An Implementable Guideline for Developing Ethical AI Systems: The Evaluation of Child Abuse and Neglect Prediction

Yuzhang Han
California State University San Marcos
yhan@csusm.edu

Aviv Landau
University of Pennsylvania
landauay@upenn.edu

Paritosh Kulkarni
University of Pennsylvania
and Columbia University
pck2123@columbia.edu

Minoo Modaresnezhad
University of North Carolina Wilmington
modaresm@uncw.edu

Hamid Nemati
University of North Carolina at Greensboro
nemati@uncg.edu

Abstract

Artificial Intelligence (AI) is becoming a crucial part of our lives. Although AI applications, such as facial recognition, autonomous driving and ChatGPT, can benefit different industries, users are more and more concerned about the ethical issues associated with AI systems. As a result, various ethics frameworks and standards have been proposed for regulating AI systems. Nevertheless, existing ethics frameworks and standards are hardly actionable or implementable for AI developers. To fill this gap, the current study proposes an actionable ethics-aware guideline for AI developers, as well as a set of quality metrics for ethical AI systems. Further, we implement the guideline using numerous AI predictive models constructed on a national big data set that estimates children's risk of experiencing abuse and neglect in the United States. Evaluation results indicate that the proposed guideline can effectively enhance the quality of predictive models in utility, ethicality and cost dimensions.

Keywords: Artificial Intelligence, machine learning, implementable ethical guideline, AI quality metrics, child abuse and neglect

1. Introduction

Artificial Intelligence (AI) has increasingly impacted many aspects of our lives in the recent decade. While we embrace AI, we are concerned about its ethical challenges, primarily how it impacts and harms marginalized communities (Noble, 2018). The research community has thus created various ethical AI frameworks that introduce ethical AI application criteria, such as fairness, transparency, and non-maleficence. Nevertheless, almost all the existing frameworks are neither implementable nor evaluable, as they provide little actionable suggestions for AI developers on satisfying those ethical criteria (e.g., Floridi & Cowls, 2022; Jobin et al., 2019). This is particularly concerning when AI is developed to address real-world problems, such as identifying child abuse and neglect (CAN) in health settings (Landau et al., 2022), or by applying welfare data to develop predictive models assessing risk for CAN (Vaithianathan et al., 2023).

CAN is a public health concern that has reached epidemic proportions, with nearly 4 million referrals to child protection services (CPS) in 2021 (Bureau, 2023). CAN is defined as any action (physical, emotional, and/or sexual) taken or failure to act by a caregiver that results in harm, potential harm, or threat of harm to a child (Bureau, 2022). A lack of "gold standard" objective assessments makes it challenging to identify CAN in clinical practice (Kuruppu et al., 2023). Furthermore, there is an agreement among clinicians that existing racial and socioeconomic biases may directly impact clinicians reporting, identification, and intervention practices for CAN (Najdowski & Bernstein, 2018). For example, the likelihood of Black children being investigated by CPS by adulthood is twice that of white children (Kim et al., 2017). These disproportionate rates have raised questions about reimagining social and child support systems and how CAN is defined in the United States (Dettlaff et al., 2020). In recent years, the widespread adoption of digital medical and welfare data has led to the development of AI tools aiming to identify CAN. These tools mainly use supervised machine learning methods to predict children's risk of experiencing various types of CAN. Prediction results can assist CPS in planning

URI: https://hdl.handle.net/10125/107205
978-0-9981331-7-1

the actions to take for the children at risk. Studying the prediction results, policy makers and researchers can also develop insights in the patterns of CAN at a location in advance, facilitating the production of timely child protection policies and research (Han et al., 2021; Negriff et al., 2023). Despite these innovations, there remains a gap in implementing ethical standards that do not cause further harm or exacerbate biases (e.g., racial, socioeconomic).

In this paper, we extend the idea of adaptive machine learning system by (Han et al., 2021) and propose a guideline for ethical AI development which consists of implementable recommendations for AI developers. We also develop a set of quality metrics evaluating the utility, ethicality and cost of AI systems. Further, we conduct a comprehensive evaluation using numerous AI predictive models trained on a national big data set, the National Child Abuse and Neglect Data System (NCANDS) (Han et al., 2021; Kim et al., 2017), that estimates children's risk of experiencing recurrence of abuse and neglect in the United States. Evaluation results indicated that the recommendations in our guideline could effectively enhance the quality of the predictive models in the above three dimensions.

2. Literature review and theoretical foundation

2.1. Ethics frameworks and guidelines for AI

Prior studies have proposed many ethics frameworks, guidelines, standards, and principles for AI systems and their development. In this paper, we denote them uniformly as *AI ethical frameworks*. Jobin et al., (2019) provided an overview of the global landscape of AI ethics guidelines by analyzing 84 documents from international companies, organizations, and governments and identified 11 common concerns. On top of the prevalence list are *transparency* (including explainability and interpretability): understanding how decisions are made by individuals; *justice & fairness* (including equality, (non-)bias): not perpetuating or amplifying existing biases and inequalities; *non-maleficence* (including security and integrity): not causing harm or negative effects to individuals, society, or the environment. Floridi and Cowls (2022) added *autonomy* to the list above as “respect(ing) the autonomy of individuals and allow(ing) them to make informed decisions”.

There are also AI regulations for auditing AI systems proposed by different countries, such as The AI and Data Act (AIDA) (Beardwood, 2023) in Canada, AI HLEG’s Ethics Guidance for Trustworthy AI in Europe (*Trustworthy AI*, 2019), and NIST AI Risk Management Framework (RMF) (Tabassi, 2023), The Organization for Economic Cooperation and Development (OECD) AI recommendation (*OECD AI Principles Overview*, 2019), and Executive Order (EO) 13960 in the US. More broadly, the European-sourced Ethics Guidelines for Trustworthy AI (*Trustworthy AI*, 2019) promotes a human-centric approach to AI and prioritizes lawfulness, ethicality, robustness/safety, human oversight, privacy, and data governance, among others, as the principles for trustworthy AI. These concerns are echoed by other agencies.

Nevertheless, it has been noted that the existing "high-level" ethical frameworks lack "substantive ethical analysis and adequate implementation strategies" (Jobin et al., 2019), and that they often lack "common aims and fiduciary duties, professional history and norms, proven methods to translate principles into practice, and robust legal and professional accountability mechanisms" (Mittelstadt, 2019). It is argued that In-depth technical instructions should accompany normative AI ethics guidelines (Hagendorff, 2020).

2.2. Measurement of AI quality

So far, little work has been done to define and measure the quality of AI systems in general. However, specific aspects of AI quality have been explored, such as machine learning (ML) models’ prediction performance (Williams et al., 2006), their cost (Gómez-Carmona et al., 2020), and their ethical properties such as fairness (Grgić-Hlača et al., 2018; Mehrabi et al., 2021) and explainability (Burkart & Huber, 2021).

Accordingly, numerous metrics measuring specific aspects of AI quality have been proposed, such as the various metrics measuring the prediction performance of ML algorithms (Tosun & Bener, 2009; Williams et al., 2006), the output quality of language models (Das & Verma, 2020), and explainability of ML models (Angelov et al., 2021; Burkart & Huber, 2021).

Based on the analysis above we identified two major gaps in research. First, the existing AI ethical frameworks lack actionable and implementable guides for developers. Second, a comprehensive, multi-dimensional metrics system for measuring the quality of AI systems is still missing. This study addresses these gaps and makes four contributions to the literature. First, we propose a novel hierarchy of metrics that measures the quality of AI systems in three major dimensions--their *utility*, *ethicality* and *cost*. Second, based on the AI quality metrics, we propose an innovative AI development guideline consisting of recommendations for AI developers, helping them develop AI systems of desired quality in the three dimensions above. The guideline is implementable, actionable and evaluable.

Third, we propose a novel, ethics-aware software framework for automated AI development satisfying user-desired quality goals. Finally, we prototype and evaluate the guideline and software framework. *To our knowledge, this study is the first to empirically analyze the impact of development decisions on AI systems' ethicality.*

3. Guideline for Developing Ethical AI Systems

3.1. Ethics-aware metrics of AI quality

We propose that the overall quality of an AI system can be measured by three metrics--*utility*, *ethicality* and *cost*. This study focuses on measuring the quality of ML predictive models.

3.1.1. Utility. On the basis of the utility of machine learning models (Zhang et al., 2020), we define the utility of an AI system as the degree of help or benefit it offers its stakeholders. In the case of ML predictive models, for example, a model has higher utility if it can make more correct predictions in any class. Decrease in utility may be caused by model underfitting or decrease in accuracy.

Equation 1 states that the utility score, U, is the average of $recall$ (also called sensitivity or true positive rate) and $specificity$ (also called selectivity or true negative rate). They are popular metrics measuring a predictive model's capability to identify positive and negative cases correctly. The purpose of incorporating recall and specificity is to allow the utility score to reflect both capabilities of AI systems.

$$U = \overline{recall + specificity} \quad (1)$$

3.1.2. Ethicality. We define the ethicality of an AI system as how much it can satisfy standard ethical criteria, such as fairness, explainability, beneficence and non-maleficence (Jobin et al., 2019). While there are many ethical criteria proposed in different AI ethical frameworks, fairness and explainability have been commonly recognized as among the main ethical concerns of AI systems (e.g., Floridi & Cowls, 2022; Jobin et al., 2019). Accordingly, we propose that the ethicality score of an AI system is mean of its fairness score, $fair$, and explainability score, exp (see Equation 2). All variables in the equation are normalized into real numbers between 0 and 1.

$$E = \overline{fair + exp} \quad (2)$$

The fairness score, $fair$, measures the group fairness of an AI system, namely, if the system delivers the same performance for different groups of its data instances. A data instance can be a human subject, such as a victim of a CAN incident. Groups of instances can be age groups or people of different income levels.

Technically, group fairness for a ML predictive model is defined as follows: A model is considered fair if the probability of a data instance "in the positive class being correctly assigned a positive outcome" and the probability of a data instance "in a negative class being incorrectly assigned a positive outcome" are the same between any groups of instances (Mehrabi et al., 2021). This type of fairness is also called the equalized odds status (Mehrabi et al., 2021). Quantitatively, a predictive model with group fairness should deliver an equal true positive rate (also called recall or sensitivity) and equal false positive rate (also called fall-out) for all the groups (Mehrabi et al., 2021). Following the above description of fairness, we propose the function of fairness score in Equation 3:

$$fair = \overline{{\sigma_{recall_i}}^{-1} + {\sigma_{FPR_i}}^{-1}} \quad (3)$$

$recall_i$ is the recall of the instance group i. ${\sigma_{recall_i}}$ is the standard deviation between recalls of all the groups, which represents the degree of spread of group recalls. Thus, the multiplicative inverse of the standard deviation, ${\sigma_{recall_i}}^{-1}$, measures the closeness of the group recalls. Similarly, ${\sigma_{FPR_i}}^{-1}$ measures the closeness of different groups' false positive rates. All the variables in Equation 6 are normalized into real numbers between 0 and 1.

exp in Equation 2, the explainability score of an AI system, measures the possibility and difficulty of explaining the reason for the AI's behavior, to gain trust of users, and to produce insights about the causes of the AI's prediction (Gilpin et al., 2018). As a crucial part of AI ethics, explainability is essential whenever a predictive model "needs to be validated before it can be implemented and deployed." (Burkart & Huber, 2021) It is also pointed out that "domains that demand explainability are characterized by making critical decisions that involve, for example, a human life." (Burkart & Huber, 2021) In our study, child risk estimation is among such domains.

In our study, exp refers to the quality score of Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), a classic explanation for ML models, provided in the AI Explainability 360 library.

3.1.3. Cost. Based on the concept of computational cost (Justus et al., 2018), we define "cost" of an AI system as the amount of resources consumed when operating the system. These resources are critical components of scalability and implementability of the system. The most common types of resource consumption considered in AI and advanced computing include the consumption of time, computational resources, and energy (Van Wynsberghe, 2021). Hence, we propose that the cost score of an AI system, C, consists of the mean of the costs in time (c_{time}), computation resources (c_{comp}), and energy (c_{enrg}), as

shown in Equation 4. All variables are normalized into real numbers between 0 and 1.

$$C = \overline{c_{time} + c_{comp} + c_{enrg}} \quad (4)$$

c_{time} is the number of nanoseconds spent in constructing the AI system, such as training and evaluating a ML predictive model.

c_{comp} is further decomposed into the consumption of three major types of computing resources in a non-GPU environment: CPU, memory and disks (Equation 5). u_{CPU} is the average CPU utilization as a percentage. u_{mem} is the average size of data (measured in bytes) stored in memory. u_{disk} is the total size of data (measured in bytes) read from and written to local disks. They were measured using the *psutil* library in Python and normalized between 0 and 1.

$$c_{comp} = \overline{u_{CPU} + u_{mem} + u_{disk}} \quad (5)$$

Finally, c_{enrg} in Equation 4, the score of energy cost, is expressed as the normalized sum of energy consumed by CPU and memory (see Equation 6). These are the major energy sources considered by prior studies in non-GPU, single computer environments (Prieto et al., 2022). e_{CPU} and e_{mem} are provided in Joules measured using the software package Intel Power Gadget 3.6.

$$c_{enrg} = e_{CPU} + e_{mem} \quad (6)$$

3.2. Ethics-aware AI development guideline

We propose an ethics-aware guideline for AI developers. The guideline consists of five recommendations which help developers construct AI systems with desired quality. The recommendations have covered the three major components of any AI system, including its input (i.e., data), procedure (i.e., algorithm) and output (i.e., result). Each recommendation addresses a unique aspect of AI development known to be able to influence the quality of AI systems substantially. Built upon the AI quality metrics above, all recommendations aim to optimize AI systems' utility, ethicality and/or cost.

3.2.1. Recommendation I. Data Representativeness. One of the major factors that could affect AI systems' ethicality, especially fairness, is data unrepresentativeness (Nargesian et al., 2021). Data provided for training and testing AI systems, such as ML predictive models, are typically samples of certain populations. A population might be underrepresented or overrepresented in data samples for various reasons, such as technical errors, bias in data collection, or an ill-design of the sampling process (Kountur, 2011).

Unrepresentativeness of data samples can cause the subsequent predictive models to deliver disparate results for data instances from different populations, posing many technical, socio-economical, and ethical challenges in different fields, such as medicine (Mac Namee et al., 2002) and finance (Fuster et al., 2022). As a result, the ethicality and fairness of models might be affected. The most popular method to resolve data unrepresentativeness during AI development is over-/under-sampling the data collection to adjust data prevalence of unrepresented groups (Bilheimer & Klein, 2010).

Recommendation I. Data Representativeness – Developers need to resolve data unrepresentativeness using proper methods, such as over/under-sampling certain groups, for improved utility, ethicality and cost.

3.2.2. Recommendation II. Data Integrity. Data integrity refers to the accuracy, completeness, and quality of data or the absence of improper modification (Sandhu, 1993). Integrity of the data used in ML processes significantly impacts output models' quality (Fujii et al., 2020). Meanwhile, the enhancement of data integrity increases the costs of data preparation, including the costs of data collection, data validation, and data cleaning (Ding et al., 2019). Classic methods for managing data integrity include data cleaning and verification, among others.

Recommendation II. Data Integrity – Developers need to control integrity of data using proper methods, such as data cleaning and verification, for improved utility, ethicality and cost.

3.2.3. Recommendation III. Feature Selection. Feature selection is a crucial part of AI development. In this process, a subset of features are discovered, which allow the AI system to achieve the desired result (Li et al., 2017).

In the utility dimension, feature selection plays an essential role, such as improving the performance of predictive models (Li et al., 2017). In the ethicality dimension, feature selection impacts ML models' fairness (Grgić-Hlača et al., 2018) and explainability (Marcílio & Eler, 2020). In the cost dimension, feature selection is crucial for reducing computational costs (Li et al., 2017) of AI development and operation. Popular classes of feature selection methods include filters, wrappers, and embedded selection (Li et al., 2017).

Recommendation III. Feature Selection – Developers need to select a proper set of features using techniques, such as filters, wrappers or embedded methods, for improved utility, ethicality and cost.

3.2.4. Recommendation IV. Algorithm Selection and Configuration. An AI system's functionality is realized by its decision algorithms. Two activities associated with the algorithms can impact the quality of AI systems significantly, the algorithm selection and algorithm configuration. Algorithm selection refers to the discovery of proper algorithms which deliver desired results. Algorithm configuration, also called

hyperparameter tuning or optimization, refers to configuring an algorithm's parameters toward the desired result.

Algorithm selection and configuration can influence the performance (e.g., Williams et al., 2006), fairness (Pessach & Shmueli, 2020), explainability (Angelov et al., 2021), and computational cost (Gómez-Carmona et al., 2020) of ML predictive models.

Recommendation IV. Algorithm Selection and Configuration – Developers need to select proper algorithms and configure them appropriately for *improved utility, ethicality and cost*.

3.2.5. Recommendation V. Result Transformation. We define result transformation as a two-stage process: (1) raw computational results of the decision algorithms are converted into information the user can understand and manage; (2) the user makes decisions and performs actions in real-world practice according to the result of (1).

While the second stage is typically beyond the scope of AI development and thus not considered in the recommendation, the first stage is necessary in many scenarios. For example, the classification algorithm could output risk scores in a continuous space, such as 0.85. However, the user could only manage categorical information, such as "high", "low", "positive" and "negative". In another example, the algorithm could output price values between $500 to $10,000. But the user could only manage a price indicator between 0 and 1. In scenarios like these, AI developers need to decide how to translate algorithm output for the user properly. Different translations can heavily impact the performance (and thus utility) of the AI system (Tosun & Bener, 2009).

Recommendation V. Result Transformation - Developers need to transform algorithm results into proper information for the user for improved utility, ethicality and cost.

3.3. Ethics-aware AI development framework

We propose a software framework that develops AI systems autonomously following the AI development guideline and end-user's requirements. The framework extends the adaptive machine learning system proposed by (Han et al., 2021).

First, as shown in Figure 1, the framework translates each recommendation in the guideline into a set of AI model specifications (details of the translation of each recommendation are provided in Section 4.2.). Model specifications from all the recommendations form a space of model specifications called the Trial Model Space.

Then, the optimal model discovery process starts, where (1) Trial Model Builder selects a different model specification from the Trial Model Space and trains an AI model, called trial model, following the specification using a small sample of the data; (2) Model Evaluator evaluates the trial model for its utility, ethicality and cost. These two steps are typically executed iteratively multiple times resulting in a part of the Trial Model Space being evaluated. The discovery process terminates when a trial model, called the optimal trial model, has been evaluated (or discovered) which satisfies the end-user's requirements about model utility, ethicality and cost. Finally, the model specification of the optimal trial model is taken to the Full Model Builder and trained into the final, full model using the full data ready for real-world deployment.

Thanks to the algorithmically efficient sizes of the data samples used in step (1), the optimal model discovery process is fast and thus can be repeated many times as an experiment in each development cycle. Furthermore, selection of the model specification to be evaluated in (1) can be guided by an intelligent algorithm, such as an instance of heuristic search or reinforcement learning in order to minimize the length of the discovery process and maximize the quality of the full model.

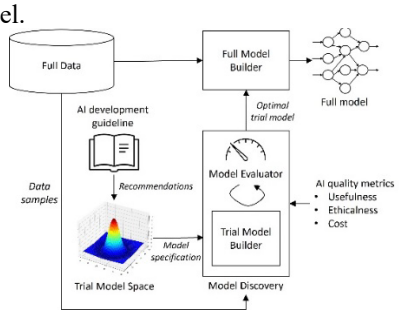

Figure 1. AI Development Framework

4. Implementation and evaluation

4.1. Settings of implementation and evaluation

4.1.1. Settings and environment. In this study, the AI development framework has been prototyped in the way that (1) the guideline with all the recommendations were implemented; (2) all the possible trial models were trained and evaluated instead of discovering the optimal trial model. Implementation and evaluation results are presented to demonstrate (1) how applying the guideline recommendations in AI development would potentially impact the quality of AI systems regarding their utility, ethicality and cost, and (2) the diverse effects of the same recommendation executed in different ways.

All the "models" mentioned in Section 4.2 refer to trial models in the AI development framework. Each trial model was trained and evaluated on two random samples from the full data of N=1,000. Each data point presented in the figures is an average of measurements

from five identical trial models constructed on different data samples.

All implementation and evaluation were conducted in Python v3.11.1. All ML components were built using scikit-learn v1.1.3. Evaluation of the prototype and guideline was performed on a Dell computer with 16 GB RAM and 12th Gen Intel i7 CPU (14 Cores) at 2300 Mhz. The operating system is Windows 11 Pro. No GPU or virtual machines were employed.

4.1.2. Real-world problem and data. The guideline was implemented to address a crucial social-technological problem, predicting the recurrence of child abuse and neglect (RCAN) in the United States. A vast national data set, the National Child Abuse and Neglect Data System (NCANDS) (Han et al., 2021; Kim et al., 2017), is available, which includes case reports of CAN incidents since 2003. In the evaluation, the framework prototype was used to develop predictive ML models that predict the risk of victims (namely, children who have experienced CAN and been included in the data set) to experience RCAN, namely, further incidents, within 24 months after their initial incidents. The problem was implemented as a supervised classification with two classes (True – victim will experience RCAN, False – will not).

The data resampled, class-rebalanced, cleaned and used in the evaluation was a portion of the NCANDS data from 2011 to 2013 (*N*=583,938). We included 20 features presenting information in three areas: (1) victim demographics (gender, race, age, etc.); (2) family information (financial difficulty, caretaker suffering from alcohol, etc.); (3) incident information (victim with prior cases, source of the report, type of abuse and neglect, etc.).

4.2. Evaluation results

Each of the subsections below describes how a recommendation was translated into trial models and evaluated in the implementation.

4.2.1. Recommendation I. Data Representativeness. We examined the variable "Victim Has Prior Case" to explore the data set between children with prior cases and children with no prior cases of abuse or neglect at CPS. Five data samples were created, each assigned a different set of group weights. The group weights are children with prior cases vs without prior cases at 1:10, 1:5, 1:1, 5:1 and 10:1 (see Figure 2). Moreover, a data sample (denoted as None) without group weights was created, representing the original data distribution among children with or without prior cases. Finally, six trial models were trained and evaluated on these data samples.

Evaluation results presented in Figure 2 demonstrate how applying Recommendation I impacts the quality of AI systems. Figure 2a indicates that applying no group weight (None) outperforms any group weights on the model's utility (outperforming by 37.3% to 40.5%) and ethicality (by 12.5% to 31.9%). By contrast, applying no group weight incurs a higher cost than three of the group weights, including 1:10, 5:1 and 10:1 (higher by 6.8% to 22.5%). This result implies that applying a group weight may not help improve the models' utility or ethicality however potentially reduces models' cost.

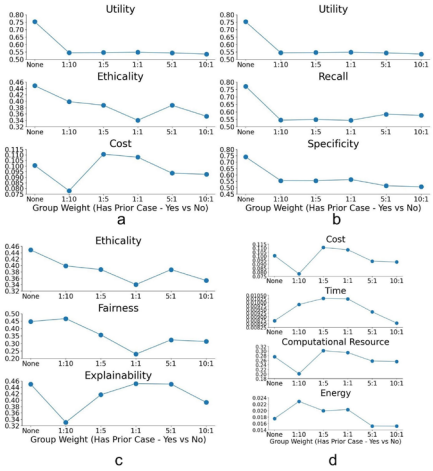

Figure 2. Impact of data representativeness

Breaking down the utility metric (Figure 2b), we see that applying any group weight would result in almost equally low recall and specificity. This means using a group weight would deduct the models' capability in detecting positive (victims with risk of being revictimized) and negative (victims without risk) instances. In the ethicality dimension (Figure 2c), the ethicality behavior of the models is influenced by their fairness more than by explainability. In the cost dimension (Figure 2d), the cost reduction of applying certain group weights resulted mainly from the decrease in computational resource consumption.

4.2.2. Recommendation II. Data Integrity. First, 10 data samples were created with different proportions of data replaced by random errors. As a result, these samples contain an increasing proportion of valid (namely, non-error) data, ranging from 10% to 100%, representing an increasing degree of data integrity. Then, a model was constructed on each data sample.

Figure 3a suggests that maintaining very low (10% and 20%) or very high (90% and 100%) data integrity potentially affects models' utility and ethicality. Meanwhile, increasing data integrity reduces models' cost monotonously.

Further, we can see in Figure 3b to Figure 3d that the top-level metrics (utility, ethicality and cost) behave similarly to their low-level constituent metrics, and that different low-level metrics in the same group also behave similarly. Therefore, the ups and downs in the

top-level metrics resulted from the changes in multiple low-level metrics.

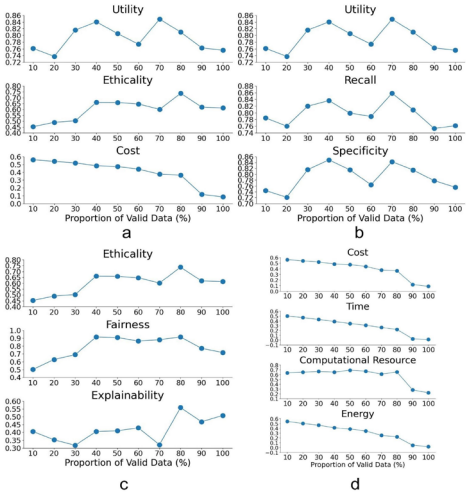

Figure 3. Impact of data integrity

4.2.3. Recommendation III. Feature Selection. As shown in Figure 4, five data samples were created, each with an increasing number of features selected and maintained, ranging from 4 (meaning 20% of features selected) to 20 (meaning all features selected). Each of these feature sets was selected using one of the most popular feature selection methods, the filter-based method, which selects the most important features based on a user-specified importance score (Ambusaidi et al., 2016). Following prior studies (e.g., Khandezamin et al., 2020; Thakkar & Lohiya, 2021), we employed the coefficients of multiple logistic regression as the importance scores for the features. In this scenario, the point at 8, for example, represents selecting eight features with the highest absolute values of their regression coefficients.

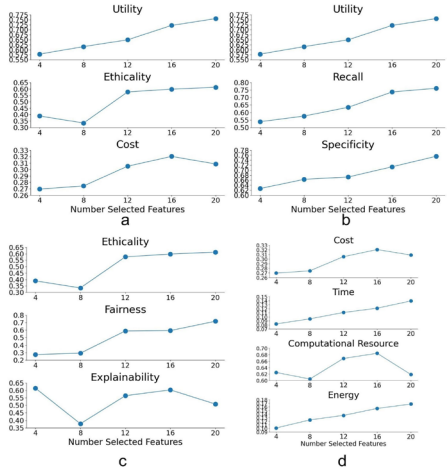

Figure 4. Impact of feature selection

We did not adopt the other major feature selection method, the wrapper method, for its potentially higher computational cost due to the invocation of model training when evaluating each candidate feature set (Li et al., 2017). The wrapper method would slow down the optimal trial model discovery in our AI development framework (see Section 3.3), making the discovery process impossible to be executed easily and frequently as an experiment in a development cycle. The embedded method was not adopted either because it requires developing a unique feature selection process specific to the choice of the prediction algorithm selection (Li et al., 2017). Since our AI development framework is expected to accept a wide range of prediction algorithms, the embedded method would increase the development cost hugely.

Figure 4a suggests that, in general, selecting more features could increase models' utility and ethicality as well their cost. In the utility dimension (Figure 4b), selecting more features improves both recall and specificity, namely, models' capabilities of detecting positive and negative instances. In the ethicality dimension (Figure 4c), selecting more features only benefits model fairness, not explainability. In the cost dimension (Figure 4d), the growth of cost can be more attributed to the growing consumption of computational resources and energy as the feature count increases.

4.2.4. Recommendation IV. Algorithm Selection and Configuration. Five models were constructed with supervised ML algorithms commonly applied in many fields (Jiang et al., 2020), including Artificial Neural Network (ANN), Gradient Boosting (GB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) (see Figure5). Their hyperparameters were determined through a classic grid search with five-fold cross validation maximizing the utility score.

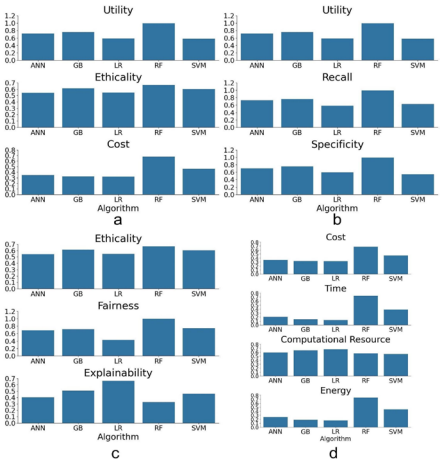

Figure 5. Impact of algorithm selection and configuration

We can see in Figure 5a that these algorithms have a diverse impact on models' utility, ethicality and cost. Among them, the ensemble algorithms, GB and RF, outperformed the others on utility (outperforming by

5.5% to 70.8%) and ethicality (by 1.6% to 22.6%). However, RF also incurred the maximum cost.

In the utility dimension (Figure 5b), algorithms tend to have the same effect on recall and specificity, resulting in the same pattern in these two metrics. However, in the ethicality dimension (Figure 5c), some algorithms have opposite effects on fairness and explainability, such as RF (high in fairness but low in explainability) and LR (low fairness, high explainability). In the cost dimension (Figure 5d), algorithms pose opposite effects between timing/energy cost and cost of computational resources. For example, algorithms of the highest timing/energy cost, RF and SVM, present the lowest cost of computational resources.

4.2.5. Recommendation V. Result Transformation. We chose to optimize models' result transformation by tuning their decision threshold. Five models were constructed with an increasing decision threshold, ranging from 0.1 to 0.9 (see Figure 6). A decision threshold of 0.1, for example, offers the highest chance of making positive predictions. It means that data instances estimated to have a higher than 0.1 likelihood to be positive will be predicted positive.

It can be seen in Figure 6a that tuning the decision threshold could pose opposite, close-linear impacts on models' utility and ethicality. A decision threshold of 0.5 results in the maximum of utility and minimum of ethicality at the same time. Meanwhile, higher threshold values (meaning higher difficulty for positive prediction) tend to introduce higher costs.

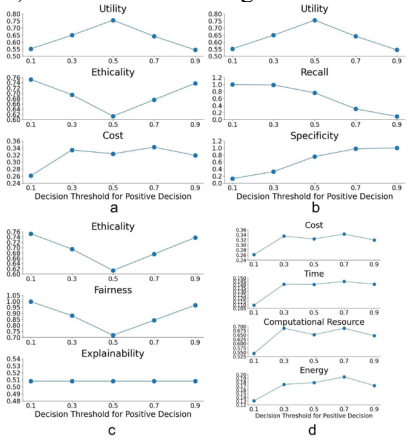

Figure 6. Impact of result transformation

In the utility dimension (Figure 6b), increase of decision threshold leads to a decline in models' recall or capability of making (correct) positive predictions. Simultaneously, it also results in a rise of models' specificity or capability of making (correct) negative predictions. Both effects contribute to the roof-shape pattern of the utility score.

In the ethicality dimension (Figure 6c), decision thresholds only impact models' fairness, not explainability. Using either an extremely high or low threshold would improve fairness. Finally, in the cost dimension (Figure 6d), using the lowest threshold (0.1) minimizes models' costs in time, computational resources and energy.

5. Discussion and implication

Results of the evaluation have revealed several insights. First of all, *applying the proposed guideline can enhance the quality of AI systems in each of the utility, ethicality and cost dimensions*. For example, applying Rec. IV. Alg. Sel. & Con. with the Random Forest (RF) algorithm resulted in the globally maximum utility score of 0.992 in the entire evaluation (see Figure 5a), which is significantly higher than the global mean utility of 0.692 (t-test result: $t(156.19) = 27.505, p < 2.2e-16$). Similarly, applying Rec. V. Res. Tran. with decision threshold = 0.1 resulted in the global maximum ethicality of 0.753 (Figure 6a), which is significantly higher than the global mean ethicality of 0.557 (t-test result: $t(4.908) = 4.335, p = 0.008$). Furthermore, applying Rec. I. Data Rep. with prior case vs no prior case = 1:10 resulted in the global minimum cost of 0.078 (Figure 2a), which is significantly lower than the global mean cost of 0.314 (t-test result: $t(7.847) = -9.438, p = 1.486e-5$)).

Second, *effects of applying the guideline are complex*. *To improve AI quality, the guideline has to be applied in well-considered ways.* It might benefit one quality metric while sacrificing the other metrics. For example, as mentioned above, applying Rec. IV Alg. Sel. & Con. with Random Forest (RF) produced the globally maximum utility (Figure 5a). However, the same recommendation also produced a high cost of 0.679, which is higher (but not statistically significantly) than the cost of any other algorithm, such as Support Vector Machine (SVM) with the second highest cost of 0.46 (t-test result: $t(7.028) = 1.772, p = 0.1196$). These results are aligned with the literature. Fu et al. (2022) claim that "fair" algorithms, which require impact parity, may not always have the anticipated results in the long run, especially in profit-maximizing firms.

Furthermore, *applying the guideline has diverse effects on the low-level quality metrics*. For example, both recall and specificity in the utility dimension received a positive influence from Rec. III. Fea. Sel. (Figure 4b). But the same metrics received opposite influences from Rec. V. Res. Tran. (Figure 6b).

With their enhancing effect on AI quality, the proposed guideline and software framework offers a powerful tool for improving AI development efficiency, productivity, and product quality. Thanks to its ethics awareness, the guideline would help introduce more and

more advanced AI applications into fields holding high ethical standards for IT products, such as healthcare and social work. On the other hand, researchers concerned with developing AI for real-world problems suggest that it must be used carefully, ethically, and in suitable scenarios. In response, clinicians and researchers working with children suffering from CAN have the opportunity to adapt, develop, and extend ethical guidelines, such as those aimed at reducing racial or socioeconomic biases.

6. Conclusion, limitation and future work

This paper provides a first-step guideline for developing AI systems to estimate the risk of child abuse and neglect (CAN). The guideline was implemented, prototyped and evaluated to demonstrate its impact on utility, ethicality and cost of AI systems. Results show that the proposed guideline and software framework were able to enhance the quality of ML predictive models in each of the three dimensions significantly. Meanwhile, the guideline influences model quality in diverse ways, making it an interesting research topic to further explore the utilization of the guideline in specific scenarios.

One of the limitations is that, due to the space limit of the paper, we are not able to present further evaluation of the proposed guideline to demonstrate, for example, discovery of optimal models and comparison between selected and baseline models with full data. These evaluations will be presented in the future. Future research should also include the participation of domain experts with lived and professional experience, such as caregivers, social workers, and youth, in designing and evaluating AI for CAN prediction. Specifically, we recommend that CPS survey these parties and use their requirements as limiting conditions in the optimal model discovery process in Figure 1 in addition to the AI quality metrics. In this way, the framework would be able to produce AI systems meeting custom quality standards required by its stakeholders. Furthermore, researchers should continuously examine the databases used and review the ethical metrics incorporated in the model development. Finally, stakeholders using AI models developed for real-world problems, such as CAN, should be responsible for their final decisions and continuously discuss the harms that can potentially occur when using these innovative technologies.

7. References

Ambusaidi, M. A., He, X., Nanda, P., & Tan, Z. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Transactions on Computers, 65(10), 2986–2998.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424.

Beardwood, J. (2023). Heads Up: The Companion Document To The Canadian Artificial Intelligence And Data Act—AIDA Companion provides answers to some key questions but then raises others. Computer Law Review International, 24(3), 65–72.

Bilheimer, L. T., & Klein, R. J. (2010). Data and measurement issues in the analysis of health disparities. Health Services Research, 45(5p2), 1489–1507.

Bureau, C. (2022). Child Welfare Information Gateway. https://www.childwelfare.gov/topics/can/defining/

Bureau, C. (2023). Child Maltreatment. https://www.acf.hhs.gov/cb/data-research/child-maltreatment

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245–317.

Das, A., & Verma, R. M. (2020). Can Machines Tell Stories? A Comparative Study of Deep Neural Language Models and Metrics. IEEE Access, 8, 181258–181292.

Dettlaff, A. J., Weber, K., Pendleton, M., Boyd, R., Bettencourt, B., & Burton, L. (2020). It is not a broken system, it is a system that needs to be broken: The upEND movement to abolish the child welfare system. Journal of Public Child Welfare, 14(5), 500–517.

Ding, X., Wang, H., Su, J., Li, Z., Li, J., & Gao, H. (2019). Cleanits: A data cleaning system for industrial time series. Proceedings of the VLDB Endowment, 12(12), 1786–1789.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine Learning and the City: Applications in Architecture and Urban Design, 535–545.

Fu, R., Aseri, M., Singh, P. V., & Srinivasan, K. (2022). "Un" fair machine learning algorithms. Management Science, 68(6), 4173–4195.

Fujii, G., Hamada, K., Ishikawa, F., Masuda, S., Matsuya, M., Myojin, T., Nishi, Y., Ogawa, H., Toku, T., Tokumoto, S., & others. (2020). Guidelines for quality assurance of machine learning-based artificial intelligence. International Journal of Software Engineering and Knowledge Engineering, 30(11n12), 1589–1606.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5–47.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89.

Gómez-Carmona, O., Casado-Mansilla, D., Kraemer, F. A., López-de-Ipiña, D., & García-Zubia, J. (2020). Exploring the computational cost of machine learning at the edge for human-centric Internet of Things. Future Generation Computer Systems, 112, 670–683.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.

Han, Y., Modaresnezhad, M., & Nemati, H. (2021). An Adaptive Machine Learning System for predicting recurrence of child maltreatment: A routine activity theory perspective. Knowledge-Based Systems, 227, 107164.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: A brief primer. Behavior Therapy, 51(5), 675–687.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Justus, D., Brennan, J., Bonner, S., & McGough, A. S. (2018). Predicting the computational cost of deep learning models. 2018 IEEE International Conference on Big Data (Big Data), 3873–3882.

Khandezamin, Z., Naderan, M., & Rashti, M. J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. Journal of Biomedical Informatics, 111, 103591.

Kim, H., Wildeman, C., Jonson-Reid, M., & Drake, B. (2017). Lifetime prevalence of investigating child maltreatment among US children. American Journal of Public Health, 107(2), 274–280.

Kountur, R. (2011). The ethical issue of response bias in survey data collection and its solution. International Forum Journal, 14(2), 55–60.

Kuruppu, J., Humphreys, C., McKibbin, G., & Hegarty, K. (2023). Navigating the grey zone in the response to child abuse and neglect in primary healthcare settings. Children and Youth Services Review, 107029.

Landau, A. Y., Ferrarello, S., Blanchard, A., Cato, K., Atkins, N., Salazar, S., Patton, D. U., & Topaz, M. (2022). Developing machine learning-based models to help identify child abuse and neglect: Key ethical challenges and recommended solutions. Journal of the American Medical Informatics Association, 29(3), 576–580.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 1–45.

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. Artificial Intelligence in Medicine, 24(1), 51–70.

Marcílio, W. E., & Eler, D. M. (2020). From explanations to feature selection: Assessing SHAP values as feature selection mechanism. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 340–347.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1–35.

Najdowski, C. J., & Bernstein, K. M. (2018). Race, social class, and child abuse: Content and strength of medical professionals' stereotypes. Child Abuse & Neglect, 86, 217–222.

Nargesian, F., Asudeh, A., & Jagadish, H. (2021). Tailoring data source distributions for fairness-aware data integration. Proceedings of the VLDB Endowment, 14(11), 2519–2532.

Negriff, S., Lynch, F. L., Cronkite, D. J., Pardee, R. E., & Penfold, R. B. (2023). Using natural language processing to identify child maltreatment in health systems. Child Abuse & Neglect, 138, 106090.

Noble, S. U. (2018). Algorithms of oppression. In Algorithms of oppression. New York University Press.

OECD AI Principles overview. (2019, May). https://oecd.ai/en/ai-principles

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. ArXiv Preprint ArXiv:2001.09784.

Prieto, B., Escobar, J. J., Gómez-López, J. C., Díaz, A. F., & Lampert, T. (2022). Energy Efficiency of Personal Computers: A Comparative Analysis. Sustainability, 14(19), 12829.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Sandhu, R. S. (1993). On Five Definitions of Data Integrity. DBSec, 257–267.

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0).

Thakkar, A., & Lohiya, R. (2021). Attack classification using feature selection techniques: A comparative study. Journal of Ambient Intelligence and Humanized Computing, 12, 1249–1266.

Tosun, A., & Bener, A. (2009). Reducing false alarms in software defect prediction by decision threshold optimization. 2009 3rd International Symposium on Empirical Software Engineering and Measurement, 477–480.

Trustworthy AI. (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Vaithianathan, R., Cuccaro-Alamin, S., & Putnam-Hornstein, E. (2023). Improving Child Welfare Practice Through Predictive Risk Modeling: Lessons from the Field. In Strengthening Child Safety and Well-Being Through Integrated Data Solutions (pp. 115–126). Springer.

Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1(3), 213–218.

Williams, N., Zander, S., & Armitage, G. (2006). A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Computer Communication Review, 36(5), 5–16.

Zhang, Y., Bellamy, R., & Varshney, K. (2020). Joint optimization of AI fairness and utility: A human-centered approach. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 400–406.