Ethics & Bias in AI Coursework

Nam Le
jssb25
1650 words

I. Introduction

This report focuses on a machine learning task for emotion recognition on human face images. Whilst not novel, it is a useful tool in a variety of applications, such as in marketing, human-robot interactions, healthcare, and security [1].

A. Description of tasks

This is a classification task which predicts the emotion of a person from an image of their face, and outputs the emotion(s) associated with it.

The range of possible emotions is typically decided by a dataset, though we aim to pick a dataset which contains the emotions of: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. According to Plutchik's wheel of emotions, a widely accepted model in discrete emotion theory, these are the universally recognized basic emotions [1].

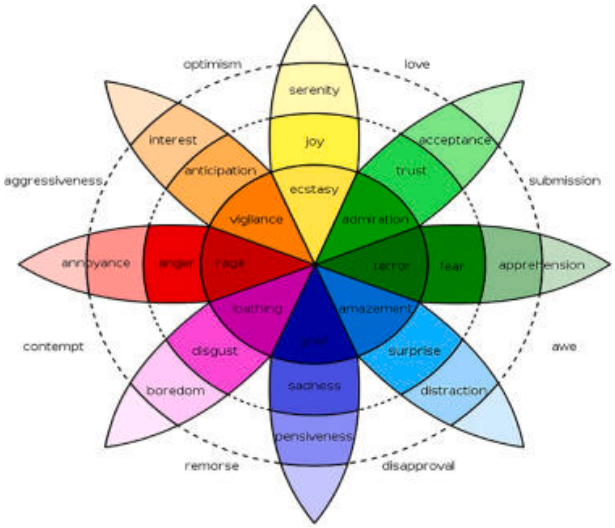

Figure 1: Plutchik's wheel of emotions, with the base emotions as well as their amplified/attenuated versions. Intensity increases towards the center and vice versa [2].

We do not use the full range of emotions as it is not practical to train a model to recognize all of them. The base emotions are sufficient for most applications.

For the output format, we want a probability distribution over the classes. This has the benefit of giving more information than single class outputs, can be used to calculate a confidence score, is easily transformed into a positive/negative/neutral output and is easily done (softmax layer). With this, we allow for more model flexibility, which is crucial for deployment to various organization and/or be made accessible via open sourcing.

The dataset used for training is custom-made from online datasets, and the process of producing it will be elaborated in section III.

B. Ethical impact, Value Sensitive Design, and use case

AI ethics refers to a set of values and principles that guide the responsible use of AI technologies [3], whereas AI bias refers to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals" [4]. An ethical impact assessment aims to root out any potential biases in the model by considering stakeholders' values/principles, allowing for fair outcomes for all.

To do this, we employ the use of Value Sensitive Design (VSD). Value in VSD refers to value, defined as "what a group of people consider important in life" [5], hence VSD is a methodology which considers the values of both direct and indirect stakeholders in the design of technology.

VSD involves three investigation steps.

1. Conceptual: Identifies stakeholders, what values they hold, how they are affected, discuss trade-offs between values.
2. Empirical: Using quantitative/qualitative methods to expand on the concepts found in the previous step.
3. Technical: Analysis of existing technological mechanism on, and proactive design supporting of human values.

A potential use case of this model is for healthcare surveillance system [6]. In hospitals, the model could be used to detect signs of depression or anxiety in patients, and alert the clinic to administer medicine. Compared to traditional methods, such as human observation, the model could be less prone to error and be more available to patients.

In the hospital example, we must make sure the model is only used on images of patients who have given explicit consent, but how do we get the consent of people who are not mentally well enough to make decisions? Issues like this is beyond our scope of responsibility, but it is important to consider nonetheless.

In the following sections (section II and III), we focus solely on the healthcare use case in hospitals.

II. Ethical impact assessment informed by VSD

Stakeholders	Values	Potential risks/harms
Healthcare providers, such as a doctor or a nurse (Direct)	Respect for human autonomy - • The model's decisions should not be absolute, and it's limits needs to be recognized • The model should act as an aid to help healthcare providers make their decisions, providing an additional point of view	A misdiagnosis might be given, in which case the healthcare provider might be forced to blindly follow the model's decision. If this results in harm to the patient, the healthcare provider could be held accountable for medical negligence, potentially leading to revocation of their medical license.
Patient (Indirect)	Privacy, informed consent - • A patient should have absolute authority over their data • Hospitals/medical institutions are expected to not record, store and process data unless given explicit consent	Personal data stored in hospital databases could be breached and leaked. Bad actors could then use this data for nefarious purposes, such as identity theft or blackmail. In rare cases, institutions could (mis)use this data for their own benefit, such as selling it to third parties despite GDPR regulations.
Medical institution (Indirect)	Human welfare, freedom from bias, efficiency - • Existing system must work well for all individuals despite their differences to minimize casualties • All forms of bias must be prevented to ensure equal treatment of all patients • The model should be able to provide diagnoses in a timely manner	A biased diagnosis, whether it be due to the model's false output or misinterpretation of the model's output by a healthcare provider, could lead to lives lost. The same applies to a delayed diagnosis. The institution could then face lawsuits, have to compensate patients financially, or even have their funding rescinded. In less grave situations, the institution's reputation could be damaged.
Pharmaceutical Company, model output data consumer (Direct)	Accuracy, environmental sustainability - • Production and distribution chains could be disrupted based on the model's output data. • Overproduction of medicine could also lead to resource wastage as well as environmental pollution	Bad quality data could cause the company to produce and distribute the wrong amounts/types of medicine, leading to a loss of profit. The company could also be held accountable for any environmental damage caused by overproduction.
Insurance companies (Indirect)	Explicability - • The outcome of the model must be clear to prevent any ambiguity as to whether or not insurance must cover for a client	A correct, but unexplainable diagnosis could lead to a refusal to cover the patient's medical expenses. Disputes over this would deteriorate the relationship between all parties involved.
Government institution, regulators and policymakers (Indirect)	Trust, accountability - • A government's aim is to maintain the trust of its citizens, and so it must regulate use of the technology • The government must also be partially accountable for the technology's use and its consequences	Any technologies, in its inception, are prone to misuse [7]. If the government fails to regulate the technology, lots of inhumane Even with regulation, if the technology is misused, the government has a duty to take some responsibility for the consequences.

Table 1: Ethical impact assessment using VSD in a hospital setting.

The values used in Table 1 are from both traditional [5] and AI-specific [8] sources.

The values in Table 1 have been acquired purely through conceptual investigations, which are prone to researcher bias and tend to misrepresent the stakeholders' actual views [9].

Only after closer examination of the stakeholders' understandings, contexts, and experiences through empirical investigations [10], can we update the previously acquired values and guarantee their trustworthiness. This can be done using traditional methods such as interviews and surveys, or more VSD-specific methods such as participatory design and envisioning cards [9].

An important step of the process is analyzing value conflicts, and deciding which stakeholders' needs should be prioritized (known as the dams-and-flows method [11]).

One value conflict is between patients' need for privacy and the insurance companies' desire for explicability. In order to explain the decisions made by the model, the insurance companies will need access to the patient's data, which is a clear violation of the patient's privacy. An approach to resolve this is to have the model explain its decisions in a general but informative way, like providing a descriptive summary of the patient's condition.

Another conflict is between the medical institution's need for efficiency and the healthcare providers' want for autonomy. Healthcare providers require lots of time and resources to make decisions, which would make diagnosis and treatment slower. The institution could use this time and resources to treat more patients instead. This is a difficult tension to resolve, but we believe that the model should be used as a tool to assist the healthcare providers, not replace them. As such, the healthcare providers' interest should be prioritized.

Whilst emotion recognition technology has benefits in diagnosing mental health issues, using it for crime prevention and/or in public spaces is a completely different matter. From past research, it is clear that the technology is not accurate enough to be used in these scenarios. Even if it were, using it would be ethically questionable (see [12], [13]).

III. Recommendations & Considerations

From the ethical impact assessment in Table 1, we have identified additional features which should be built on top of the model to satisfy the stakeholders' needs.

These include, but are not limited to:

- Mandatory approval of decisions by human overseers
- GDPR-compliant data handling
- Overviews of performance and bias
- General, descriptive explanations of outcomes

We also have to be prepared to update the model in an integrative manner during the design process, as results from our empirical investigations may reveal new values and tensions that needs to be addressed.

Deployment should also be rolled out in a controlled manner. By releasing the model to a small group of hospitals first, we can deal with any potential emerging biases and other issues before they become widespread.

A. Datasets

Choosing a dataset is the most important part of the process, especially in avoiding bias. Our goal is to have a dataset that:

- Contains at least 7 of 8 emotions mentioned in Figure 1
- Is reliable, meaning accurate labels and little noise [14]
- Contains GDPR-compliant [15] data collected ethically
- Contains realistic, and preferably medical data
- Is representative of the user base of the technology

This last point is the hardest to handle, as most facial emotion recognition (FER) datasets simply are not representative. This can be seen in Figure 2, Figure 3, and Figure 4.

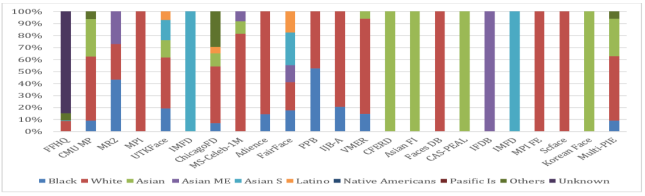

Figure 2: Racial composition in common FER datasets with "White" (in red) being the majority class [16].

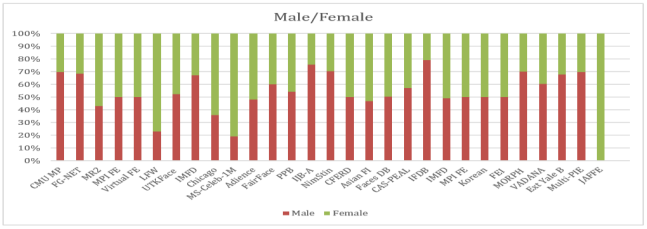

Figure 3: Gender composition in common FER datasets with "Male" (in red) being the majority class [16].

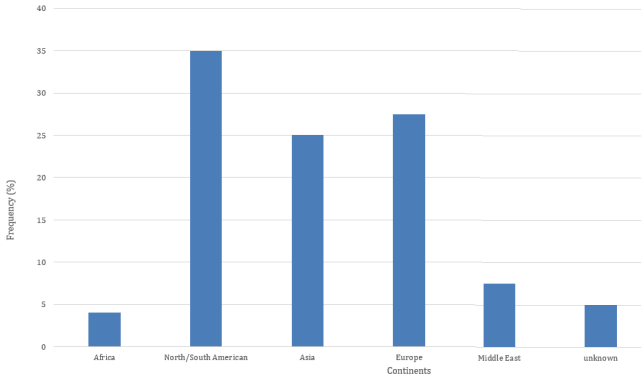

Figure 4: Source distribution of FER datasets. North/South America, Asia and Europe make up 87.5% of all frequencies [16].

Some datasets are made to be representative, however these tend to be much smaller compared to mainstream ones like AffectNet. Hence training a model solely on these is not a viable option.

To deal with this issue, we decided on aggregating multiple datasets to create a more representative one. We chose this option since it ensures a decent quality of data, whilst taking up much less time and resources than creating creating a dataset from scratch. We will document our dataset using the structure given in [7], but removing some sections like "Data Collection Process" and "Data Preprocessing".

Algorithms in [16] will be used to measure the bias in our new dataset.

B. *Risk and bias mitigation measures*

When training the model, we resort to the process in Figure 5.

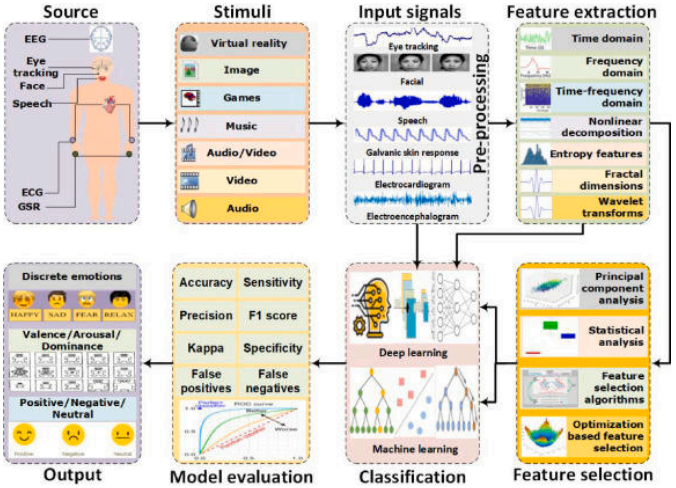

Figure 5: Process of training an emotion recognition model to be used for our model from choosing a dataset to outputting a verdict [1].

For risk and bias mitigation during the last four steps of Figure 5, we look to Figure 6.

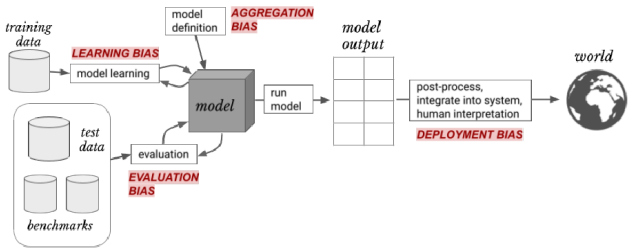

Figure 6: Potential areas of bias during the model training and implementation process [17].

We will now analyze a source of bias shown in Figure 6.

Learning bias is an issue where the model is optimized for a specific metric, like accuracy, while disregarding fairness metrics like f1 score, precision or recall.

To mitigate this, we can assign different weights to metrics depending on their importance. For our use case (medical diagnosis), precision would be valued more, as a false-negative output could be life-threatening. Pharmaceutical companies would be most interested in mitigating this bias, as they require the model's data to be as accurate as possible.

Other sources of bias can be mitigated in a similar and relatively simple manner, as long as model engineers are aware of them.

Another method of mitigation is seen in [18], where the model is tested on a predefined groups, such as men and women, light and dark skinned individuals, etcetera. Afterwards, the performance for each group is listed out for comparison purposes. An unbiased model should perform equally well on every group. Note that while useful, this is not a complete solution, as it is impossible to test the model on every possible group. This is useful for medical institutions as they require the model to be as unbiased as possible for the equal treatment of all patients.

After training the model, we aim to document the process in a model card, which has the added benefit of increasing transparency [19]. This would greatly benefit policymakers as they can quickly grasp key information about the model, allowing more informed decision-making on how to regulate such technologies. Greater transparency would also help with explainability, which insurance companies are interested in.

C. *Critical assessment and limitations*

The recommendations given in this section aims to provide an unbiased and fair model using well-researched methodologies, however there are clear limitations.

The main issue is the lack of empirical investigations. The values identified in Table 1 are purely speculative, and may not reflect the actual views of the stakeholders. Seeing as this entire section is based on these values, the entire process may be flawed. There is a simple solution to this, however, which is to carry out empirical investigations and update our process accordingly.

An issue which is not as easily solved is with regards to the dataset. Aggregating multiple datasets for better representation is a good idea, but it is not perfect.

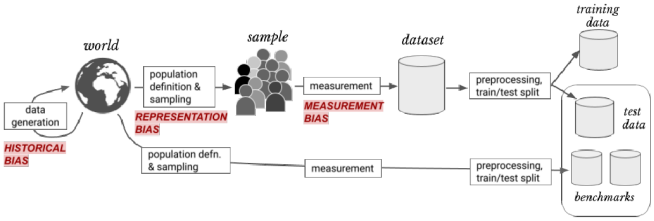

Figure 7: Potential areas of bias during the data collection process [17].

Bias can be introduced at every step of data collection, but as we do not have control over the original datasets, we cannot mitigate these biases. The only solution, besides creating a dataset from scratch, is to be aware of these biases and to document them in the model card.

In this paper, we cover the ethics of the technology, but failed to provide any implementation details, such as providing what algorithms to use since there are plenty of existing resources on this topic [20]–[23]. Regardless, before any definitive recommendations is made, it would be helpful to cover these in more detail.

References

[1] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations." 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523003354

[2] R. Plutchik and H. Kellerman, *Theories of Emotion, Vol. 1*. 2013. [Online]. Available: https://books.google.co.uk/books?id=Ie1GBQAAQBAJ&lpg=PP1&ots=MA0_sk5AGf&lr&pg=PP1#v=onepage&q&f=false

[3] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector." 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3240529

[4] B. Friedman and H. Nissenbaum, "Bias in Computer Systems." 1996.

[5] B. Friedman, P. Kahn, and A. Borning, "Value Sensitive Design and Information Systems." 2008. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-007-7844-3_4

[6] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey." 2021. [Online]. Available: https://arxiv.org/abs/2107.05989

[7] T. Gebru *et al.*, "Datasheets for Datasets." 2018. [Online]. Available: https://arxiv.org/abs/1803.09010

[8] S. Umbrello and I. van de Poel, "Mapping value sensitive design onto AI for social good principles." 2021. [Online]. Available: https://doi.org/10.1007/s43681-021-00038-3

[9] J. Davis and L. P. Nathan, *Value Sensitive Design: Applications, Adaptations, and Critiques*. 2015. [Online]. Available: https://doi.org/10.1007/978-94-007-6970-0_3

[10] B. Friedman and P. H. Kahn Jr., *HUMAN VALUES, ETHICS, AND DESIGN*. 2007. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.1201/9781410615862-78/human-values-ethics-design-batya-friedman-peter-kahn-jr

[11] J. K. Wolk, B. Friedman, and G. Jancke, "Value Tensions in Design: The Value Sensitive Design, Development, and Appropriation of a Corporation's." 2007. [Online]. Available: https://www.researchgate.net/publication/220729345_Value_Tensions_in_Design_The_Value_Sensitive_Design_Development_and_Appropriation_of_a_Corporation's

[12] L. Podoletz, "We have to talk about emotional AI and crime." 2022. [Online]. Available: https://link.springer.com/article/10.1007/s00146-022-01435-w

[13] D. Thomas, "The cameras that know if you're happy - or a threat." [Online]. Available: https://www.bbc.co.uk/news/business-44799239

[14] Google, "The Size and Quality of a Data Set." [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality

[15] M. Louw and M. Curl, "Data ethics and GDPR." [Online]. Available: https://www.csp.org.uk/professional-clinical/digital-physiotherapy/data-ethics-gdpr

[16] A. M. Udefi, S. Aina, A. R. Lawal, and A. I. Oluwafemi, "An Analysis of Bias in Facial Image Processing: A Review of Datasets." 2013.

[17] H. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." 2021. [Online]. Available: https://arxiv.org/abs/1901.10002

[18] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." 2018. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[19] M. Mitchell *et al.*, "Model Cards for Model Reporting." 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3287560.3287596

[20] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human Emotion Recognition: Review of Sensors and Methods." 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7037130/

[21] P. Naga, S. D. Marri, and R. Borreo, "Facial emotion recognition methods, datasets and technologies: A literature survey." 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2214785321048987

[22] F. Z. Canal *et al.*, "A survey on facial emotion recognition techniques: A state-of-the-art literature review." 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0020025521010136?via%3Dihub

[23] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review." 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1566253519302532