machine learning include:

- Perpetuating and aggravating societal inequalities through the use of biased data sets
- Aggravating privacy and security concerns in Fragility, Conflict and Violence (FCV) settings through a combination of previously distinct data sets
- Limiting opportunities for public participation in disaster risk

1 GFDRR. (2018) 'Machine Learning for Disaster Risk Management', *GFDRR Knowledge Hub.* https://www.gfdrr.org/en/publication/machine-learning-disaster-risk-managemen

management due to the increased complexity of data products
- Reducing the role of expert judgment in data and modelling tasks in turn increasing the probability of error or misuse
- Hype and availability of private sector funding for artificial intelligence technology in DRM leads to overstatement of the capacities of these tools and the deployment of untested approaches in safety-critical scenarios

These are risks that need to be weighed seriously against the potential benefits before introducing new technologies into disaster information systems. While there are some relevant guidelines being produced in adjacent domains, the conversation is still evolving. In some cases, like facial recognition, experts have begun to recommend not using it at all, and it has been banned in a number of jurisdictions. It is too early to know how this debate will play out in the field of disaster risk management, so it is worth proceeding with caution.

There are many ways by which, as we will discuss, through technical means as well as improved project design and management, machine learning projects can be more responsibly developed and applied. However, we will also note some of these issues go deeper than machine learning, and are rooted in how we collect disaster risk and impact data, and how we design DRM projects more broadly. The heightened focus on the social impacts of AI tools offers an opportunity to draw attention to some of these questions. We will return to this issue in Section 7.

To develop this document, a team comprised of disaster data experts from the World Bank Global Facility for Disaster Reduction and Recovery (GFDRR), University of Toronto, and Deltares convened 7 online meetings between January and March of 2020, where machine learning experts and other researchers working in the area of disaster data discussed the opportunities offered by machine learning tools in disaster risk management, potential risks raised by these tools, and opportunities for mitigation. On average, 17 individuals participated in each 90-minute session. These interdisciplinary conversations were shaped by joint readings of relevant research and presentation of detailed case studies. In addition, the project team conducted 14 in-depth interviews with data scientists working on these topics for their views.

We present the preliminary findings of this process here, for wider review by the community. We have organized the body of the work around 4 key sources of threat: bias, privacy and security risks, lack of transparency and explainability, and hype. For each, we have presented an overview of the topic, realistic threat models along with hypothetical examples, strategies for managing these risks, and suggested further reading. Wherever possible, we keep the text here short and provide plenty of footnotes and links to more information. Through this process, we sought to produce these recommendations collectively as members of the community of expert researchers and practitioners working to create and use disaster data effectively and responsibly. We look forward to continued discussions with the DRM community on the contents of this report and to continued exploration into these important issues.

CHAPTER

02.

MACHINE LEARNING FOR DRM:

The Potential

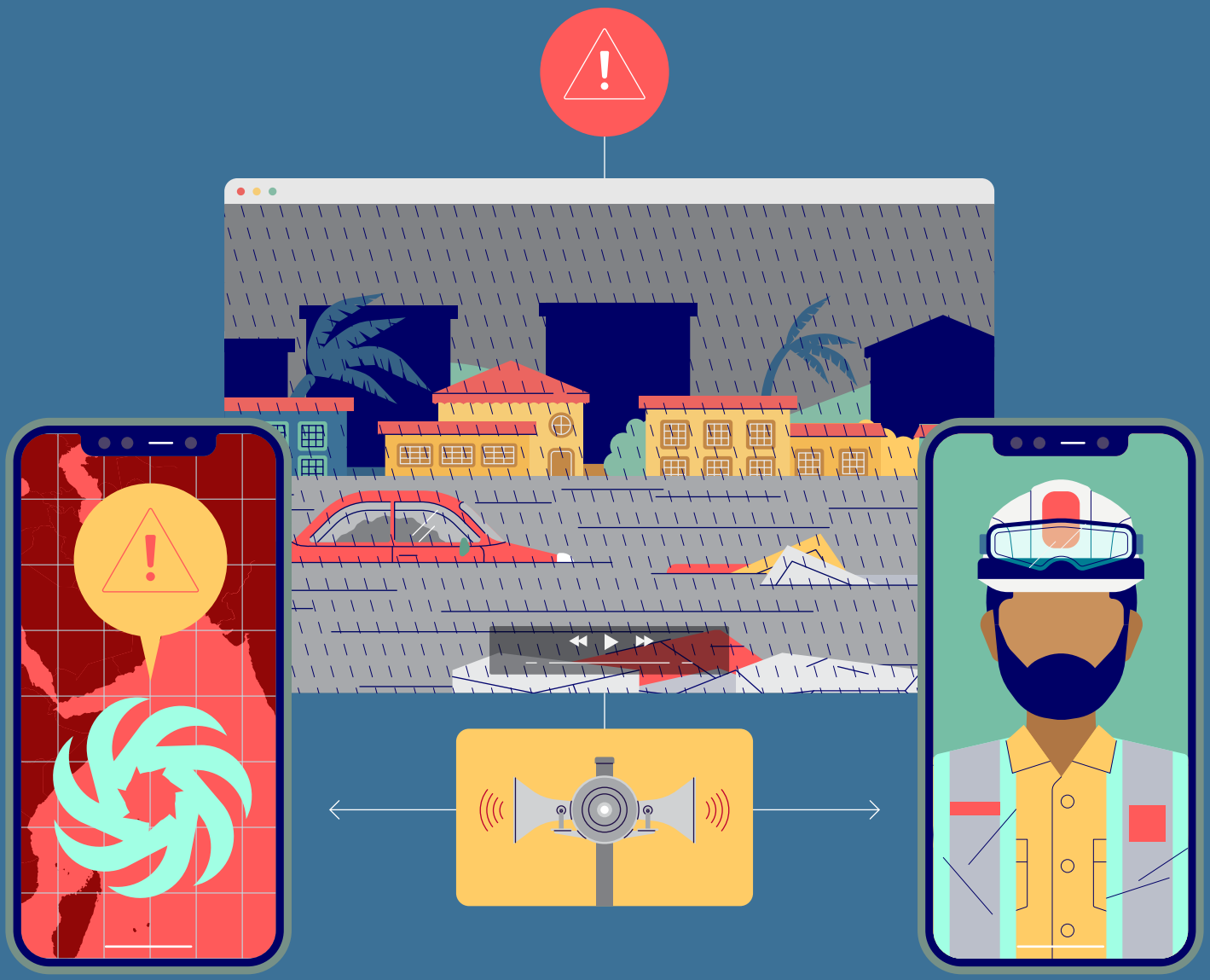
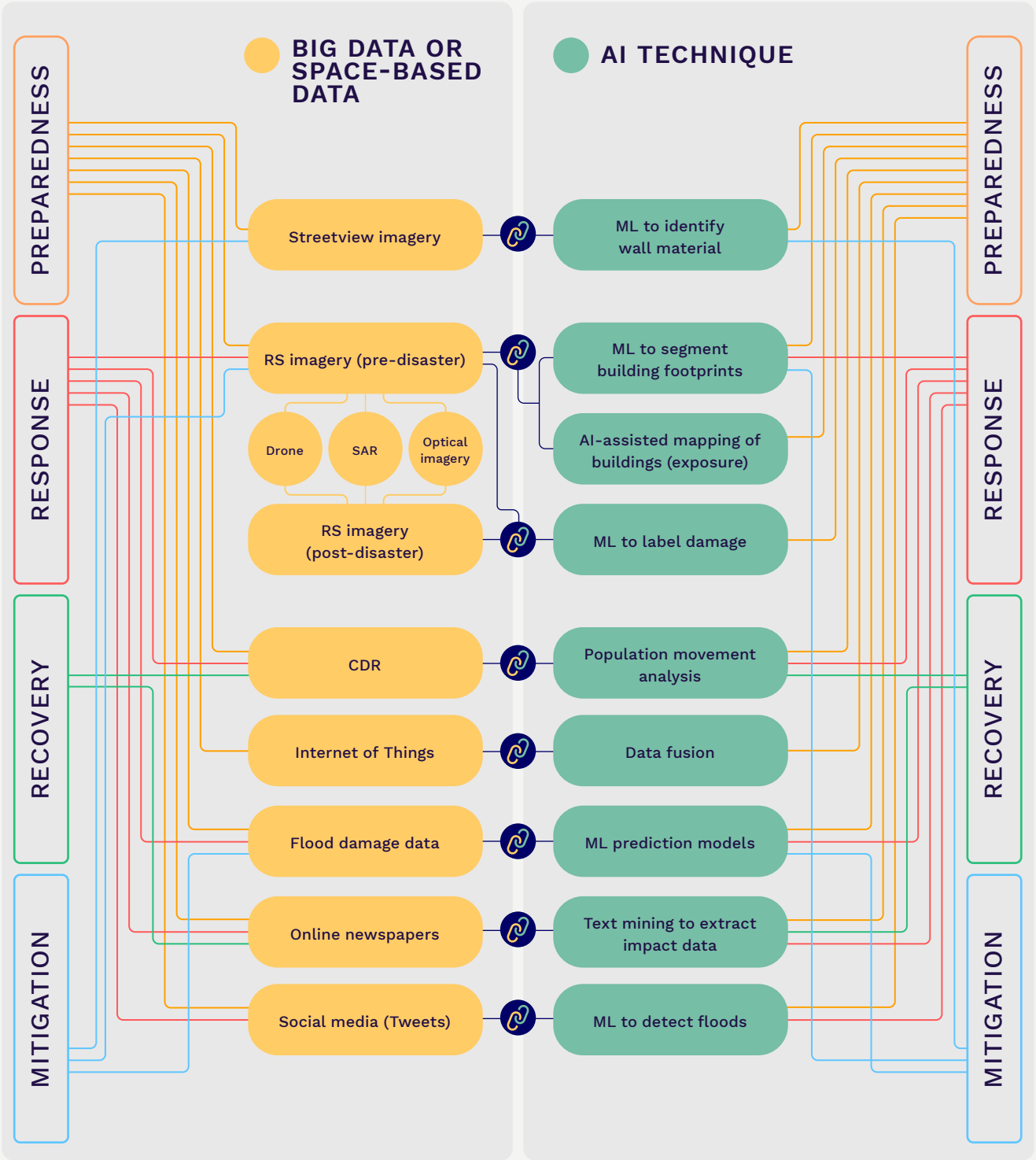


Figure 1 | The potential of ML techniques in combination with novel big data and space-based data along the DRM cycle.

RS Remote Sensing
ML Machine Learning
SAR Synthetic Aperture Radar
CDR Call Detail Records
AI Artificial Intelligence



Artificial Intelligence (AI) in combination with increasing amounts of data becoming available is likely to change the way we model and manage disaster risk. Disaster models are currently used in a range of products across the DRM cycle including infrastructure planning, insurance products, and early warning early action systems. These models typically consist of a combination of risk (hazard, vulnerability, and exposure information) and historical impact data. The last decades have shown significant technological advances, predominantly in the ICT and remote sensing domain, such as the increased use of satellites, drones, street view imagery, social media, smartphones and the Internet of Things. These technological advances have led to an exponential increase of Big Data that provide additional information on risk and impact. In addition to Big Data, Small Data based on sampling techniques from a wide variety of stakeholders, is becoming increasingly accessible online and can be analyzed with novel AI techniques.

When modelling disaster risk using traditional approaches, experts rely on digital tools and many different sources of data such as historical water levels, ground elevation or building information. This data is derived from sources including field measurements, remote sensing, and citizen science. Such measurements are then used to develop physics-based models to describe hazards (e.g. floods) or vulnerability (e.g. how a building responds to an earthquake). The expert judgment of modellers is used to design and calibrate such models throughout the process, as well as provide suggestions of the suitability of a model for various use cases. Machine learning offers a different approach.

Machine learning (ML) is the most common form of AI used in DRM applications, and this report focuses on these tools. ML techniques work by developing programmatic means to find patterns in given data and then using these patterns to make predictions in other comparable situations. For example, ML tools are used to look at historical records of water depths and damage and produce a model that can predict damages given a water depth. In other fields, these methods are being applied for technologies used in healthcare, self-driving cars, recommender systems and speech recognition.

For disaster risk management, machine learning is being used both to extract features from raw sensor data or to establish the relationships between this data. For example, building information is typically required for risk modelling. The acquisition of this data in the past often required considerable manual

labour, such as digitizing each building from a top-view image. Work to automate a task like this using machine learning methods is now achieving reasonable success. Examples include: recent SpaceNet challenges using Synthetic Aperture Radar (SAR) and optical imagery as data sources,[2] other projects relying on drone imagery,[3] and Microsoft's AI assisted mapping of vulnerable areas together with the Humanitarian OpenstreetMap Team (HOT).[4,5]

Today, exposure data is collected with machine learning techniques applied to "street view" imagery, or 360 degree photography taken at street level. For example, machine learning techniques have already been applied in this way to identify vulnerable buildings for earthquakes in a case study in Guatemala City and a number of other cities across Latin America.[6] Similar techniques are being explored to find the ground floor elevation of buildings in street view images. This holds the potential to improve flood risk information because despite flood risk models being very sensitive to ground floor elevation, they are often neglected because this data is often prohibitively expensive to collect.

Machine learning is also applied as an alternative to physics-based models and expert estimates. For example, rather than using detailed models of how water moves through a river based on data about the physical characteristics of a given system, machine learning might be applied to find patterns between previous rainfall records and measured water levels. It will also probably increasingly be used for weather forecasts in general.[7] It can and has been applied for disaster impact estimates. For example, machine learning has

2 SpaceNet. *Challenges*. https://spacenet.ai/challenges/

3 DrivenData. *Open Cities AI Challenge: Segmenting Buildings for Disaster Resilience*. https://www.drivendata.org/competitions/60/building-segmentation-disaster-resilience/

4 Microsoft. *AI for Humanitarian Action*. https://www.microsoft.com/en-us/ai/ai-for-humanitarian-action

5 Humanitarian OpenStreetMap Team. *HOT is an international team dedicated to humanitarian action and community development through open mapping*. https://www.hotosm.org/

6 GFDRR. (2018) 'Machine Learning for Disaster Risk Management', *GFDRR Knowledge Hub*. https://www.gfdrr.org/en/publication/machine-learning-disaster-risk-management

7 Palmer, T. 'A Vision for Numerical Weather Prediction in 2030', *arXiv Preprint, 2007.04830*. https://arxiv.org/ftp/arxiv/papers/2007/2007.04830.pdf

been applied to predict the impact of typhoons based on historic records of wind speeds, building materials and damage and in forecast-based financing schemes in the Philippines.[8] It is also possible to detect and estimate disaster-induced building damage severity by comparing pre-and post-event remote sensing imagery of an area (e.g. xView2 competition: https://xview2.org/),[9] detect floods occurring using Tweets,[10] or detect flood depths from images posted on social media.

This field is changing rapidly, and new use cases for the role of ML in DRM are being continuously identified. ML tools have potential to contribute to DRM efforts by making the collection and analysis of disaster information more accurate, timely, and cost-effective. In recognition of the problems with AI approaches that have been demonstrated in other fields, the working group on Responsible AI for DRM worked to identify potential consequences if these tools are not utilized in a measured and responsible fashion, as well as practical steps that can be taken to mitigate such potential. While this conversation is ongoing, this document summarizes the working group's discussions to date and related work such as the output of the Responsible AI Open Cities AI Challenge.[11] The purpose of this document is to highlight potential negative aspects of AI systems. However, AI can be very beneficial for DRM and in the suggested reading we refer to some documents highlighting the positive aspects.

Suggested Readings

- GFDRR. (2018) 'Machine Learning for Disaster Risk Management', GFDRR Knowledge Hub. https://www.gfdrr.org/en/publication/machine-learning-disaster-risk-management

8 510. (2019) Automated Impact Map Sent 120HRS Before Typhoon Kammuri Arrives. https://www.510.global/automated-impact-map-sent-120hrs-before-typhoon-kammuri-arrives/ van den Homberg, Marc JC, Caroline M. Gevaert, and Yola Georgiadou. "The changing face of accountability in humanitarianism: Using artificial intelligence for anticipatory action." Politics and Governance 8, no. 4 (2020): 456-467 https://www.cogitatiopress.com/politicsandgovernance/article/view/3158.

9 xView2. Computer Vision for Building Damage Assessment. https://xview2.org/

10 de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., & Aerts, J. C. (2019). A global database of historic and real-time flood events based on social media. Scientific data, 6(1), 1-12.

11 DrivenData. *Open Cities AI Challenge: Segmenting Buildings for Disaster Resilience*. https://www.drivendata.org/competitions/60/building-segmentation-disaster-resilience/

CHAPTER

03.

CONCERNS:

Bias in Machine Learning Systems

Inevitably, all systems will have some form of bias. All models of the world are, by necessity, incomplete and inherit, sometimes unknowingly, the priorities, prejudices, and perspectives of their creators and the wider society in which they are developed and used. Understanding and mitigating the sources, as well as the consequences, of bias in the information used to guide disaster risk management initiatives is therefore a necessary component of ensuring that machine learning tools are used in a responsible and fair manner.

Concerns over bias occupy center stage in many of the ongoing discussions over the societal consequences of machine learning tools. One challenge of addressing bias is that varying definitions of bias and fairness are used in these discussions. Statisticians and systems developers often have significant expertise in identifying and correcting measurement errors that can lead to biased data sets through over- or under-representation of certain phenomena. Meanwhile, legal scholars and social scientists point to unfair representations of disadvantaged groups in data sets or the reinforcement of societal inequities as a result of decision-making processes based on algorithms. Biases can appear in every major stage of an AI model's development, as illustrated by Suresh and Guttag (2019). In this guidance document, we choose to address most (if not all) of these aspects of bias and fairness in ML systems.

Sources of Bias and their Harms

Figure 2 | Types of bias - Data biases can be classified as historical biases, representation biases, measurement biases, aggregation biases and evaluation biases.

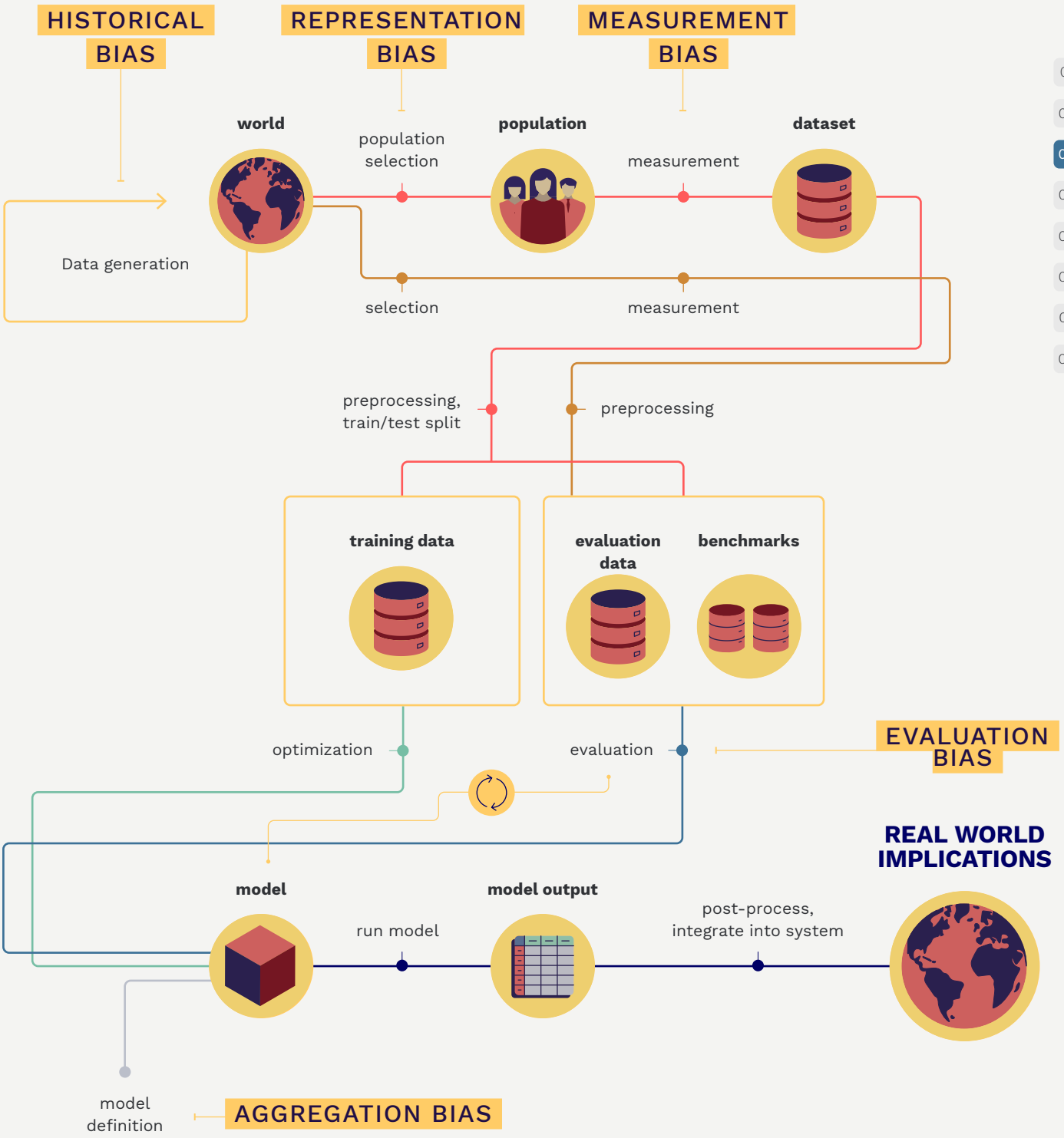


Image source: Suresh, H. and Guttag, J. (2019) "A Framework for Understanding Unintended Consequences of Machine Learning." *arXiv preprint arXiv:1901.10002*. https://arxiv.org/pdf/1901.10002.pdf

Historical biases occur when a model perfectly represents the world but copies unwanted patterns from the real world into the model (e.g. stereotypes).

Representation biases occur when the data used to train the model isn't representative for the problem that needs to be solved. DRM models often need to predict extremes, however available data typically focuses on the common situations and will perform poorly on extremes. A variation on this problem is that data is only available for another region or another time and hence needs to be transferred to the area or time of interest.

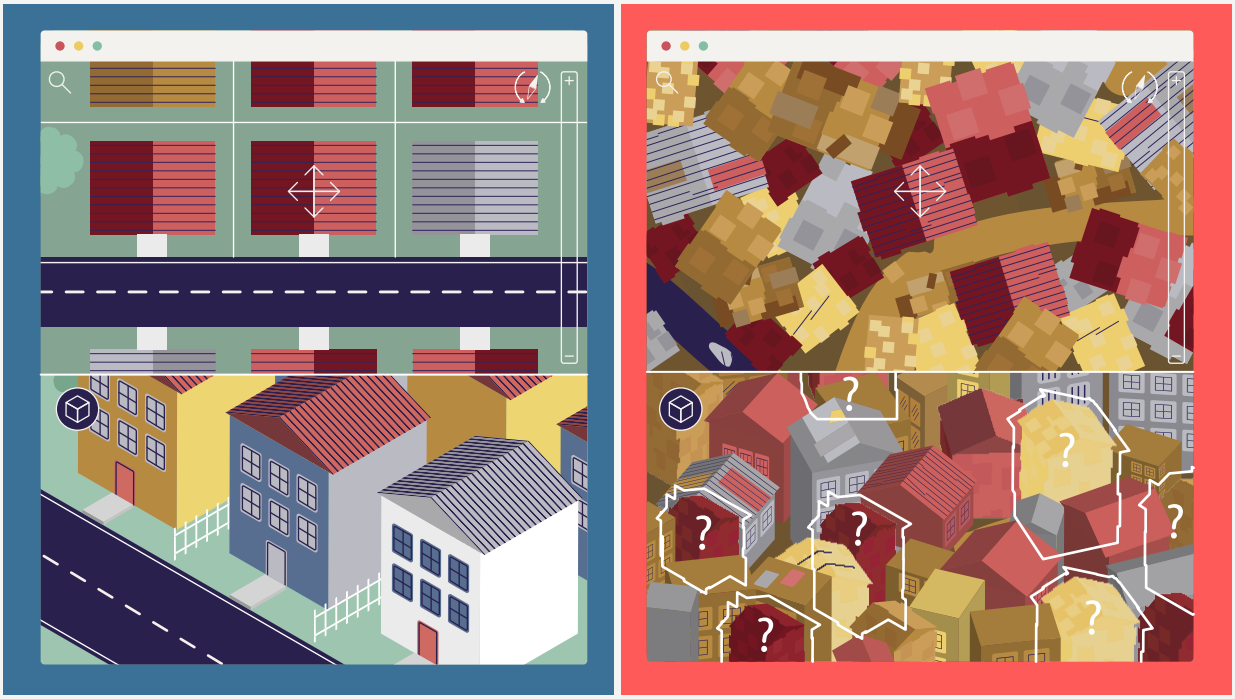
Measurement biases occur when the wrong input or output for the model is picked. For example, when important predictors for storm damages are not included in the model or when the model predicts wind speed while decision makers need wind damage.

Aggregation biases occur when data are aggregated in such a way that useful information is lost. ML models often predict mean values but hide the variability around that mean. For example, the mean earthquake damage in a district may be low but some vulnerable buildings of interest may still have a lot of damage. This information may be lost by the aggregation in the model but could be very important to take action.

Evaluation biases occur when a model isn't evaluated correctly. For example, it is evaluated based on the wrong parameters or based on data that isn't representative of the way the model will be used.

Biased Training Data -

One of the most common sources of bias in a machine learning project is the training data that is used to train algorithms. A training data set used for DRM projects is biased if the data it contains doesn't reflect the context to which the algorithm is meant to be used. Such mismatches can be rooted in geography, demography, based on structure type, or other kinds of data. They can result in poor or biased outcomes in model results or attempting to transfer a model from one setting to another.



Example 1: Assessing Earthquake Impacts

Null Island has recently undergone a high-intensity earthquake, affecting the majority of the country. Experts scramble to pull together the best data they have in order to conduct a rapid impact assessment aimed at guiding relief and early recovery work. As is common in many parts of the world, high resolution pre-event aerial imagery is primarily available for the capital city center. The available imagery is used to train a computer vision model that compares pre-event data with imagery taken just after the earthquake. When that model is then used to estimate building damage across the whole country, it significantly undercounts impacts in rural areas and informal settlements. This bias in the damage assessment unfairly leads to the capital city receiving a disproportionate share of recovery and reconstruction assistance.

Biased approaches to measurement -

Measurement biases are often difficult to identify or evaluate. This source of bias stems from decisions made in how we decide to represent extraordinarily complex phenomena like disaster risk or impacts with data. Data standards like the 100-year floodplain, or measures such as average annual loss serve as useful but incomplete indicators for often much broader processes or events in the world. While data scientists are intimately familiar with this issue, this awareness or nuance is often lost when models and data are brought into project planning and policy-making. Also an understanding of the quality of underlying data and the limitations inherent in that are often

lost when a model is handed over from data scientists to project planners or policy makers. For example, the understanding of spatial variability in the quality of rainfall data. Bias in measurement approaches runs deeper than machine learning, but is important to consider as it will impact project outcomes and, in some cases, machine learning may exacerbate them.

Example 2: Calculating Risk in DisasterLand

DisasterLand is kicking off a national multi-hazard risk assessment in order to set priorities for disaster-risk management investments. The cost-benefit analysis that is used to set the priorities relies on a ML model using a lot of variables including property values in order to determine the "benefit" of protecting buildings and places against hazards like earthquakes or floods. This inherently puts communities with lower economic status at an unfair disadvantage; as disaster risk reduction measures for wealthy neighbourhoods will come out as more cost-effective, which could lead to under-protection of less well-off areas.

Example 3: Evacuation Planning in Alphabet City

Alphabet City, the capital of Null Island, is updating its evacuation plans and wants to use machine learning models along with transportation and population data sets to plan key routes and determine areas that will need extra assistance. Unfortunately, the latest estimates on population don't include important details on disability status. In addition, the team is planning to use Call Detail Record (CDR) data in order to understand variable population density across the city at different times of day. As a result of privacy regulations on Null Island, CDR data isn't disaggregated by gender or age. Evacuation planning models that assume all people are adult able-bodied men (as is common), or don't otherwise account for the great range of needs and capacities of potential evacuees will misallocate resources and be improperly prepared should the time come when evacuation is necessary.

Recommendations for Addressing Bias

1. **Remember that, in a general sense, all models are biased.** All models of the world are necessarily incomplete. This simplification is a necessary part of how they function. Even if perfectly unbiased data sets were possible, this would not resolve all potential problems such as when, for example, the ML tools are supporting unjust systems. As developers and designers of ML systems, we need to understand and continually interrogate the limits of our systems and what that means for decisions based upon them.

2. **Technical solutions** such as continuous validation, fairness tests, or scrubbing training data sets can help mitigate or reduce biases in training data sets. The techniques differ depending on the type of bias encountered and should be routinely deployed as part of every project.

3. **Participatory strategies** can be used to include residents of areas that ML models describe in training data collection and validation processes. These strategies can help to identify sources of bias, potential misuses and add understanding of local context to the project planning and design.

4. **Diversify project teams.** A recent report has shown that, as an industry, AI is unfortunately quite homogenous.[12] Diversifying our project teams over relevant dimensions such as gender, race, expertise and socio-economic background is necessary because, as developers of these systems, we often have unexamined assumptions based on our own backgrounds. A more diverse team could find and address blindspots in order to identify potential problems early in the process of e.g., data set or code reviews. This will also help to improve documentation for future maintenance of models or potential regulatory auditing.

12 West S.M., Whittaker, M. and Crawford, K. (2019) 'Discriminating Systems: Gender, Race, and Power in AI', *AINow* https://ainowinstitute.org/discriminatingsystems.pdf

5. **Provide users with information they need to evaluate the results properly.** When the results of ML projects are shared or published, sufficient information about the training data, models, uncertainties, and development process should be included along with guidance on how to responsibly interpret the results. See the section on Transparency and Explainability for more details.

Suggested Readings

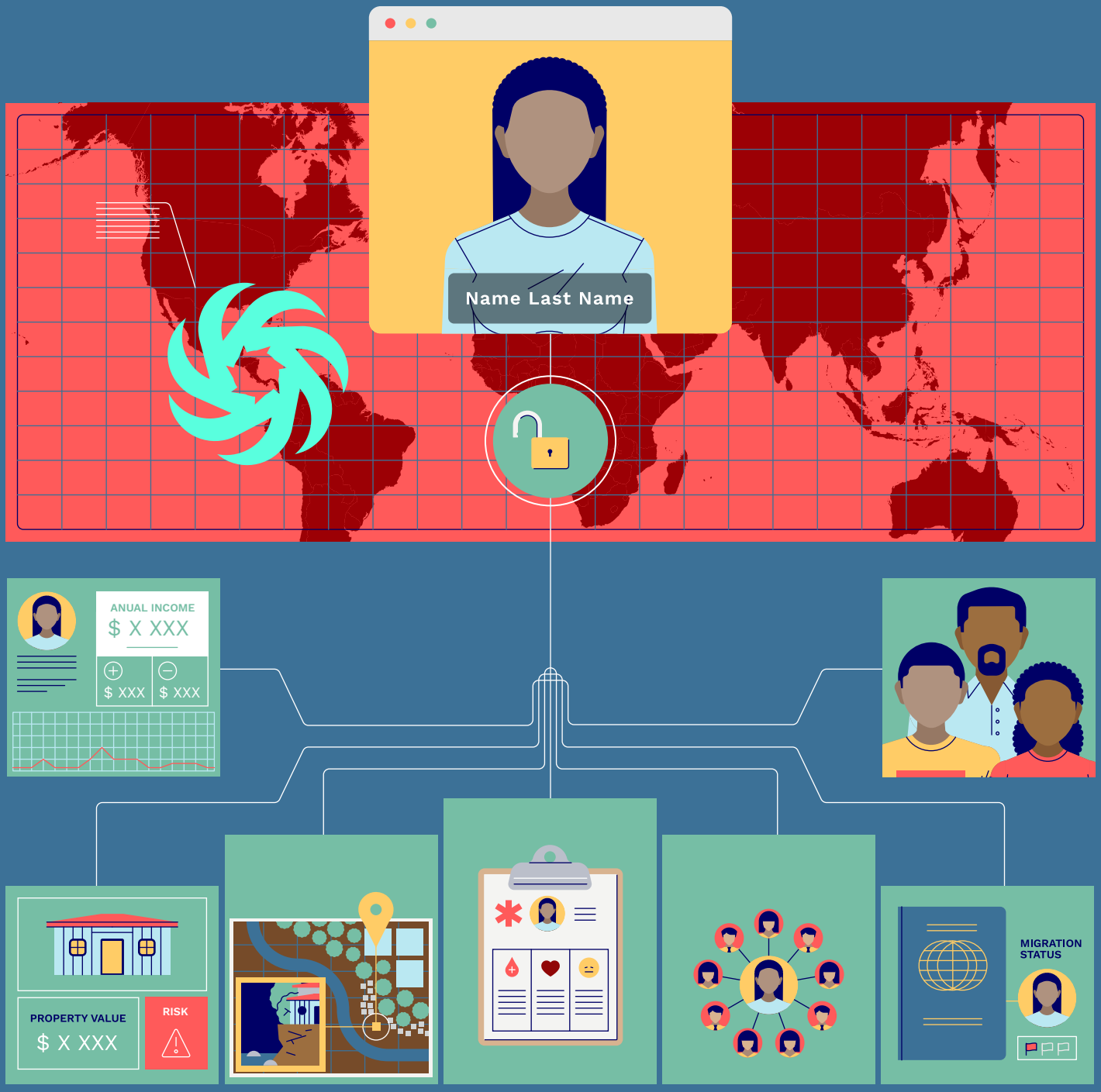
- Kumaraswamy, A. (2017) 20 lessons on bias in machine learning systems by Kate Crawford at NIPS 2017. https://hub.packtpub.com/20-lessons-bias-machine-learning-systems-nips-2017/
- Kusner, M.J. and Loftus, J.R. (2020) 'The long road to fairer algorithms', Nature, 578, 34-36. DOI: https://doi.org/10.1038/d41586-020-00274-3
- Pestre, G., Letouzé, E. and Zagheni, E. (2020) 'The ABCDE of big data: assessing biases in call-detail records for development estimates', The World Bank Economic Review, 34(Supplement_1), pp.S89-S97. DOI: https://doi.org/10.1093/wber/lhz039
- Powles, J. (2018) The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53
- Suresh, H. (2019) The Problem with "Biased Data." https://medium.com/@harinisuresh/the-problem-with-biased-data-5700005e514c

CHAPTER

04.

CONCERNS:

Privacy and Security



Machine learning tools are data intensive, and rely on the collection and analysis of large data sets describing people, places, and other aspects of the world. The kinds of data we work with—high-resolution drone imagery, CDR data, or information on economic activity or demographics of an area—can be invaluable to modelling disaster but risk violating individual privacy and may present security issues in some settings. In projects related to disaster response or prevention, these issues may be compounded as there is often a temptation to relax privacy and security measures in favor of speed or data resolution.

The informed consent of individuals whose personal data is being collected and used has long been central to research ethics and is part of GDPR, IRB, and other privacy regulations. In DRM projects, where data is often not of a directly personal nature, the role and application of informed consent is murky. Concerns over the potential negative impacts of releasing data about the exposure and vulnerability of communities to floods or other hazards are part of ongoing debates and court cases, where harms that might be caused by such release are weighted against the values of openness and transparency in disaster information. In many cases, governments act as proxies for the public in providing consent to DRM projects to collect, analyze, and distribute risk data, but in some cases, in FCV settings for example, this may not always be appropriate.

One of the challenges for ML practitioners in trying to ensure that our work does not cause these kinds of harm is the fact that both privacy and security issues are highly dependent on context. What counts as sensitive information in one context or culture will not be the same as another. For example, the information that, if released, could present a security risk to communities living in a conflict zone may not be obvious without detailed understanding of that context of the conflict.

How does GDPR and other privacy laws impact ML for DRM?

In response to privacy concerns about the use of data, the European Union enacted stringent privacy laws called GDPR. These laws concern the collection of personal data, that is any data that can directly or indirectly be used to identify a person. It sets requirements about consent, security, rights of data subjects, and automated decision making.

Not all data in DRM is personal data and hence the impact of GDPR on DRM is limited. For example, the most common data such as hazard data and data about building characteristics seems to fall outside the scope of GDPR and can still be used within GDPR. The GDPR does also not include any rules about collecting data on marginalized groups.

Privacy and Security Risks

Unanticipated use of overly-granular or unnecessary data -

In the examples the working group reviewed, one of the primary sources of privacy and security threats in disaster data projects was simply the collection of data in the first place. As part of modelling teams' goals to produce the most accurate information possible, it can be tempting to collect as much as possible. However once collected, the data creates the opportunity for unanticipated or unforeseen use, with potential harms to the people and places the data is about.

Example 1: Satellite imagery in conflict areas

Engineers are working with vulnerable communities on monitoring floods in a part of DisasterLand experiencing conflict. They are using high resolution satellite imagery to train models that will be able to quickly assess flood impact and extent in the event of a disaster. As part of a public briefing on the issue they make a screenshot of a satellite photo of a community near a river. Though they had carefully removed any geographic metadata from the image, combatants from the region were able to identify the location of the community through characteristic landscape features. The safety of the community was thus inadvertently compromised.

Example 2: Flood maps and property values

As part of a transparency effort, Alphabet City, began releasing detailed flood maps of the city to the public and launched a public review process as part of updating or creating any new maps in the early 2000s. While this has improved public awareness and involvement with flood issues, it has also led to an unintended outcome. Fearing, perhaps rightly, that having their homes included in a flood zone would impact property values, residents now regularly arrive at public meetings or file petitions to contest the maps. Flood mapping in Alphabet City is now a highly contested process, raising the costs and making many maps out of date as the time needed for review has increased.

Deanonymizing data -

Data sets are often anonymized by removing columns with sensitive data such as names or addresses. This technique is however not completely safe and it's often possible to deanonymize this data and reconstruct deleted columns using the other available data.

Importing standards of privacy from one context into another -

What counts as private information is culturally specific and thus varies from setting to setting. As many machine learning projects are now incorporating high resolution drone and street view data, we risk collecting and distributing information that breaches privacy. Even if proper local consultation is conducted in one place to understand these potential risks, we should not assume they are the same in another.

Addressing Privacy and Security Risks

1. **Evaluate unintended consequences** - Project teams should discuss potential unintended consequences and develop a theory of harm,[13] or a robust set of possible issues that could arise from the use of a particular in the project context. These conversations should be informed by an in-depth understanding of local context and examples of where other projects have gone wrong, as well as speculative design techniques that aim to create detailed scenarios where such potential may be explored.[14]

2. **Fitness for purpose** - Data collection projects have high overhead costs; they are expensive and time-consuming just to plan and set up. This, combined with reasonable goals of developing the most detailed models possible, can encourage project designers to collect unnecessarily granular information or additional details about people and places of concern. Unless the aim of the project is to produce fundamental data sets, projects should instead try to focus on specifically what is necessary. This will both reduce the potential for unintended consequences or misuse as well as limit the necessary work to understand this potential for the data that is being collected. Sometimes, the value of data is not always clear in earlier stages of a project. Therefore, a trade-off exists between reducing potential privacy risks and potentially reducing biases.

3. **Beyond data collection** - Consider privacy and security not only in data collection, but throughout the whole process. The full lifecycle analysis of ML projects for privacy and security concerns suggest looking for opportunities to: (i) avoid unnecessary or overly granular data collection, (ii) remove sensitive parts before releasing it,

13 Sandvik, K.B. and Raymond, N.A. (2017) 'Beyond the Protective Effect: Towards a Theory of Harm for Information Communication Technologies in Mass Atrocity Response', *Genocide Studies and Prevention: An International Journal*, 11(1), p.5 DOI:http://doi.org/10.5038/1911-9933.11.1.1454

14 CHI4Evil. *Creative Speculation on the Negative Effects of HCI Research* https://chi4evil.wordpress.com/

(iii) mitigate misuse of sensitive data. (iv) consider how presentation of the data (e.g. removing or blurring certain information), and (v) destroying sensitive data when the project is completed.

4. **Technical solutions** to anonymize individual data. Data sets shared among researchers are often anonymized to protect the privacy of data subjects. The most common way to do this is to remove the most personally identifiable data. This makes it more difficult to identify an individual but it is often still possible to reconstruct who a particular person is based on the unique combination of identifiers. A safer way to do this is to use "differential privacy", this is a technique to add random perturbations to the data set without removing the statistical properties.[15] This technique can be safe against many types of attacks but a drawback is that considerable information can be lost and fewer questions can be answered from the data. Hence this technique comes with a trade-off between bias and privacy.

15 Nguyen, A. (2019) *Understanding Differential Privacy* https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a

Suggested Readings

- Dorschel, A. (2019) Rethinking Data Privacy: The Impact of Machine Learning. https://medium.com/luminovo/data-privacy-in-machine-learning-a-technical-deep-dive-f7f0365b1d60
- Garcia, C. (2018) Everything You Need to Know About Informed Consent. https://humansofdata.atlan.com/2018/04/informed-consent/
- Greenwood, F., Howarth, C., Poole, D.E., Raymond, N.A. and Scarnecchia, D.P. (2017) 'The signal code: A human rights approach to information during crisis', Harvard Humanitarian Initiative. https://hhi.harvard.edu/publications/signal-code-human-rights-approach-information-during-crisis
- Prabhu, M. (2019) Security & Privacy considerations in Artificial Intelligence & Machine Learning — Part-6: Up close with Privacy. https://towardsdatascience.com/security-privacy-in-artificial-intelligence-machine-learning-part-6-up-close-with-privacy-3ae5334d4d4b
- Wright, J. and Verity, A. (2020) 'Artificial Intelligence Principles For Vulnerable Populations in Humanitarian Contexts', Digital Humanitarian Network. https://www.digitalhumanitarians.com/artificial-intelligence-principles-for-vulnerable-populations-in-humanitarian-contexts/

CHAPTER

05.

CONCERNS:

Lack of Transparency & Explainability

Predictions made by machine learning systems are often difficult to explain, even for the developers of the system. This "black-box" problem stems in part from the property of ML systems that the models they produce do not have a physical basis and are instead solely based on relationships found in data. This raises several related problems that the field of "explainable AI[16]" seeks to address:

- Sanity checks are different than in traditional models
- Reduced feeling of ownership of the system.
- Difficulty in motivating decisions to stakeholders.
- Possibility of misused spurious relationships.

Sanity checks are different -

Sanity checks are tests engineers apply to check whether calculations are correct. In traditional modelling, a modeller carries out the analysis in explicit steps. At each step, there are intermediate results that have a meaning and on which sanity checks can be carried out. This can be used to build confidence in the results and to spot potential problems. When final model predictions don't make sense, the modeller can work backwards to figure out where the model goes wrong. In a ML system, such sanity checks are different. The inherent "black-box" characteristic of ML systems makes it harder to check intermediate results, as the model is "self-learning" and not explicitly programmed. This makes it sometimes more difficult to debug the model or to gain confidence in it. This will also make it sometimes more difficult to find and mitigate problems of bias in a ML system.

Reduced feeling of ownership -

In a well-designed DRM project, it is common that models are built collaboratively in order to create a feeling of ownership of the model by the users. This creates trust and makes it more likely that model results are accepted. In such a process, users may be asked for input in the form of expert estimates and interaction between the modellers and decision makers takes place. It is clear where expert judgement feeds into the model and what the effect of this input is on the results and intermediate results. This leads to a transparent modelling process. As machine learning systems are not explicitly programmed, it is difficult to feed expert judgment

16 Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp.82-115.

into these systems and intermediate results sometimes can't be validated with many ML approaches. In these cases, expert judgement can only be used for validation of model outcomes and improving the training data of the model. These difficulties can lead to both expert and non-expert stakeholders feeling further removed from the actual modelling process, which could lead to a reduced sense of ownership for the model results.

Motivating decisions to the public and decision makers -

DRM models can lead to adverse decisions for some stakeholders. Such stakeholders are likely to want to challenge these decisions. When such a decision is made based on an assessment using a traditional approach, model results are (usually) easily explained. Someone who is adversely affected by the decision can use the choices and assumptions made in the modelling process to attest the decision. The explainability weaknesses of machine learning models make it harder to explain the outcomes of the model, which in turn will make it harder to attest decisions made on ML model outcomes. This can lead to a dispute or lack of trust in the modelling process leaving decision-makers in a difficult position. Decision-makers (may) prefer to use models that can be easily explained to avoid situations where the public distrusts the decision making process. To avoid decisions being made on unexplainable models, the GDPR law in the European Union states that data subjects can't be subject to a decision based solely on automated processing.

Example 1: Building a dike in Alpha town

Alpha town lies in a river valley and they want to protect themselves from flooding with a dike. The neighbouring town Beta, does not like this plan because it may increase water levels in their town. Both Alpha and Beta develop their own analysis using different ML systems. The model produced by Alpha town shows that the flood risk increase from their dike is negligible in Beta town. However the model produced by Beta town shows a considerable increase in flood risk in their town. The modellers of the two towns meet to discuss the differences but because they can't explain their model results to each other they cannot resolve this problem. Beta town proceeds to block the dike construction plan.

Example 2: Buyout program in Disasterland

Disasterland has a high level of flood risk and the government is responsible for compensating disaster victims for their losses. To reduce risk, they developed an ML-powered system to identify and buy out flood prone properties. Some homeowners have lived in the area for their entire life and really don't want to move. The ML system advises that some of these homeowners need to leave. The homeowners want to challenge this decision because some of their neighbours in seemingly similar situations are allowed to stay. Disasterland cannot provide a clear explanation and this results in a reduction of trust in the government and the buyout program in particular.

Misused spurious relationships -

ML systems are likely to find spurious relationships. These are relationships between variables that are not directly related. For example, cooking fuel may be an indicator of poverty, and poverty may be an indicator for disaster damage. So the ML system may assign a relationship between cooking fuel and disaster damage. Though such relationships are not necessarily always incorrect, in cases where better data isn't available, a proxy may be useful. However, it can lead to undesirable situations such as the model suggesting nonsensical measures—for example, changing the cooking fuel to reduce disaster risk. This means that when ML models find spurious relationships, this can create unwanted and/or perverse incentives in the system, where people can (mis)use the spurious relationships to cut corners and game the system. This is a very common problem in the search engine domain where web developers often make decisions based on what is perceived to be best for the search engine ranking.

Example 3: Insurance in Disasterland

A company provides insurance policies for storm damage in Disasterland. The height of this premium is based on a storm damage model based on a ML system. The insurance company has no data on roof strength, but the ML system found that poorer people typically cook with gas and also have typically weaker roofs while rich people tend to cook with electricity and typically have stronger roofs. So cooking fuel is applied to estimate the roof strength without the insurance company being fully aware of this relationship. A large investment firm owns many properties in Disasterland and they notice that when they change the cooking fuel to electricity they pay lower insurance premiums for a property. They start taking this information into account during renovations and slowly reduce their storm insurance premiums with each yearly premium reevaluation. However, they do not actually reduce their storm risk. After years of doing this they pay too little insurance premium to cover their risk and have made many investments that make no sense from a societal perspective.

Recommendations

1. **Technical solutions for sanity checks:** There are many techniques available to make ML systems more intelligible. The most common approach is to show which variables contribute to the ML system decisions. Graphs can be made that show the relationships between input variables and the output variable given that the other variables remain constant. Another approach is to list the data points that lead to a particular modelling decision. Also, techniques such as Bayesian Networks can be applied to understand how variables relate to each other. Such technical solutions can be applied to find spurious relationships and help with sanity checks. However, these techniques are only useful when the modeller has a good understanding of the data.

2. **Provide an interactive tool:** To build trust in a ML system, it can be useful to build an interactive tool where affected users/people can play with input variables and see how it affects the results. Such interactive tools make it very easy for users to do sanity checks and help to build some confidence in the model. Such interactivity could however make a system more vulnerable to malicious use. So for some systems an interactive tool might not be the best solution.

3. **Avenues for appeal, recourse, remedy and redress:** Good processes to challenge decisions based on ML systems can counteract some of the distrust generated by using a ML system. This can involve a formal appeal procedure, including human review and maximum disclosure of technical information applied to come to a decision such as described in the technical solutions above. In some cases, it may also include methods to redress wrong decisions. However, in some cases this may not be enough because it may not be possible to determine whether a decision was correct.

4. **Consider alternative approaches:** Sometimes a ML approach is simply not the best approach because it is not transparent enough. For example, when transparency is very important to satisfy all parties involved (e.g. in a conflict situation). This is however not a binary choice. Some ML techniques are more transparent than others. It is also often possible to go for hybrid approaches whereby some parts of the modelling chain are based on ML techniques and sensitive parts of the modelling chain are done with traditional methods. It is important to take this transparency criteria into account before the modelling process starts so the right technology can be chosen.

Suggested Readings

- ARTICLE 19. (2019) Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence. https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf
- Citron, D. and Pasquale, F.A. (2014) 'The Scored Society: Due Process for Automated Predictions', Washington Law Review. https://www.semanticscholar.org/paper/[89WashLRev0001]-The-Scored-Society%3A-Due-Process-Citron-Pasquale/6471a105b6982d0f31ab6e3dd53e58b62dbeb841#paper-header
- Dickson, B. (2018) 'The Next Step Toward Improving AI', PCMag. https://www.pcmag.com/news/the-next-step-toward-improving-ai
- Goodman, B. and Flaxman, S. (2017) 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"', AI Magazine. DOI: https://doi.org/10.1609/aimag.v38i3.2741
- Lador, S.M. (2019) Navigating the of Explainability. https://towardsdatascience.com/navigating-the-sea-of-explainability-649672aa7bdd

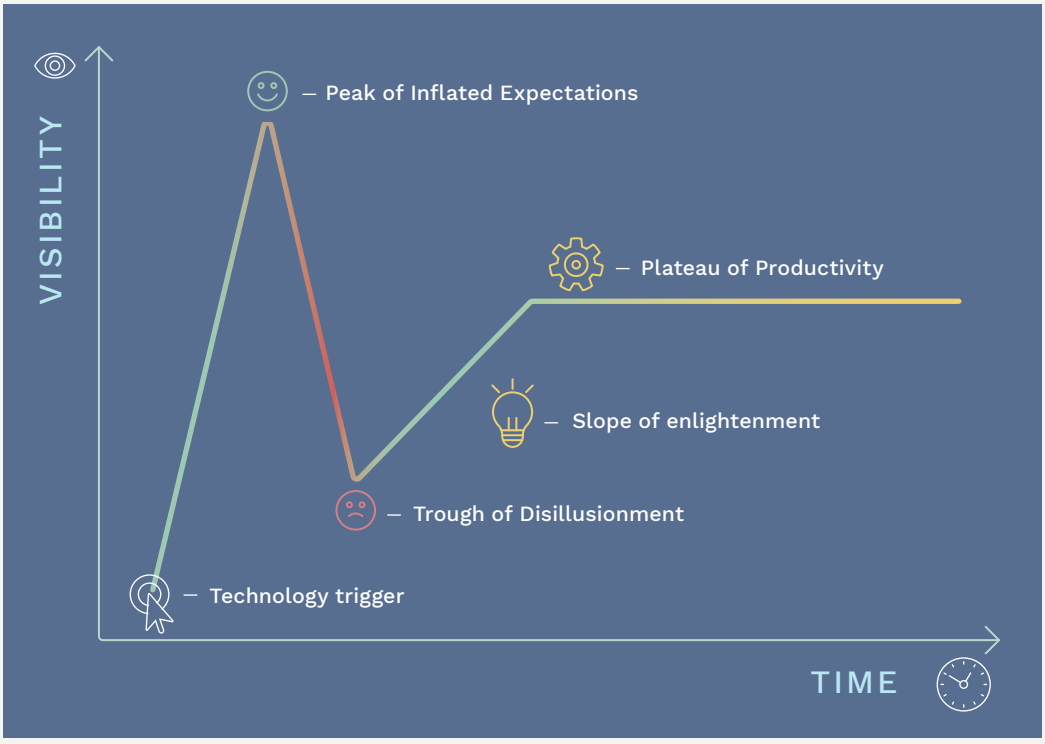
CHAPTER

06. CONCERNS: Hype

Enthusiasm in the DRM community about AI can lead to harmful hype. Hype is an inappropriate amount of publicity and/or unreasonably high expectations for the benefits or value of emerging technologies. The most well-known example of this was the dotcom bubble in the 90s when unrealistic short-term beliefs about the internet led to a trail of bankruptcies and an economic crisis. Hype is a common risk for potentially disruptive technologies such as AI.

This process of hype is described in a so-called "hype cycle," as depicted by Fenn and Raskino (2008).[17] The cycle begins with a technology trigger where stories about potential benefits of a new technology start circulating. From that point onwards, expectations increase until an unrealistic level and a peak of the hype is reached. After this peak, typically a period of disillusionment follows in which the expectations of the technology typically sink too low before they climb up again and finally reach a realistic expectation level as the technology matures, as shown in **Figure 3**. Different AI applications in DRM are at different places in the hype cycle, but most of them have not reached the realistic expectation levels from the end of the cycle yet and are often still far away from maturity.

Figure 3 | The hype cycle.[18]



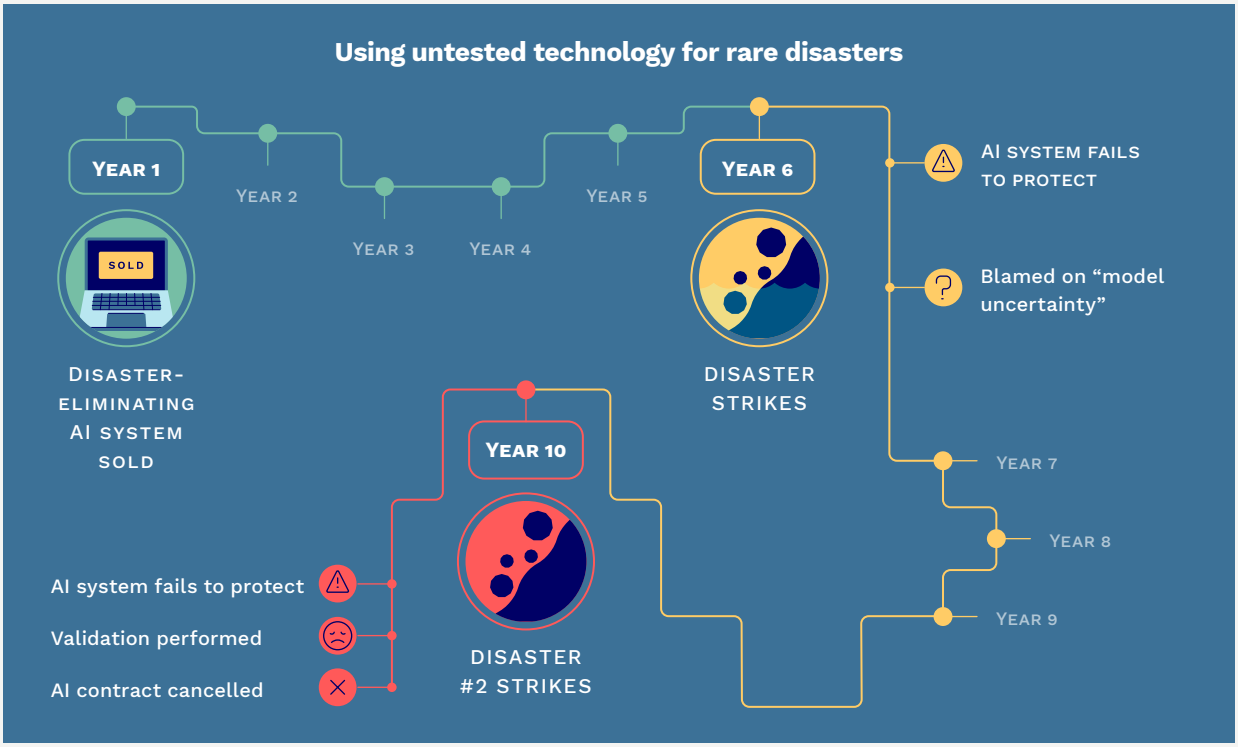
17 Fenn, J. and Raskino, M. (2008) 'Mastering the hype cycle: how to choose the right innovation at the right time', Harvard Business Press.

18 Ibid..

Unrealistic expectations of AI in DRM can be harmful in several ways. They can direct funding to projects that don't deserve it and leave more promising DRM projects unfunded. The second potential greater harm is that AI systems may be prematurely employed to undertake tasks for which they are not yet sufficiently prepared. This is especially harmful when good traditional alternatives are available (e.g. physics-based models for hazard modelling). Acting too soon could lead to decisions that may worsen the toll of disasters; decisions that may result in greater damage, and may even cost human lives.

A well-known case in another sector where hype set the stage for larger problems was the company Theranos. This company worked on developing a new technology that would be able to do cheaper and faster blood tests. The company staff had good credentials but the technology was kept hidden from the scientific community to protect its commercial interest. The hype surrounding Theranos helped it gather 700 million dollars in investments and with it, incredible expectations to fulfill. It eventually came to light that the technology wasn't working, a fact the management had kept hidden for years. In the end, the company was charged with fraud and it terminated operations. In the meantime, many people had received results from these blood tests that were later found to be untrustworthy, thus raising the likelihood of improper or incorrect diagnoses.

Disaster risk management may be especially susceptible to these problems because DRM model quality is often difficult to assess by outsiders. Models for rare unlikely events are especially difficult to validate because it may take decades before a large disaster happens. Inaccurate models could therefore be applied for years by decision-makers without them being aware of these shortcomings. In the rush to demonstrate results, hype may therefore lead to distortions in the model development process. These problems also risk discrediting genuinely innovative AI systems for DRM that do produce useful results.



Example 1: AI startup in Disasterland

A new startup launches in Disasterland touting a big promise to eliminate the impact of disasters with the help of AI. The founders have great credentials and the story sounds good to venture capitalists. The technology is hidden from the scientific community for commercial interests. In doing so, also hidden is the fact that the technology does not really work as well as traditional approaches yet. Pressured to show some results, the company starts selling their AI systems. They need to predict very rare disasters, so at first nobody notices their bad quality. More of the systems are being sold, driving away good traditional models. A large disaster happens, it is missed by the new AI system, and people die unnecessarily. It is explained away by DRM models inherently being uncertain. However, pressure builds up to get the methods peer-reviewed. It takes a second missed disaster before the company and methods are fully discredited. This experience leads to the real inventive AI applications in DRM modelling from being adopted by a decade.

Example 2: Pressure from management to be "innovative"

The board of directors of a water authority is enthusiastic about the potential of AI for innovation in their modelling department. They ask the manager of the department questions about why they are not working with the latest technologies. To satisfy the board of directors, several data scientists are hired and pressure is put on the senior engineers to be

open to the new technologies. After years of investments, it turns out the technology is still not outperforming traditional models. However, at this point a group of data scientists is already hired and the manager wants to show a result for that decision. They decide to implement the AI system anyway with the internal story that in a few years it will probably be ready. The next year a large flood happens and the new system misses it. As a result, human lives are unnecessarily lost and the water authority switches back to their traditional system having wasted years of research funding.

Recommendations

1. **Peer-review:** It is crucial that models are always peer-reviewed so they can be openly discussed by experts, and potential shortcomings can come to the surface. This is especially important when uncommon claims are being made or innovative techniques are applied.

2. **Model validation:** Like with all DRM models, it is important to validate models to check whether they would have performed well on past events not used to develop the model. Especially when model innovations are being implemented, such validations are important and should be openly communicated. However, not all modelling situations can always be validated and validation results can be misleading without a good understanding of the context. Therefore, validation should be communicated as part of the peer-review process.

3. **Technology readiness assessments (TRA):** AI is relatively new in DRM and not suitable for every type of analysis. TRAs could be used to decide whether technologies are ready to be implemented.[19] Such assessments have been successfully applied in many industries to make systematicdecisions about the adoption of new technologies or over 40 years.[20]

19 Eljasik-Swoboda, T., Rathgeber, C. and Hasenauer, R. (2019). 'Assessing Technology Readiness for Artificial Intelligence and Machine Learning based Innovations', *DATA*. DOI: https://doi.org/10.5220/0007946802810288

20 Tomaschek, K., Olechowski, A., Eppinger, S. and Joglekar, N., 2016, July. A survey of technology readiness level users. In *INCOSE International Symposium* (Vol. 26, No. 1, pp. 2101-2117).

4. **User needs assessments:** In some cases, even if the ML system functions well, it may not address the most pressing issues faced by the intended beneficiaries. Therefore, it is good to also always do a proper user needs assessment. Human-centred design techniques offerone way to accomplish this[21].

5. **Education and awareness-building:** Guidance and educational materials about potential uses and limitations of various ML approaches should be designed and disseminated for non-technical audiences who make or influence technology adoption decisions. Some examples include: the ML for DRM guidance note,[22] and DeepLearning.AI's "AI for Everyone" online course.[23]

Suggested Readings

- Burkhardt, V. (2007) Innovator Interviews: Why the Hype? https://www.ideaconnection.com/interviews/00114-Why-the-Hype.html
- Eljasik-Swoboda, T., Rathgeber, C. and Hasenauer, R. (2019). 'Assessing Technology Readiness for Artificial Intelligence and Machine Learning based Innovations', DATA. DOI: https://doi.org/10.5220/0007946802810288
- Fernandez, S. (2019) Ethical AI: Separating fact from fad. https://www.linkedin.com/pulse/ethical-ai-separating-fact-from-fad-sheldon-fernandez/

21 IDEO (Firm), 2015. *The Field Guide to Human-centered Design: Design Kit.* IDEO. https://d1r3w4d5z5a88i.cloudfront.net/assets/guide/Field%20Guide%20to%20Human-Centered%20Design_IDEOorg_English-0f60d33bce6b870e7d80f9cc1642c8e7.pdf

22 GFDRR. (2018) 'Machine Learning for Disaster Risk Management', *GFDRR Knowledge Hub*. https://www.gfdrr.org/en/publication/machine-learning-disaster-risk-management

23 DeepLearning.AI. *AI For Everyone* https://www.deeplearning.ai/ai-for-everyone/

CHAPTER

07. Beyond Artificial Intelligence

Many of the identified issues and solutions in this resource go further than machine learning, and are rooted in how we collect disaster risk and impact data, and how we design disaster risk management projects more broadly. In other parts of this resource, we have explored potential negative impacts of machine learning techniques (bias, privacy and security, reduced transparency and explainability, and hype) and solutions for a more responsible deployment of machine learning technologies in disaster risk management. When exploring these solutions, some in our working group questioned if we are not holding ourselves to a higher standard for these new techniques compared to our standards for “traditional” practices in disaster risk management.

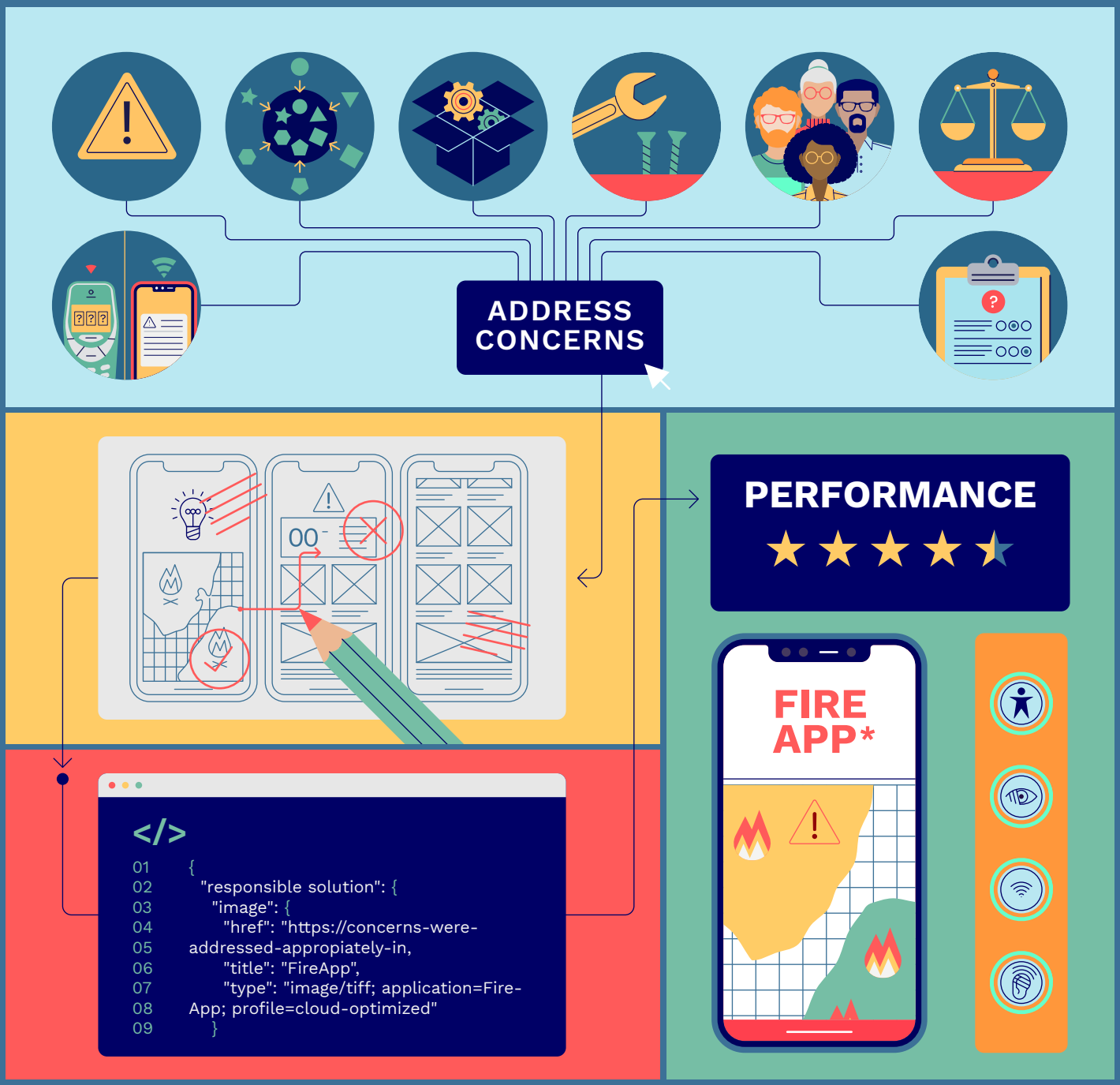
For example, we discussed how machine learning technologies reduce the role of expert judgement in the modelling process. In more traditional disaster risk management practices, we aim for a collaborative modelling process to create a feeling of ownership of the model (results) by the stakeholders. However, more often than we would like, resource constraints restrict the inclusion of stakeholders in the modelling process in “traditional” disaster risk management projects as well. Similarly, databases used for disaster risk assessment as well as outcomes of risk models can contain sensitive information that, if traced back to individuals or (targeted) minority groups, could be used for retribution by the government or the opposing side in a conflict; irrespective of the Machine Learning technologies or “traditional” practices that have been used. This also holds for the concern that the use of biased data sets may lead to the continuation and aggravation of societal inequalities. If our risk assessments, based on Machine Learning or “traditional” approaches, do not account for the disproportionate impact on vulnerable groups, how are we going to appropriately allocate resources to enhance the resilience of these vulnerable groups?

Therefore, the heightened focus on the societal impacts of AI offer us an opportunity to draw attention to some of these questions for traditional practices in the creation and use of data for disaster risk management as well. Furthermore, this report does not claim to offer the final word on these issues. Far more research and careful experimentation will be necessary to ensure that the technologies we use to make sense of and respond to the threats of disaster are fair, just, and sustainable. We would therefore want to conclude with the following call to action:

Let us hold ourselves to a higher standard and strive for a more responsible and ethical use of disaster data and design of disaster risk management projects irrespective of whether they deploy “new” AI technologies or “traditional” models and practices.

CHAPTER

08. Overarching Recommendations

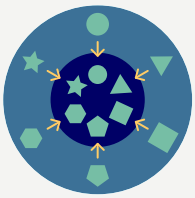


In order for machine learning technologies to be deployed in the disaster risk management context in a responsible manner, the community of experts and practitioners working on these tools urgently need to take concerns raised in this document seriously. We recommend that the following actions be taken:

1.

Proceed with caution
and conduct full life-cycle threat assessments of all new applications of AI or ML technologies. Recognize that, as a community, we haven't yet conducted enough due diligence around the potential unintended harms that our tools may cause. Give priority to alternative or traditional approaches if risks are significant or unclear.

2.



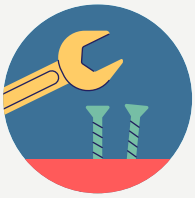
Draw on the experiences of other fields and domains.
These are inherently interdisciplinary projects and challenges. As data scientists and machine learning practitioners, we may lack necessary background or expertise to fully evaluate the potential risks of our projects. While the conversation about ethical use of machine learning in disaster risk management is nascent, there are numerous studies and cautionary examples from other areas that we can draw on when evaluating the potential consequences of these technologies.

3.

Work in transparent fashion,
in collaboration with communities and people who are represented in/by our technologies. Where possible and appropriate, support open-source and open data approaches. Provide users, decision-makers, and the public with the information they need not only to evaluate the outcomes of machine learning systems, but also to understand the limits of when and where (and in which context) the systems can be applied. Help support capacities in communities at risk of disaster

and climate impacts to cocreate machine learning projects and challenge them when necessary.

4.



Recognize the limits of technical approaches to addressing concerns.

This document has highlighted some of the tools that exist to help machine learning experts. These tools have the potential to improve our projects and to reduce many of the potential sources of harm, but many of the most urgent concerns require non-technical measures and safeguards as well.

5.

Diversify project teams.

One of the most important reasons that potential harms of machine learning projects are not caught early in development processes is that the teams producing them have overly narrow backgrounds, skill sets, and life experiences. Diversifying teams can help ensure potential negative impacts of a project are identified early and mitigated.

6.

Be aware of the trade offs and conflicts that can arise when taking measures to enhance responsible deployment of AI technologies. Recognize that actions to improve the one aspect can reduce accuracy, increase costs, or negatively impact other ethical concerns.

7.

Remember that technology is never neutral.

The design and implementation of data collection and analysis processes, especially those with close connections to society like DRM, inevitably encode preferences for some values, priorities, and interests over others. Failing to account for this in projects can serve to reinforce existing inequities and adding to the vulnerabilities of already at-risk communities.

01
02
03
04
05
06
07
08

8.

Principles are not enough.

Ethical guidelines or other sorts of recommendations for how to proceed responsibly in uncertain settings are important to helping a field establish shared norms and practices but principles are not enough. Establish and maintain warning systems, oversight, and fail-safes to protect against the most consequential threats anticipated and avenues of appeal, account, and redress/remedy for those who suffer when things do go awry in anticipated and, more likely, unanticipated ways.

Responsible AI

FOR DISASTER RISK MANAGEMENT

Working Group Summary

Deltares GFDRR Global Facility for Disaster Reduction and Recovery WORLD BANK GROUP