Applying Algorithmic Accountability Frameworks with Domain-specific Codes of Ethics: A Case Study in Ecosystem Forecasting for Shellfish Toxicity in the Gulf of Maine

Isabella Grasso
grassoi@clarkson.edu
Clarkson University
Potsdam, New York

David Russell
russeldj@clarkson.edu
Clarkson University
Potsdam, New York

Abigail Matthews
avmatthews@cs.wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin

Jeanna Matthews
jnm@clarkson.edu
Clarkson University
Potsdam, New York

Nicholas R. Record
nrecord@bigelow.org
Bigelow Laboratory for Ocean Sciences
East Boothbay, Maine

ABSTRACT

Ecological forecasts are used to inform decisions that can have significant impacts on the lives of individuals and on the health of ecosystems. These forecasts, or models, embody the ethics of their creators as well as many seemingly arbitrary implementation choices made along the way. They can contain implementation errors as well as reflect patterns of bias learned when ingesting datasets derived from past biased decision making. Principles and frameworks for algorithmic accountability allow a wide range of stakeholders to place the results of models and software systems into context. We demonstrate how the combination of algorithmic accountability frameworks and domain-specific codes of ethics help answer calls to uphold fairness and human values, specifically in domains that utilize machine learning algorithms. This helps avoid many of the unintended consequences that can result from deploying "black box" systems to solve complex problems. In this paper, we discuss our experience applying algorithmic accountability principles and frameworks to ecosystem forecasting, focusing on a case study forecasting shellfish toxicity in the Gulf of Maine. We adapt existing frameworks such as Datasheets for Datasets and Model Cards for Model Reporting from their original focus on personally identifiable private data to include public datasets, such as those often used in ecosystem forecasting applications, to audit the case study. We show how high level algorithmic accountability frameworks and domain level codes of ethics compliment each other, incentivizing more transparency, accountability, and fairness in automated decision-making systems.

CCS CONCEPTS

• **Social and professional topics → Codes of ethics**; • **Applied computing → Environmental sciences**; • **Human-centered computing → Walkthrough evaluations**;

KEYWORDS

Algorithmic accountability, ethics, ecology, forecasting

ACM Reference Format:
Isabella Grasso, David Russell, Abigail Matthews, Jeanna Matthews, and Nicholas R. Record. 2020. Applying Algorithmic Accountability Frameworks with Domain-specific Codes of Ethics: A Case Study in Ecosystem Forecasting for Shellfish Toxicity in the Gulf of Maine. In *Proceedings of the 2020 ACM-IMS Foundations of Data Science Conference (FODS '20), October 19–20, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3412815.3416897

1 INTRODUCTION

Algorithms are increasingly replacing human judgement in many important decision-making processes [34]. Algorithms influence personal decision making by suggesting where to eat, where to live, and who to date. Algorithms also control institutional decisions such as sorting resumes, determining lines of credit and interest rates, analyzing loan applications, determining prison sentencing, and sorting news feeds [35]. One consequence is a shifting of decisions away from domain experts and towards the on-the-spot choices of programmers or the output of machine learning algorithms that often absorb bias from patterns of past decisions [30]. Technologists who may not be experts in the domains of journalism, criminal justice, human resources, social science, or natural resource management are producing software that can be used to make critical decisions in these areas. Increasingly, decisions are made with a misconception that software-based decisions can be trusted to be objective and unbiased, leading to "rubber stamping" and a lack of explanation, accountability, or transparency into the how and why needed to put software results into context [41].

This increasing reliance on automated decision-making systems can have unintended consequences including development of unfair, or biased algorithms. To name a few: Amazon developed a

FODS '20, October 19–20, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8103-1/20/10…$15.00
https://doi.org/10.1145/3412815.3416897

resume sorting system that was never deployed because it downgraded resumes that included the words "women" and "women's" [21]. The Gender Shades project, a research initiative that analyzes the accuracy of facial recognition software, found that 93.6% of the errors made by Microsoft's system were of darker skinned individuals [16, 17]. There is a risk that for natural resources, reliance on algorithms for tasks such as forecasting can have unintended drawbacks as well [27, 36].

Although best practices for data and model usage under the principles of algorithmic accountability have yet to be standardized or regulated, the field of research studying fairness, accountability and transparency in algorithmic systems has grown rapidly. Mature research communities spanning industry and academia have arisen to focus on these topics beginning with Fairness, Accountability, and Transparency in Machine Learning (FATML) in 2014 [2]. Other venues arose to cover fairness, accountability and transparency issues in areas beyond machine learning, such as FATES [13], FACTS-IR [11], HWB [14], AAAI/ACM Conference on AI, Ethics, and Society (AIES) [8], FAT*/FAccT [1]. Professional societies in computing such as ACM have issued statements of transparency and accountability principles [7].

Algorithmic accountability focuses less on the selection of a single ethical standard, but rather on methods for holding a system to an ethical standard determined by domain experts. Algorithmic accountability includes an action plan for redress when things go wrong, as well as incentives for iterative development of algorithmic systems with the inevitable evolution of its intended domain [18]. Transparency refers to understanding the inner mechanisms of why an algorithm finds a particular output.

There are many incompatible ways to measure fair outcomes of algorithms depending on the principle of ethics being applied [29]. AI researcher Arvind Narayanan calls the attempt to find a single definition of fairness in computer science "a wild goose chase," and suggests that algorithms should uphold the human values in the domain for which the algorithm is used [33].

While algorithmic accountability is gaining traction in some fields, the idea has not been widely applied in environmental sciences. Ecological forecasts and projections are key tools for strong environmental policy and management [19] and are becoming increasingly important as we undergo rapid ecological change due to climate change. Ecosystem forecasters fundamentally question what will happen in the future provided different scenarios [22]. Decisions by management institutions are made daily and impact socio-environmental systems. With the rise of automation, earth and environmental science data has increased by many orders of magnitude over the last decade [15] providing ecosystem forecasters with a rich set of input data to analyze. Institutions like National Oceanic and Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA) curate large datasets collected in real time using remote sensing, which has significantly increased the power of ecological models [3]. With this data deluge, forecasters are utilizing machine learning models. High-granularity deep learning models are particularly useful given the complexity of the ecosystems [25].

Machine learning API's and packages are increasingly available in statistical software systems favored by ecologists [28]. For example, using Tensorflow's Keras library in R, ecosystem forecasters can now build a deep learning model in under ten lines of code. However, widespread use of machine learning in earth and environmental science is still fairly new. Ecosystem forecasting is also just beginning to gain momentum in the scientific literature, which provides an opportunity to set standards of algorithmic fairness, accountability, and transparency early in the ecological forecasting community [26] (Figure 1). The data itself, such as sea surface temperature, salinity, and other environmental factors, may not have the same direct ethical implications as personally identifiable data such as GPS location, photos, or purchase history. Still, ecological models do have substantial impacts on the lives of individuals. They are used by policymakers and managers to regulate businesses, manage ecosystems, and impact human health, and by industry members to make business decisions. There is a danger of an accountability deficit in the creation and use of ecological forecasts. A recent paper by Hobday et al. [27] proposes one ethical code for forecasting, including a call for openness and transparency, but codes of ethics can vary across stakeholders and through time, highlighting the need for algorithmic accountability.

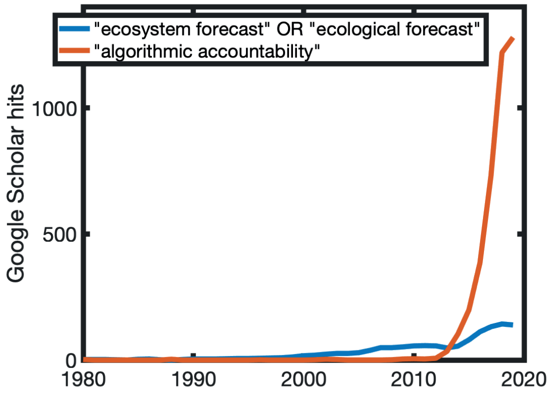


Figure 1: Literature trends through time of ecosystem forecasting and algorithmic accountability, based on yearly hits in Google scholar.

It is worth emphasizing that different stakeholders can bring different ethical considerations to a decision (e.g. valuing individual rights vs. collective good, valuing human well-being vs. valuing the well-being of all species collectively, or valuing short-term economic gains vs. valuing long-term ecosystem health). Domain experts are often trained with a code of ethics for their discipline. This is another way in which the shift of responsibility from domain experts to technologists can fundamentally change the nature of the decision-making process. Mechanisms for algorithmic accountability and transparency help to expose the critical knobs of the decision-making process to domain experts and enable them to apply the code of ethics developed for their discipline to the automated system.

Gebru et al. created templates for datasheets [23] and continues this work in Mitchell et al. [32] with model cards that encourages data creators, consumers, modelers, and machine learning practitioners to follow a standardized ethical practice and reporting process. While these templates are useful starting points, best practices should be developed at a domain specific level. For example, while ethics is included in each of these frameworks, they are less specific than the ethical principles outlined specifically for ecological forecasting in Hobday. On the other hand, algorithmic accountability is not considered in the Hobday principles. There is a need to align general principles of algorithmic accountability with domain specific codes of ethics as we illustrate in Figure 2.

In this case study, we apply Hobday et al.'s principles for ethical forecasting in conjunction with algorithmic accountability frameworks. The appropriate domain-specific code of ethics would vary with the application. For example, developers of automated decision-making systems in the domain of news amplification could apply principles of journalistic ethics [6] in conjunction with algorithmic accountability frameworks. Developers of medical applications could apply principles of medical ethics [10]. Developing and evolving these domain specific codes of ethics are an essential and long standing part of most professional fields [4, 5, 12]. In conjunction with algorithmic accountability frameworks that allow domain experts to meaningfully oversee the implementation choices of technologists, domain-specific codes of ethics offer a key way to resolve Narayanan's call to uphold the human values in each domain in which AI algorithms are used rather than search for one universal definition of fairness [33]. The same issues arise beyond AI with any complex automated decision-making system [31].

To build an algorithmic accountability framework, we utilize the Datasheets for Datasets frameworks provided by Gebru et al. [23] and also Gebru's further work on Model Cards for Model Reporting done in collaboration with a different set of coauthors, Mitchell et al. [32]. We model algorithmic accountability in ecological forecasting, specifically the use of machine learning systems by management institutions in making high stakes decisions for both human and ecological communities. We examine a shellfish toxicity model developed by Grasso et al. [25] to model algorithmic accountability in ecological forecasting, stepping through both templates. Finally, we review ethical aspects specific to ecological forecasting that are not captured in the current framework, as well as what algorithmic accountability adds to the current ecological forecasting code of ethics.

2 ALGORITHMIC ACCOUNTABILITY FRAMEWORK

The Datasheets framework includes a set of questions to be asked about any dataset. The Model Cards framework includes similar questions that encourage extensive evaluation of machine learning models [23, 32]. We pair these frameworks together for ecosystem forecasting, and apply the questions to both our dataset and model. Not all questions appropriate for this specific use. The Datasheets framework focuses on how to handle datasets that include information about individuals, often personally identifiable or private data. They ask questions focused on how data about people should be handled. However, in many other applications, like ecological forecasting, the data is not about individuals, and it is not the data that needs to be kept private. Instead the model is based on data which could be made public, but where the decisions made as a result will impact the lives of individuals. For Gebru et al. the issues of fairness revolve around ensuring that private information is protected and that groups of individuals are appropriately represented in the data set. Based on our experience, we propose some different questions for public data, which are often the basis for ecosystem forecasts. One way to think about this is as Datasheets for public data versus Datasheets for private data. Instead of looking at the way individuals are represented in our dataset, we are instead asking about how individuals are differentially impacted by the decisions that result from the use of public data. We ask who might receive an advantage or disadvantage from results of forecasts based on public data. We seek to make those trade-offs transparent and accountable to the wide set of stakeholders impacted by those forecasts. The Model Cards framework provides more broadly applicable questions, so we did not need to adapt the framework for this particular study.

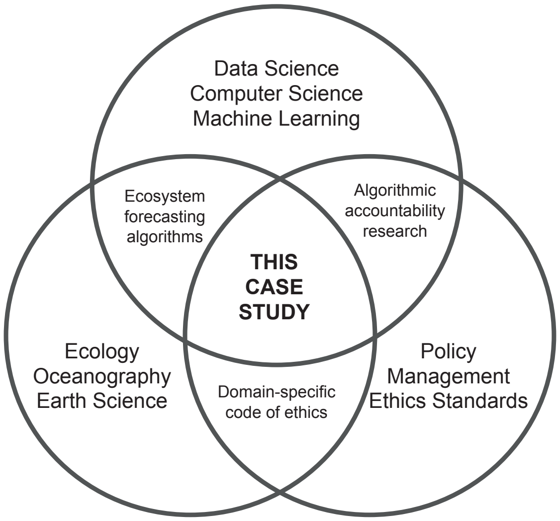


Figure 2: Conceptual visualization of how this case study fits into the context of algorithmic accountability, domain-specific codes of ethics, and ecosystem forecasting algorithm design.

In the following subsection, the headings in the boxes reorganize and regroup the questions from the Datasheets framework based on the nature of this particular case study and its intended uses. In some cases, the reduced number of questions led to some additional reorganizations. For each subheading, we provide a series of questions posed by Gebru et al., and use them to audit the ecological forecast developed by Grasso et al. We have not tested the revised questions with other datasets, but it was our goal to propose a set of questions that would work beyond this single case to ecosystem forecasting more broadly and to other similar domains that work with public data rather than private data about individuals. In the following subsections we present our updated questions and responses.

2.1 Applying and Adapting Datasheets for Datasets to Public Data

Motivation, Composition, and Collection
For what purpose was the dataset created?
Who created the dataset?
What do the instances that comprise the dataset represent?
Is there a label associated with each instance?
Is there any information missing from individual instances?
What was the sampling strategy?

The data in this case is the result of the Marine Biotoxin Monitoring Program conducted by the Department of Marine Resources (DMR) in Maine. The motivation behind this program is to protect public health. Every summer the Gulf of Maine experiences harmful algal blooms (HABs), which results in the accumulation of toxins in shellfish harvested on Maine's coastline. This program is interested in paralytic shellfish toxins (PST) in particular, which can be lethal to humans [40].

DMR staff sample the operating shellfish harvesting sites in Maine on a (semi) weekly basis. These shellfish samples are then processed by Bigelow Analytical Services (BAS) at Bigelow Laboratory for Ocean Sciences and tested for 12 toxins, specifically the PST neurotoxin saxitoxin (STX) and 11 of its derivatives (GTX4, GTX1, DCGTX3, GTX5, DCGTX2, GTX3, GTX2, NEO, DCSTX, C1, and C2) [25]. If the shellfish sampled is above a specific total toxicity threshold, the harvesting site is closed temporarily until toxicity drops to healthy levels. High toxicity measurements may also motivate higher frequency sampling in some cases. This is a long standing management procedure that is used with the knowledge of shellfish harvesters. It is required for harvesting sites to be tested by DMR.

DMR is currently responsible for the storage of the biotoxin data used in this study, which is publicly available upon request. Each instance represents a sample at a particular harvesting site. The data includes Location ID, date, species, latitude, longitude, total toxicity, and the weighted values of 12 different toxins generated by BAS. The weights, which are set by BAS, correspond with the chemical significance of the toxins in relation to the total toxicity of the sample [25].

Preprocessing, Cleaning, and Labeling
Was any preprocessing/cleaning/labeling of the data done?
Was the raw data saved in addition to preprocessed/cleaned/labeled data?
Is the code available?

For the forecasting system, the dataset was filtered for relevant data. From the full dataset (years 2014-2017 across all species and sites sampled), data was filtered to include only samples of blue mussels (*Mytilus edulis*). Each record included date, location ID, total toxicity, and the weighted values of the twelve toxins available. Rows with missing values for any of the toxicity data were omitted. Then each instance was represented as a two dimensional array containing toxicity information for each of the twelve toxins over a five week period. So there were twelve rows, with a row representing a toxin, and five columns. Each column represented the recorded toxicity level of the twelve toxins in a particular sample, with five columns indicating five samples collected (semi) weekly, so five weeks worth of data. These instances were filtered to remove any instances with gaps longer than ten days between sampling or less than five weeks worth of sampling data or missing data. The remaining instances were labeled based on the total toxicity of the subsequent week (i.e. five previous weeks of data were used to predict the toxicity of the sixth week). The instances were normalized and then binned into four distinct severity levels, with the highest level indicating a closure. The data was binned to mimic the common practice of categorizing natural events such as earthquakes or hurricanes by severity. For more information on the data refer to [25]. The code has not been released yet, but will be prior to forecast deployment.

Uses
Has the dataset been used for any tasks already?
What other tasks could the dataset be used for?
Are there any tasks for which the dataset should not be used?

The measurements that comprise the dataset are used for the day-to-day decisions made by DMR on whether to close or reopen sites due to toxicity levels. This is the first use of this dataset.

The dataset could also be used to track and understand the monitoring methods used by DMR. Further investigation could be done to study the biological and oceanographic mechanisms that associate the timing, intensity, and frequency of HABs with the timing, intensity, and frequency of the resulting shellfish toxicity outbreaks. It can also be used to prioritize the limited sampling resources on areas where HABs are predicted. In that case, it is important to consider the impact of a resulting feedback loop in which more problems may be detected in areas where more sampling is done, reinforcing the predictions themselves. This is similar to problems with predictive policing encountered in criminal justice applications [20, 24]. This and other ethical considerations regarding data use will be discussed in more detail below. The data, while available on request, is generally not made publicly available in real time because of the potential for misuse.

Distribution and Maintenance
Who is hosting and maintaining the dataset?
How can the owner/curator/manager be contacted?
Will the dataset be distributed to third parties outside of the entity?
When will the dataset be distributed?
How will the dataset be distributed?
Do any export controls or restrictions apply to the dataset?

The raw data is managed by DMR, but it is publicly available data if a request is made to DMR's Marine Biotoxin Monitoring Program. Processed data is managed by the Center for Ocean Forecasting within Bigelow Laboratory for Ocean Sciences and is not released publicly, but the code used to process the data will be made available prior to model deployment.

2.2 Applying Model Cards for Model Reporting

Model Details
Who is the person or organization developing the model?
Model Date? Version? Type?
Paper or other resources for more information?

This model was developed by the Center for Ocean Forecasting within Bigelow Laboratory for Ocean Sciences. A Keras sequential model was utilized, including an input layer with dropout, a fully connected layer with dropout, and an output layer. The optimizer used was Adam and the loss function categorical cross-entropy. All model specifications are reported in [25].

Intended Use
Primary intended uses?
Out-of-scope uses?

This model was developed to forecast shellfish harvesting site closures on a weekly, site-specific basis to aid DMR in making sampling plans and shellfish harvesters and wholesalers in making more informed business decisions. DMR lacks the resources to sample every harvesting site weekly, so this tool can help prioritize sampling sites. The intended users are both DMR shellfish sanitation and management staff as well as shellfish harvesters and wholesalers. This model is intended to be used for shellfish toxicity forecasting only. It is possible that the forecast could be incorporated as part of an ensemble forecast if other similar forecasting products become available.

Forecast deployment can advantage or disadvantage certain groups relative to each other. For example, an ecological forecast that predicted when lobsters would arrive on Maine's shoreline influenced the supply chain and impacted dealers contracts leading to unexpected costs and benefits for different stakeholders within the industry [27, 36]. This forecast is also vulnerable to out-of-scope uses that could result in similar unintended consequences.

Factors
What are foreseeable salient factors for which model performance may vary, and how were these determined?
Which factors are being reported?

The relevant factors to monitor are equitable sampling among the harvesting sites, performance with major environmental regime shifts and extreme events, and impacts to locations and businesses as a result of forecasts of toxicity even if the business is not required to close. These factors will be monitored as part of ongoing research when the forecast is deployed.

Metrics
What measures of model performance are being reported?
Why were they selected?

Overall testing and evaluation accuracy were used to tune hyperparameters, but sensitivity and specificity were also reported. For this particular model, false positives result in DMR sampling that particular site, determining that the shellfish is safe to eat and they remain open. False negatives could result in the site being overlooked and toxic shellfish to remain unsampled, therefore, for this model, false negatives indicated poor model performance. Uncertainty distributions around forecasts are also reported.

Evaluation Data and Training Data
What datasets were used to evaluate the model?

One dataset was used, as described above. The model was tested over several training-testing splits, omitting a year's worth of data as the test set and using the other three years as training data, as well as a test with 20% of the training data randomly sampled to be used as validation data. The ultimate evaluation of the forecast performance is running the model in a forecasting mode, where only data prior to a certain date have been used in training and testing, and evaluation is done using the forecast.

Quantitative Analysis
How did the model perform with respect to each factor?

Forecast performance was very good. Through the years 2015-2017, when run in a simulated forecasting mode, the forecast only failed to predict one closure level event out of a total of 49, with only two false positives. For testing sets, accuracy was high, generally greater than 95% and as high as 98%. This accuracy persisted for forecasts out to two weeks. When the forecast range was extended to three weeks, accuracy dropped sharply. For accurate forecast instances (true positive and true negative of closure-level toxicity), confidence measures were much higher than for inaccurate forecasts instances [25]. The reasons for the very strong forecast performance were not entirely clear from the forecast design, as the neural network operates on some level like a black box. This raises issues of interpretability.

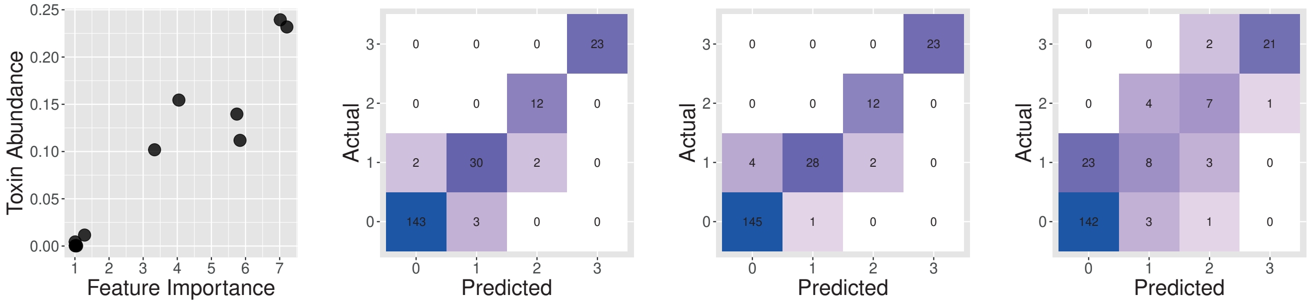
Ethical Considerations
Does the model use any sensitive data?
Was the model intended to inform decisions about human life or flourishing?
What risks may be present in model usage?

While the model uses only shellfish toxicity data, it is intended to inform decisions about human life. There is an agreement between the forecasting team and the intended users that shellfish harvesting sites are not to be shut down based solely on model output. Forecasts can be used to inform DMR's sampling strategy, as well as to help industry make proactive decisions, but site closures and openings rely on direct measurement. However, even in this context, there is information that informs decisions relevant to human life or flourishing. Whenever automated systems are interacting with human decision-makers there is risk for rubber stamping. The continued collaboration with both stakeholders and decision-makers greatly reduces this risk. There is also a risk that forecasts of high shellfish toxicity could be used to drive business to or away from a particular grower or location. This risk can be mitigated with open lines of two-way communication with stakeholders. Finally, there is a risk involved in the feedback between the forecast and the system itself. For example, DMR is intending to use this forecast to inform management strategy, meaning the forecast will impact the sampling distribution of future data to be used by the forecast, creating a feedback loop which can impact model performance [39].

2.3 Forecast Modifications

Stepping through an algorithmic accountability framework can highlight areas where the forecasting system could be modified. We discuss a few such modifications here.

First, there is the issue of unintended consequences (discussed under Intended Use). Whether a technology could be used outside of its intended use is a question that applies across science, and is one that bears repeated asking. For forecasts like this, working with forecast users as the forecast is developed, and soliciting feedback on reliability, clarity, and utility, is one way to hedge against possible unintended consequences.



(A) (B) (C) (D)

Figure 3: (A) Feature importance of random forest model using the twelve toxins to predict the total toxicity vs relative toxin abundance within samples. (B) Forecast results using all toxins. (C) Forecast results using the eight most important toxins. (D) Forecast results using six least important toxins, for comparison.

Second, there is the issue of interpretability (discussed under Quantitative Analysis). The forecasting algorithm is highly accurate, even down to the scale of individual sites, but it operates in some ways like a black box. This could lead to drops in accuracy if there are shifts in the ecology that change the underlying empirical relationships. Post hoc analysis can help uncover system dynamics that were not previously apparent, which can help zero in on the important signals within the forecasting algorithm. In the case of this forecast, performance evaluation simplified the information underlying the predictive signal in the data. DMR shellfish sanitation and management staff have observed that high levels of GTX1, GTX3, and GTX4 tend to result in closures. Correlation analysis was performed on the data and confirmed that the total toxicity of the sample was highly correlated with GTX1, GTX3, and GTX4 as well as STX and NEO. To further support this hypothesis, a random forest was trained with 50 trees. The twelve toxins were used to predict the total toxicity, and the features were ranked by importance. The three most important features were GTX3, GTX1, and NEO, followed by GTX2, GTX4, and STX. Upon further analysis it was found that these toxins are also the most abundant (Figure 3a). This information was then used to evaluate the performance of the neural network. The neural network was trained first on all twelve toxins (Figure 3b), then on the eight most important features as determined by the random forest (Figure 3c), and, for comparison, the six least important (Figure 3d). The neural network operated with very high accuracy with both twelve and eight most important toxins, and, as expected, had the lowest accuracy when trained on the six least important toxins. In addition to improving interpretability, streamlining the input data could be used to pare down the number of toxins necessary for BAS to measure, potentially making the forecast less costly and more efficient, and opening the potential for broader spatial and temporal coverage.

Finally, there are the issues around feedback mentioned earlier (discussed under Uses and Ethical Considerations). This particular forecast system requires monitoring of the coupling of human actions and forecast dynamics. When humans change their behavior in response to forecast output, then the forecast becomes a component of the system, interacting dynamically with human response. Since the intended use of this model is to inform business and management decisions, we can expect that certain changes will occur with human response to the forecast. For example, sampling distribution, site selection, or market dynamics could all respond to forecasts. This type of feedback is referred to as reflexive prediction or forecast feedback. Reflexive prediction can be either self-fulfilling or self-defeating. An example of self-fulfilling reflexive prediction is the use of algorithms for predictive policing [20, 24]. The interaction of police behavior and the algorithmic system can lead to a feedback loop of arrests. The predictions directed police to patrol certain geographic areas, which produced more arrests in that area, and in turn generated crime data for the area, therefore directing the police to the same area. An example of self-defeating reflexive prediction is epidemic forecasting, where a dire forecast can motivate behavior that slows the epidemic, or vice versa [37]. Reflexive prediction can be a result of ecological forecasts because forecasts are a key tool in developing environmental policy, which can lead to shifts in ecosystems.

A few safeguards can help this forecasting system avoid reflexivity. One can maintain a version of the forecast that is tuned consistently on the same sampling sites, regardless of how sampling sites or usage of the environment change over time. This can serve as a baseline against which to monitor trends or patterns in the full forecasting system. A second safeguard is to regularly solicit input from forecast users on how the forecasts inform their decision making.

3 DISCUSSION

In this section, we will address two questions. First, how should the algorithmic accountability framework be adapted to the domain of ecological forecasting? Second, how can algorithmic accountability add to the current code of ethics for ecological forecasting?

Algorithmic accountability frameworks are mechanisms for confirming that models are in accordance with the different concerns or even different codes of ethics brought to the table by different stakeholders. Models should reflect the human values or ethical principles defined by the domain in which they are operating, but different disciplines involved in model creation can bring different professional codes of conduct or ethics. We note that ethical principles for forecasting proposed by Hobday et al. [27] overlap with, but are not the same as, the Association for Computing Machinery

(ACM) Code of Ethics [9] or its Statement on Algorithmic Transparency and Accountability [7]. For example, both Hobday et al. and ACM include calls for openness and transparency. However, Hobday et al. include principles such as "Do not deliver forecasts that would lead to unregulated impacts on the ocean (e.g. for fisheries without clear catch limits and/or enforcement)." This is the type of ethical principle that is likely to come from certain domain experts, but might not represent the ethics of all groups of stakeholders. An algorithmic accountability framework can incorporate different codes of ethics, and can provide a path for correction if a particular code of ethics leads a forecasting program astray.

The algorithmic accountability framework is designed to be general, but adapting it with domain knowledge can be advantageous. In the case of ecological forecasting, involving stakeholders as collaborators and maintaining two-way communication can avoid some of the pitfalls that can arise around forecasting communication and unintended consequences. Ongoing dialogue can help inform decisions about the best evaluation procedures and performance metrics to use, which data collection methods are most appropriate, and any risks, or asymmetries in risks, that would not be obvious to a technologist. The most useful and reliable forecasts are likely to be those where the goals and understanding of the forecasters and the forecast users are aligned. In the domain of ecological forecasting, one could add questions on how forecasts will be communicated, and how feedback from stakeholders will be incorporated into algorithm design. Finally, the ecological forecasting code of ethics urges technologists to consider when the algorithm should be abandoned–e.g. in cases of substandard results or unintended consequences. There are even circumstances where automation might not be advantageous at all, as forecasts are often best used as decision-support tools, not decision-making tools.

The algorithmic accountability framework offers other useful perspectives to the domain of ecological forecasting. The specificity and granularity of the questions make accountability clear and traceable at each step. They force the algorithm developer to be explicit in intention and approach, and lay the groundwork for reproducibility.

Machine learning has seen explosive development and adoption over the last decade, and the lessons learned from that field can act as a guide to aid forecasters in avoiding common pitfalls. User-friendly frameworks and APIs have lowered the barriers to entry and have allowed people who are not machine learning experts to easily create applications for their domains. While machine learning represents a powerful tool in the toolbox of ecosystem forecasters, there are subtleties which could give a false sense of high performance to someone without domain expertise in machine learning. Algorithmic accountability frameworks help orient those without a background in machine learning to potential problems such as using a model on production data that deviates substantially from the training data and issues such as algorithmic and machine learning bias [38]. The development and deployment of any model or automated decision-making system should be a community effort involving technologists, domain experts, front line decision-makers, and stakeholders broadly defined. Each of these groups may bring different concerns and even different codes of ethics to the process. Applying algorithmic accountability is essential to understanding and managing the impact of these systems on individuals and ecosystems.

At a general level, a road map for accountability is a tool for identifying where algorithm development went wrong and where it can be improved in subsequent iterations. Algorithmic accountability adds incentives and tools for iterative improvement [31]. It can be safely assumed that complex models and systems always contain errors of some kind. A key part of algorithmic accountability is giving a wide range of stakeholders the tools they need to identify errors throughout the life-cycle–in design, in implementation, in deployment.

Beyond clear errors or bugs, every model simplifies reality in some ways and these simplifications can advantage or disadvantage some stakeholders. Algorithmic accountability gives stakeholders the point of reference they need to advocate for changes in models necessary to protect their interests. Fairness is difficult to define even within a domain-specific context, but the accountability framework empowers stakeholders to advocate for different fairness definitions and for outcomes they believe to be fair using evidence, rather than being faced with an impenetrable black box.

While none of the approaches were exhaustive in addressing all of the concerns of algorithmic decision-making, when used together, algorithmic accountability frameworks and domain specific codes of ethics help technologists think critically about the systems they are designing and make continuous iterative improvements on their models as a result (Figure 2).

4 CONCLUSIONS

We have used a case study to adapt algorithmic accountability frameworks to the field of ecosystem forecasting. Algorithmic accountability systems can shed light on possible improvements for domain specific ethical codes and vice versa. This is increasingly important in the domain of ecosystem forecasting due to the increased demand for ecosystem forecasts in a rapidly changing environment, coupled with the increasing accessibility of machine learning toolkits. Machine learning allows ecosystem models to forecast on finer temporal and spatial scales, which is useful to management institutions making day-to-day decisions. However, these decisions impact both human and ecological communities, and are often highly regulated. Co-evolution of ecosystem forecasting science with fairness, accountability, and ethics in machine learning as a field of research will result in more powerful forecasts, deeper understanding of biological mechanisms, and more equitable outcomes for stakeholders.

Operating under the standards of algorithmic accountability, ecosystem forecasters take ownership not just for the predictive power of their models, but also the human and environmental consequences. Algorithmic accountability practice by ecosystem modelers is a continuous process that requires attention for the entire life of the model. By adapting existing frameworks such as Datasheets for Datasets from its focus on personally identifiable private data to the types of public data often used in ecosystem forecasting application, we illuminate the importance of algorithmic accountability frameworks for all machine learning systems, even those that do not use personal information.

In this paper, we have focused on ecological forecasting. The task of building domain specific ethical codes into algorithmic accountability frameworks applies more generally: essentially anywhere that algorithms are increasingly replacing, or supporting, human decision making. Ethical codes can be highly context dependent, with implicit desired outcomes that may align with the value systems of particular groups of stakeholders. An algorithmic accountability framework compliments domain specific codes of ethics by incorporating the domain expertise of machine learning researchers into auditing a system. We demonstrate how the combination of algorithmic accountability frameworks and domain-specific codes of ethics offer a key way to answer calls to uphold fairness and human values in each domain in which AI algorithms are used rather than search for one universal definition of fairness [33]. At a high level, the complexities of machine learning, and the speed at which the field is advancing, leaves the possibility of the implementation of systems which have hidden biases. Pairing ethical codes of the field in which a system is deployed with a high level frame of reference for which to audit an algorithmic system leads to the creation of systems that uphold human values and perform effectively.

ACKNOWLEDGMENTS

We thank the Maine Department of Marine Resources and Bigelow Analytical Services for providing the data, for being collaborative partners on the forecasting model, and for helping to build links with the stakeholder community. Funding for IG and NRR was provided by NOAA grant #NA19NOS4780187. We gratefully acknowledge the input and assistance of our wider team at Clarkson University researching algorithmic accountability and transparency and the students in CS 649 in Fall 2019, especially Marzieh Babaeianjelodar.

REFERENCES

[1] [n.d.]. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT/FAT*. https://fatconference.org

[2] [n.d.]. Fairness Accountability Transparency in Machine Learning. http://fatml.org

[3] [n.d.]. National Oceanographic and Atmospheric Administration (NOAA). https://www.ncdc.noaa.gov/cdo-web/datasets

[4] 1975. National Education Association Code of Ethics. http://www.nea.org/home/30442.htm

[5] 2013. American Psychiatric Association Ethics. https://www.psychiatry.org/psychiatrists/practice/ethics

[6] 2014. SPJ Code of Ethics. https://www.spj.org/ethicscode.asp

[7] 2017. ACM Statement on Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[8] 2018. AAAI/ACM Conference on AI, Ethics, and Society (AIES). https://dl.acm.org/doi/proceedings/10.1145/3278721

[9] 2018. ACM Code of Ethics. https://ethics.acm.org/code-of-ethics/

[10] 2019. Code of Medical Ethics overview. https://www.ama-assn.org/delivering-care/ethics/code-medical-ethics-overview

[11] 2019. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. http://sigir.org/wp-content/uploads/2019/december/p020.pdf

[12] 2020. Lawyer Ethics Regulation. https://www.americanbar.org/groups/professional_responsibility/resources/lawyer_ethics_regulation/

[13] 2020. Workshop on Fairness, Accountability, Transparency, Ethics and Society on the Web. http://fates.isti.cnr.it

[14] Ricardo Baeza-Yates and Jeanna Neefe Matthews. 2019. Handling Web Bias 2019: Chairs' Welcome and Workshop Summary. In *Companion Publication of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 5. https://doi.org/10.1145/3328413.3328417

[15] Annie Brett, Jim Leape, Mark Abbott, Hide Sakaguchi, Ling Cao, Kevin Chand, Yimnang Golbuu, Tara J Martin, Juan Mayorga, and Mari S Myksvoll. 2020. Ocean data need a sea change to help navigate the warming world. *Nature* 582, 7811 (June 2020), 181—183. https://doi.org/10.1038/d41586-020-01668-z

[16] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades. http://gendershades.org/overview.html

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[18] Robyn Caplan, Joan Donovan, Lauren Hanson, and Jeanna Matthews. 2018. *Algorithmic accountability: A Primer.* https://datasociety.net/library/algorithmic-accountability-a-primer/

[19] James S Clark, Steven R Carpenter, Mary Barber, Scott Collins, Andy Dobson, Jonathan A Foley, David M Lodge, Mercedes Pascual, Roger Pielke, William Pizer, et al. 2001. Ecological forecasts: an emerging imperative. *science* 293, 5530 (2001), 657–660.

[20] Kate Crawford and Jason Schultz. 2014. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review* 55 (2014), 13. Issue 1. https://lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=3351&context=bclr

[21] Jeffrey Dastin. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women.* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[22] Michael C Dietze, Andrew Fox, Lindsay M Beck-Johnson, Julio L Betancourt, Mevin B Hooten, Catherine S Jarnevich, Timothy H Keitt, Melissa A Kenney, Christine M Laney, Laurel G Larsen, et al. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences* 115, 7 (2018), 1424–1432.

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for Datasets. *ArXiv* abs/1803.09010 (2018).

[24] Sharad Goel, Maya Perelman, Ravi Shroff, and David Alan Sklansky. 2016. Predictive Policing refererence, Combatting Police Discrimination in The Age of Big Data. https://www.fatml.org/schedule/2016/presentation/combatting-police-discrimination-age-big-data

[25] Isabella Grasso, Stephen Archer, Craig Burnell, Benjamin Tupper, Carlton Rauschenberg, Kohl Kanwit, and Nicholas Record. 2019. The hunt for red tides: Deep learning algorithm forecasts shellfish toxicity at site scales in coastal Maine. *Ecosphere* 10 (12 2019). https://doi.org/10.1002/ecs2.2960

[26] Celine Herweijer, Benjamin Combes, Pia Ramchandani, and Jasnam Sidhu. 2018. Fourth Industrial Revolution for the Earth: Harnessing Artificial Intelligence for the Earth. https://www.pwc.com/gx/en/news-room/docs/ai-for-the-earth.pdf

[27] Alistair Hobday, Jason Hartog, John Manerson, Katherine Mills, Matthew Oliver, Andrew Pershing, and Samantha Siedlecki. 2019. Ethical considerations and unanticipated consequences associated with ecological forecasting for marine resources. *ICES Journal of Marine Science* (2019). https://doi.org/doi:10.1093/icesjms/fsy210

[28] Robert Kwok. 2019. AI empowers conservation biology. https://www.nature.com/articles/d41586-019-00746-1

[29] Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. AAI/ACM.

[30] Jeanna Matthews. 2020. Patterns and Anti-Patterns, Principles and Pitfalls: Accountability and Transparency in AI. *Association for the Advancement of Artificial Intelligence (AAAI) AI Magazine* (2020). https://www.aaai.org/ojs/index.php/aimagazine/article/view/5204

[31] Jeanna Neefe Matthews, Graham Northup, Isabella Grasso, Stephen Lorenz, Marzieh Babaeianjelodar, Hunter Bashaw, Sumona Mondal, Abigail Matthews, Mariama Njie, and Jessica Goldthwaite. 2020. When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 102–108. https://doi.org/10.1145/3375627.3375807

[32] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[33] Arvind Narayanan. 2019. *21 Fairness Definitions and their Politics.* https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

[34] Cathy O'Neil. 2016. *Weapons of Math Destruction.* Crown.

[35] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group.

[36] Andrew Pershing, Katherine Mills, Alexa Dayton, Bradley Franklin, and Brian Kennedy. 2018. Evidence for Adaptation from the 2016 Marine Heatwave in the Northwest Atlantic Ocean. *Oceanography* 31 (June 2018). https://doi.org/10.5670/oceanog.2018.213

[37] Nicholas Record and Andrew Pershing. 2020. A note on the effects of epidemic forecasts on epidemic dynamics. *PeerJ* (2020). https://doi.org/10.7717/peerj.9649

[38] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 539–544. https://doi.org/10.1145/3308560.3317590

[39] D Sculley, D Holt, D Golovin, E Davydov, T Phillips, D Ebner, V Chaudhary, M Young, and J.F Crespo. 2015. Hidden technical debt in machine learning systems. In *Proceedings of the Twenty ninth Conference on Neural Information Processing Systems* (Monteal, Canada) *(NIPS '16)*. Neural Information Processing Systems, Montreal, Candada, 9. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

[40] Anke Stüken, Russell JS Orr, Ralf Kellmann, Shauna A Murray, Brett A Neilan, and Kjetill S Jakobsen. 2011. Discovery of nuclear-encoded genes for the neurotoxin saxitoxin in dinoflagellates. *PLoS One* 6, 5 (2011), e20096.

[41] Ben WagnerS. 2019. Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy Internet* 11, 1 (2019), 104–122.