PEER REVIEW FRAMEWORK FOR PREDICTIVE ANALYTICS IN HUMANITARIAN RESPONSE

MARCH 2020

THE CENTRE FOR HUMANITARIAN DATA

OCHA centre for humdata

1. INTRODUCTION

Humanitarian decision-makers have called for the increased use of predictive analytics to inform anticipatory action. However, translating the outputs of predictive models into timely and appropriate responses remains a challenge for several reasons:

- First, there is no common standard for **documenting predictive models and their intended use** which highlights the critical aspects for the application of models in the humanitarian sector.
- Second, there is no common standard or mechanism for **assessing the technical rigor and operational readiness** of predictive models in the sector.
- Third, the development of predictive models is often led by technical specialists who may not consider important **ethical concerns** that the application of models in humanitarian contexts may entail.

The Centre for Humanitarian Data ('the Centre') has been working with our partners to understand the current state of model development and use in humanitarian operations. We have noted a clear desire for quality assurance of models by partners, with the Centre identified as having a unique role to facilitate a peer review process.

The following *Peer Review Framework for Predictive Analytics in Humanitarian Response* aims to create standards and processes for the use of models in our sector. It is based on research with experts and stakeholders across a range of organizations that design and use predictive models. The Framework also draws on best practices from academia and the private sector.

Quality assurance is one of three areas of focus for the Centre's predictive analytics workstream. We also work on developing new models and supporting existing partner models for use in humanitarian operations, and on community and capacity building. Learn more about our work on predictive analytics on the Centre's website: **https://centre.humdata.org**.

2. ROLES IN THE PEER REVIEW PROCESS

The Centre will lead the process and work with experts to complete the peer review. The Client will be asked to identify a single focal point for the process although different colleagues may need to be involved for each step.

The roles are:

- The Client: an organization submitting a model for peer review.
- Technical Reviewer: an individual with demonstrated expertise in data science, statistics and their application in the relevant context.
- Ethical Reviewer: an individual with demonstrated expertise in practical and humanitarian ethics.
- Moderator: a member of the Centre's predictive analytics team designated by the Team Lead on a case-by-case basis.
- Stakeholders: individuals who could be affected by the use of the model or who may use model outputs to take decisions.

2.1 The Client

The Client initiates the process through a direct request to the Centre's Predictive Analytics Team Lead. Models may be submitted for peer review by organizations developing a model, or by organizations planning to use model outputs for decisions.

2.2 Reviewers

The Centre will invite experts to submit a brief application to become a Reviewer in one of the domains (technical or ethical). Once accepted, the Reviewer will become part of a Reviewer Pool which will be managed by the Centre. The Moderator will select Reviewers based on availability and a match of skills for the model in reference. Reviewers will not be assigned to review models submitted by their own organization. The Reviewer role is unpaid.

2.3 The Moderator

The Centre's Predictive Analytics Team Lead will appoint a member of the Centre's team to act as the Moderator for each review. Once the review begins, the Moderator is the point-of-contact for the Client and the Reviewers.

2.4 Stakeholders

Stakeholders can be identified by the Client with support from the Moderator and the Reviewer. Stakeholders may be consulted on model design and use to inform the review.

3. STEPS IN THE PEER REVIEW PROCESS

The Framework consists of five steps:

Step 1: Model Submission

Step 2: Technical Review

Step 3: Implementation Plan Submission

Step 4: Ethical Review

Step 5: Final Report

The Framework aims to support two types of models:

- First, models that are fully developed and have an implementation plan defined. These are models that can respond to the requirements and specifications of the stakeholders who will implement the related actions. These models will go through the complete peer review process.
- Second, models that are exploratory. For these models, it is not possible to evaluate the ethical risks as the actions that the model would generate are not defined. In this case, we will only perform step 1 and 2 of the peer review process.

The duration for review will depend on the completeness and quality of the submission by the partner organization. A complete review process may take anywhere from one to two months.

PEER REVIEW STEPS	ROLES INVOLVED
Step 1: Model Submission Output: Model Card	Client Moderator
Step 2: Technical Review Output: Model Evaluation Matrix	Client Moderator Technical Reviewer
Step 3: Implementation Plan Submission Output: Implementation Card	Client Moderator
Step 4: Ethical Review Output: Ethical Matrix	Client Moderator Ethical Reviewer
Step 5: Final Report	Moderator

3.1 Step 1: Model Submission

In the first step, the Client submits documentation of the model. The Moderator may support the Client to document the model objectives, development, evaluation, and operational readiness. The output of this step is a completed **Model Card**.

Building upon the work done in **other sectors**, the Model Card highlights the relevant information for decision makers, including:

- The scope and the intended use of the model.
- The methodology of the model, including the reliability of input datasets, potential biases and the model assumptions.
- The evaluation process of the model and the reliability of the model output.
- The operational readiness of the model for humanitarian contexts.

The Model Card will help to standardize documentation, allowing stakeholders to compare candidate models for deployment. The Model Card is outlined below; a Google doc version of the template will be shared with the Client during the review process.

MODEL CARD

Model Summary

- Person/organization developing the model:
- Model date:
- Model version:
- Model type:
- License:
- Main contacts:
- Link (if available):

1. Intended Use

A. In-scope use cases: what is the actual and potential scope of the model? Describe the situations in which the model output is expected to be reliable.

B. Out-of-scope use cases: what are the model's limits and constraints?

C. Describe the situations in which the model output may not be reliable?

D. Model interpretation: what does the output represent?

2. Model Development

A. Details of the datasets used to build the model.

- Describe the sources of data, size and scope of the datasets.
- Is it representative of the population being sampled?
- How accurate or reliable is the training data?
- How is missing data treated?
  (e.g., exclusion, single imputation, multiple imputation)

B. What are the model assumptions and approximations?

C. Methodology - provide a description of the different analysis steps and how the input datasets are used to train the model.

3. Model Evaluation

A. Are there other similar existing models and how does this model compare?

B. Have there been any past peer reviews, such as published papers or other review exercises?

C. Is there any reference or benchmark used to evaluate the performance?

D. What are the metrics used for model evaluation? Why have these metrics been selected? Describe current model performance.

E. How does performance depend on forecast lead time?

F. What is the risk tolerance of the model? What happens if the model produces false positives, false negatives?

4. Operational Readiness

A. Is the model ready to be used to inform humanitarian response?

- Is the model kept up-to-date with the latest datasets?
- Who is responsible for model updating and/or recalibration?
- Is the model ready to be deployed? If not, what are the additional steps needed (e.g. further research, validation, updates)?

B. Has the model been developed in collaboration with operational partners? Has the model previously been used in humanitarian situations?

C. What happens if the model is inaccurate, produces false positives or false negatives? Describe expected impact according to the in-scope use cases highlighted in Section 1A above.

D. List additional considerations for the use of the model in humanitarian response.

Step 2: Technical Review

In the second step, the model will undergo a technical review based on the information contained in the Model Card. The output of this step is a completed **Model Evaluation Matrix**. The evaluation will highlight the strengths and weaknesses of the model and identify the scope within which the model should be considered for anticipatory action.

The Technical Reviewer will use the following quality criteria for the assessment:

Intended Use:

- Clarity in the definition of use-cases.
- Understanding of model's boundaries, limitations and constraints.

Model Development:

- Quality and representativeness of model's input datasets.
- Robustness of model methodology.

Model Evaluation

- Reliability of projections with respect to existing benchmarks.
- Understanding of model failures.

Operational Readiness

- Usability and timeliness of model.
- Clarity in the definition of roles and responsibilities.

In addition to these criteria, the Technical Reviewer may use additional quality standards depending on the type of the model. The Model Evaluation Matrix is outlined below; a Google doc version of the template will be shared with the Client during the review process.

MODEL EVALUATION MATRIX			
Model Card Section	Strengths	Weaknesses	Recommended Actions
1. Intended Use			
2. Model Development			
3. Model Evaluation			
4. Operational Readiness			

[1] O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group New York, NY, USA, 2016.

Step 3: Implementation Plan Submission

In the third step, the Moderator will support the Client to develop documentation of the intended use of the model. The output of this step is a completed **Implementation Plan** which summarises the concrete actions that the model will trigger or inform and clarifies the roles and responsibilities for implementation of the model in humanitarian action.

The Implementation Plan is outlined below; a Google doc version of the template will be shared with the Client during the review process.

IMPLEMENTATION PLAN

Model Use

- What model will be used?
- What is the signal expected from the model?
- Who is monitoring the model's output and who will produce a signal?
- Have the model signals been used in the past to trigger action?
- Has a risk, harm, and benefit assessment of the model been undertaken?

Model-generated Actions

- Explain when the following actions are expected to happen
  - How long before the shock will the model be used to trigger the action plan?
  - When will financing be released?
  - When will actors start implementing a response plan?
- Explain who the stakeholders are
  - Who is responsible for coordinating the response plan?
  - Who is responsible for implementing the response plan?
  - Who is responsible for monitoring the response plan?

Proposed Response Plan Activities

- Describe the intervention that is envisioned.
- What are geographic areas and the population groups that will be targeted?
- Potential number of households or people to be reached?

Step 4: Ethical Review

In the fourth step, the Ethical Reviewer will assess the ethical risks related to intended use of the model. The output of this step is a completed **Ethical Matrix**.

The Ethical Matrix has been adapted from the work of Cathy O'Neil[1] and aims to identify all stakeholders and concerns regarding the intended use of the model. For instance, a model may need to be adjusted in a scenario in which a false negative is unacceptable for affected populations or a false positive is unacceptable for a donor. The matrix will be completed in collaboration with the Client.

The Ethical Matrix is outlined below; a Google doc version of the template will be shared with the Client during the review process.

ETHICAL MATRIX				
Stakeholder/ Concern	Inaccuracy	False positive	False negative	Gaming…
Modeller				
Affected population				
Donor…				

To complete the Ethical Matrix, the Ethical Reviewer will go through the following steps:

Step 1: Identify the stakeholders in the context of the algorithm and its intended use

Step 2: Identify potential algorithm failures

Step 3: Fill out specific stakeholder concerns

Step 4: Color-code specific concerns according to priority level

Step 5: Develop recommendations to address priority concerns and reduce the overall balance of consequences to an acceptable level

A detailed description of the steps and process for completing the Ethical Matrix is provided in Annex A.

Step 5: Final Report

In the final step, the Moderator will summarize the findings from the Model Card, Model Evaluation Matrix, Implementation Plan and the Ethical Matrix into a final report. The report will be shared with the Client and Reviewers, after which it may be shared with Stakeholders.

FEEDBACK

The Centre invites individuals and organizations working in the humanitarian, academic, research and private sector to engage with us on the peer review process. Please send feedback on the framework to centrehumdata@un.org.

ACKNOWLEDGEMENTS

The initial draft Peer Review Framework was developed by Dani Poole who worked as a Data Fellow with the Centre during the 2019 Data Fellows Programme in June and July The Hague. As part of her research, Dani conducted interviews with over 20 experts including data scientists, researchers, ethicists, and decision makers spanning the humanitarian, academic, and private sectors.

The draft Framework was further detailed with input from Leonardo Milano, Evan Tachovsky, Madeline Lisaius, Stuart Campo, Jos Berens, Sarah Telford, Eleonore Fournier-Tombs, Manu Singh, and Kirsten Gelsdorf, among others. We worked with the Dutch Red Cross 510 Global Team and the Danish Refugee Council to test the draft Framework on their models. This collaboration informed the updated version released in March 2020. We appreciate the time and consideration of the many people who have contributed to this process and will continue to update the Framework as it becomes more widely used.

ANNEX A

ETHICAL REVIEW GUIDE

Step 1: Identify the stakeholders in the context of the algorithm and its intended use

For whom does this algorithm fail?

Stakeholders are actors who could be affected by the use of the algorithm. To identify who is a stakeholder, consider:

- Who benefits from this algorithm when it works as intended?
- Who would be at risk if the algorithm fails?
- Who is directly involved in the development and deployment of the algorithm?

Stakeholders may include individuals, groups of individuals (e.g. specific demographics or populations), organizations or specific teams within them (e.g. data team, program delivery team), or government agencies. The identified stakeholders are inserted in each row of the Ethical Matrix. Record them in the template provided.

Examples of stakeholders are listed below. Stakeholders that are not affected in any way by the process should not be included in the matrix:

- **Affected population** (split into subgroups if those different subgroups have different concerns)
- **Modeller**
- **Senior decision-maker** (e.g. ERC, Head of Agency/Office, etc.)
- **Host government**
- **Donor**
- **Local NGOs / response organizations**
- **International NGOs / response organizations**
- **Model reviewers**

Step 2: Identify potential algorithm failures

How can the model fail?

Based on the content in the Model Card, the Model Evaluation Matrix, and the Implementation Plan, generate an initial list of specific ways in which the algorithm can fail for different stakeholder groups. For example, the data team likely cares about accuracy, while affected people likely care more about false negatives than false positives. All three concerns — accuracy, false positives and false negatives — should be on the list.

Supplement and validate this initial list by engaging stakeholder representatives where possible and necessary. Engagement of stakeholder representatives is particularly important for models and related implementation plans that have not been through the peer review process before. Try to be comprehensive in stakeholder engagement while balancing this with the effort required for engagement. In the initial phase, wide stakeholder engagement will be required for each ethical review. Over time, less stakeholder engagement might be required as the Moderator and Ethical Reviewer can rely more on experience with similar models.

Stakeholder representatives should be identified by the Client with support from the Moderator and the Reviewer. If representatives are not available because of access constraints or other challenges, try to find people who can represent their interests (e.g., a field team member who has extensive interaction with affected people may be able to represent them).

The identified concerns are inserted in the columns of the Ethical Matrix. Record them in the template provided.

A non-comprehensive list of potential algorithm concerns includes:

- **Inaccuracy**. The model's output is inaccurate. *A cyclone impact model estimates that 100,000 people are severely affected by the event but the true number is 180,000 people.*
- **False positive**. The model produces a false alarm. *A food insecurity model wrongly estimates that a district will face a food security crisis in the next 3 months.*
- **False negative**. The model misses an event. *A food insecurity model does not catch an upcoming food security crisis.*
- **Lack of transparency**. The algorithms are not made available or are not transparent. *The model has been developed by a private company which uses a proprietary software which is not made available.*
- **Gaming**. Stakeholders can act according to the algorithm's rules to achieve an intended outcome. *An organisation systematically inflates the severity of needs of the beneficiaries they are assisting to have greater humanitarian financing.*
- **Statistical Bias**. The data used to train the algorithm does not reflect the reality on the ground. *The displacement data available to train a displacement model only covers 7 out of 10 districts because of limited humanitarian access.*
- **Systematic Bias**. The datasets used to train the algorithm don't reflect the full complexity of the reality. *A drought impact model doesn't take into account factors related to ongoing conflict and insecurity but only environmental factors in a country with ongoing conflict.*

- **Corruption**. Offering goods/services to make the algorithm present a more favorable outcome. *A private company developing an impact model for heat waves is offered compensation from a service provider to tune the parameters of the model to increase the probability of triggering financing.*

- **Politics**. The outcomes of the model are not acted upon due to competing political interests. *A model is developed without an agreed response plan and local actors are not involved in the implementation.*

- **Missing attributes**. The data used to train and validate the model has missing attributes. *The input data for population is not disaggregated by age and the number of children in the affected areas is obtained by extrapolating from the national average.*

- **Insufficient data**. Historical data on previous impacts is not available for proper testing of the model. *An epidemic model is only verified on one historical outbreak.*

- **Lack of consent**. The model doesn't align with the consent provided at data collection. *The model uses population density data from a private data provider released under a non-disclosure agreement.*

- **Privacy infringement**. The model allows individuals or vulnerable groups to be re-identified. *A flood impact model uses microdata on displacement population allowing identification of vulnerable groups.*

- **Lack of data security**. The model's inputs and outputs are accessed by a non-authorised entity. *The model is using sensitive population data that is stored in a non-secured data warehouse.*

- **Ossification**. The algorithm learns the current state of the world and predicts according to it. *Conflict and insecurity escalated in a country where a drought food security model is implemented but the model was tuned on historical data and only considers environmental factors.*

Step 3: Fill out specific stakeholder concerns

What does it mean to fail?

For each potential failure (column), go row-by-row and consider the implications for each stakeholder group. Ask: If the algorithm failed in this way (e.g., if there were many false negatives), would this stakeholder group be adversely affected? If yes, how? These implications become the cells of the Ethical Matrix. Record them in the template provided.

Some stakeholders might need to be split into subgroups if those different subgroups have different concerns (e.g., Hindu disaster victims might care more about certain kinds of failure than Muslim disaster victims). You can add rows to the Ethical Matrix accordingly and follow steps 3 and 4. Note that not all algorithm failures would affect all stakeholders, so some cells of the matrix may be blank.

This step is complete when the participants in the conversation – which ideally includes representatives from all stakeholder groups – believe that all major stakeholders and major concerns are represented in the Matrix.

If possible, the recommended approach is a 'case conference' style process with all stakeholders involved rather than an individual assessment of the ethical implications based on inputs provided by the client.

Step 4: Color-code specific concerns according to priority level

How bad is the model's failure for a specific stakeholder?

Review the cells of the Ethical Matrix one-by-one. Highlight cells in red that constitute 'existential risks', where the given stakeholder has a concern that could impair or cripple the use of the algorithm. Alternatively, a cell of the matrix that raises some ethical worries for a stakeholder would be highlighted yellow and a cell which satisfies the objectives for the stakeholders and raises no worries would be highlighted green.

Judging the magnitude of risks and prioritizing among them can require a variety of sub-steps, including but not limited to:

- Considering independently (a) the severity of a particular concern, and (b) the likelihood that it will materialize;
  - Complementary information to assess risk may include probability estimates of model errors (e.g. false negative rate). The Client may provide these numbers (and how they were derived), to support the Ethical Reviewer.
  - Catastrophic consequences may be acceptable if probabilities are negligible.
- Assessing how to rank and weigh the various objectives of the algorithm;
- Data analysis or other research to validate/measure the extent to which specific concerns are being realized;
- Deciding how the competing concerns of the algorithm's stakeholders will be balanced, usually in the form of balancing the different kinds and consequences of errors that fall on different stakeholder groups;
- Recognizing hard legal and procedural constraints that exist;
- Considering the potential consequences of using the algorithm versus not using it, i.e. of continuing the process that the algorithm would replace, including the possibility that there is no such process in place.

Ethical Matrix Template

Note: The row and column titles below are illustrative; they will vary depending on the algorithm and use case being considered.

Stakeholder/ Concern	Inaccuracy (e.g. the model underestimates the severity of a food security crisis)	False positive rate (e.g. the model predicts an escalation in food insecurity that does not actually occur)	False negative rate (e.g. the model doesn't predict an escalation in food insecurity and humanitarian response is not triggered)	
Modeller				
Affected population				

Priority levels:
- Not concerned
- Combination of severity and likelihood make us somewhat worried
- Existential risk: Combination of severity and likelihood make us very worried

Step 5: Develop recommendations to address red cells

The ultimate goal is to turn all the red cells to yellow or green, which would imply they are no longer existential risks, or that there is at least an ongoing monitoring system to alert the algorithm's users to any unacceptable problems.

In this final step the Ethical Reviewer develops recommendations for each red cell in the Ethical Matrix. What it takes to turn a red cell yellow or green varies widely – and there may be more than one solution for each. Tactics could include:

- Updating the algorithm itself (e.g., redefining the objective function, using different training and/or validation data);
- (Re)defining *how* the output of the algorithm will be used to inform decisions or resource allocation;
- Other changes to the programme or service of which the algorithm is part;
- Experiments, simulations, or other research to gauge the impact of design changes on key areas of concern;
- Developing automated monitoring systems and tools to continuously track the algorithm's performance in key areas of concern;
- Specifying purposes or settings for which the algorithm can and cannot be used;
- Explaining the algorithm to users, customers, or other stakeholders;
- Engaging with stakeholder groups so they can help decide how their concerns will be addressed.

The list above makes clear that not all red cells will be 'resolved' quickly or finally. For some risks, the best we can do is create robust monitors and stay vigilant. In any case, new concerns will evolve as community values and constraints change over time. For these reasons, the Ethical Review process can be seen as ongoing: recommendations for each red cell are carried out and updated, new stakeholders and concerns arise, and circumstances change. An Ethical Matrix should never be considered a finalized project, but an artifact of an ongoing discussion.