A Guide to

Challenges and Recommendations for the Implementation of a Fairer AI

Fundación Vía Libre

Index

Executive Summary 2

1. Introduction 3

Methodology and profile of the interviewees 4

Structure of this document 5

2. Challenges and recommendations 7

2.a. Previous considerations 7

2.b. Problem definition 8

What does this part of the process consist of? 8

Possible problems 9

Recommendations 9

2.c. Data compilation and curation 10

What does this part of the process consist of? 10

Possible problems 11

Recommendations 11

2.d. Model training 13

What does this part of the process consist of? 13

Possible problems 14

Recommendations 14

2.e Evaluation and model selection 15

What does this part of the process consist of? 15

Possible problems 15

Recommendations 16

2.f. Putting the model into production 16

Possible problems 17

Recommendations 17

Final words 19

Executive Summary

In this document, we elaborated a list of technical recommendations for the development of Artificial Intelligence (AI) projects, specifically Machine Learning based systems. These recommendations are the result of structured interviews with people who work in practical applications of data-based systems in various roles and organizations within the Argentine technological ecosystem, and have been elaborated through the lens of our AI ethics team, composed of activists, social scientists and computer science researchers.

The main recommendations we propose are:

1. Before proceeding with any development, analyze the need and social value of it. Consider the potential benefits and drawbacks of implementing the project.
2. Define the problem to be solved in non-technical terms, based on the joint reflection of the parties interested in developing an AI application. Consider the possible secondary uses of the application of the project and its possible harmful effects.
3. Analyze the data sources with which the model may be trained, taking into consideration the protection of personal data and intellectual property. Pay close attention to potential biases to ensure that all groups are adequately and non-discriminatorily represented, with a special focus on minorities.
4. When training and evaluating predictive models, incorporate a methodology for checking well-known limitations of models that lead to pernicious results: underrepresented classes, majority classes, a tendency to overfit, etc. Incorporate also more general error analysis to detect previously unknown limitations.
5. Carry out deployment in phases, accompanied by the corresponding monitoring mechanisms for early detection of pernicious effects of the implementation of the system in actual contexts, incorporating different perspectives into the analysis of the behavior of the system.

Equity analysis must be transversal, i.e., a reflective resource present throughout the entire system development and construction process, rather than being applied only when evaluating the discriminatory impacts of a system after it has been deployed.

This document intends to contribute to the understanding of the development processes of AI systems and the problems associated with them. Besides, it serves as a guide for those interested in developing more human rights-conscious technological products.

1. Introduction

Systems based on Artificial Intelligence (AI) can replicate and amplify social inequalities. Models that are inferred from data can acquire and amplify harmful patterns, such as abusive language or stereotyping.[1] For example, Internet search engines such as Google search have often been observed to reinforce gender or race stereotypes. In 2018, for the query "black women", the Google search engine offered much more pornographic content than for white men.[2] For this reason, the need to establish clear rules for the use of Artificial Intelligence is being claimed by governmental and academic institutions as well as by human rights activists in the digital environment. These rules should ensure that these technologies are respectful of the fundamental principles of societies, giving rise to an area known as the Artificial Intelligence Ethics (AI Ethics).

The European Commission designed proposals for regulations for the development and use of artificial intelligence.[3] The OECD (Organization for Economic Cooperation and Development) also presented AI regulation initiatives.[4] However, to guarantee fundamental rights such as non-discrimination or respect for privacy, and in general principles of fairness in the collection, representation, processing and availability of data and in the development and deployment of machine learning algorithms, it is essential to create AI systems that include, translate or implement these guidelines throughout their life cycle. AI ethics should not be an afterthought when evaluating the impacts of the system once it is in production, when it may already have harmed a large number of people.[5]

Instead of thinking of AI ethics as a separate initiative from the construction of the system or one more requirement that must be met, it is more organic to maintain a perspective centered on the rights of the people throughout the entire process of construction and development of the system. This is especially important when AI is used in critical processes or sensitive social aspects, as is the case with systems

[1] Emily M. Bender et al.: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 21), 2021. Available at https://dl.acm.org/doi/10.1145/3442188.3445922 (Consulted: July 4, 2021)

[2] Safiya Noble. "Google Has a Striking History of Bias Against Black Girls", Time, March 26th 2018, https://time.com/5209144/google-search-engine-algorithm-bias-racism/ (Consulted: July 4, 2021)

[3] digital- strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (Consulted: December 4, 2022)

[4] www.oecd.org/gov/innovative-government/hola-mundo-la-inteligencia-artificial-y-su-uso-en-el-sector-publico (Consulted: December 4, 2022)

[5] news.bloombergtax.com/tax-insights-and-commentary/we-can-all-learn-a-thing-or-two-from-the-dutch-ai-tax-scandal (Accessed: December 4, 2022)

that draft court rulings,[6] suggest possible diagnoses for a patient,[7] or propose actions to prevent school dropout[8] that directly affect the life of many end users. This document proposes a series of technical recommendations that arise from the dialogue with people who work in practical applications of data-based systems, in different roles at technological companies in Argentina. These dialogues aimed to understand and identify the needs, challenges, and possible consequences that arise throughout the development processes of AI through a human rights lens.

Methodology and profile of the interviewees

We conducted semi-structured interviews, each lasting approximately 40 minutes, through videoconferences. The script for the interviews was adapted from the one proposed in the article "Data Cascades in High Stakes AI".[9] In this work, the authors introduce and develop the concept of "Data Cascades". They define data cascades as events that, combined, cause negative effects such as damage to communities, the erosion of relationships with different actors, the discarding of entire data sets, and duplication of work due to software rewriting. These effects manifest themselves at later stages in Machine Learning applications. These events do not have to do with conscious bad intentions on the part of those who program, but rather with bad practices, bad incentives, and a series of reasons that the authors identify through interviews with practitioners from different areas (health, economy, universities, etc.).

We interviewed six AI developers and managers, who work in different roles at companies with different characteristics in terms of the number of collaborators, seniority, and areas in which they apply AI. Interviewees include a data leader, a data scientist, a technical lead, two start-up founders, and a researcher. Although all the companies (except one) are Latin American, we selected all the samples from the same region. We acknowledge that it is not the same to develop AI at a small company with a limited budget, few employees, and a handful of clients, as in a multinational company where these resources abound.

We interviewed two people from Mercado Libre, a leading e-commerce multinational company of Argentine origin, with more than 8,000 employees and offices in different Latin American countries; one person from Rappi, a multinational company of Colombian origin, very similar to Mercado Libre in terms

[6] Estevez, Elsa; Fillotrani, Pablo; Linares Lejarraga, Sebastian (2020-06). «PROMETEA: Transforming the administration of justice with artificial intelligence tools». Inter-American Development Bank

[7] Layes, María Elisabeth Silva, Marcelo Alejandro Falappa and G. Simari. "Clinical decision support systems." 4th Argentine Congress of Informatics and Health, CAIS 2013.

[8] Ignacio Urteaga, Laura Siri, Guillermo Garófalo (2020). "Early prediction of dropout using machine learning in professional online courses". RIED Ibero-American Journal of Distance Education, 23 (2), pp. 147-167.

[9] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 39, 1–15.

of size and business model; one data scientist in Argentina working at Medalia, a multinational company of American origin whose business model is based on extracting information from customer satisfaction surveys; and 2 startup founders from Argentina, one of which develops chatbots for appointment assignment in clinics, that is, the health sector, and the other develops predictive models in the education area.

The interviews were recorded and later transcribed. Tables with thematic indices of each one were generated in order to identify common topics. The interviews were analyzed to identify common problems in the AI development processes and the techniques and/or actions that the interviewees apply in their respective workspaces to address those challenges. In particular, we also discussed the tools and techniques that were missing in each of the development stages.

Structure of this document

This report is organized based on a workflow within machine learning-based development and deployment processes. In this process, broken down into five parts, possible problems and recommendations for each of them were grouped, understanding that different situations may arise at each moment of the process. This approach is based on the premise that each one of the parts of the AI development process has its particularities in terms of the actors involved and the technical tools in which they work. For example, in the data collection process, the the focus is on the data and its owners, whether they are the company's own databases, end users, or external clients.

The stages that we distinguish in the AI development process are the following. It is important to highlight that these steps are not sequential but any step may signal that is is necessary to return to a previous step and revise its outcome.

a) *Previous considerations:* We propose this stage as a reflective moment to ask if it is really necessary to build the system, if it is necessary to do it through Artificial Intelligence, taking into account who benefits and who it may harm.
b) *Definition of the problem*: This is moment of dialogue between the parties interested in developing the AI application. The problem is defined in non-technical terms and making explicit why it would be necessary to develop said application, who may be harmed and who benefits.
c) *Data compilation and curation*: This is the stage of the process where the data sources that the model will use in its training are selected and evaluated if they are consistent with the problem. Here the datasets are automatically or manually analyzed in order to verify consistency and balance.

d) *Model Training*. This is the process where the data selected in the previous stage is used to train models that address the problem defined above. Several models are trained with different strategies and algorithms.
e) *Evaluation and selection of models*. In this stage, metrics are used to evaluate the performance of the model. Different metrics measure and prioritize different characteristics of it and let designers decide whether to improve it or put it into production. This stage should include error analysis.
f) *Put into production*. At this point the model begins to interact with the real world, and this is where unexpected results can emerge.

2. Challenges and recommendations

In the following, we will group challenges and recommendations according to the standard workflow within the artificial intelligence industry. The workflow consists of five stages that were defined in the previous section. In these stages, different actors are in dialogue, and different problems arise. The stages overlap chronologically with each other; it is not a linear process but rather a cycle in which the results of the development of the stages may require redefinitions in previous or subsequent stages. For example, in the model selection stage, a modification in the training data implies reprocessing the data to adapt it to the different models that were considered.

2.a. Previous considerations

A first consideration, before any development, is to analyze if it is really necessary to build an AI system. The authors Paz Peña and Joana Varón (2021) reflect: "Instead of asking how to develop and deploy an AI system, shouldn't we be asking first "why to build it?", "is it really needed?", "on whose request?", "who profits?", "who loses?" from the deployment of a particular AI system? Should it even be developed and deployed?"[10]

Once it is decided that building the AI system is necessary, we can move on to the next points.

A cross-cutting recommendation that applies to the entire production process is the formation of diverse development teams that include members from underrepresented communities. The main objective of this recommendation is to avoid reproducing the hegemonic approaches that currently guide the development of AI systems.

From the beginning, it must be considered that data is a cutout of reality. Data is often taken as a source of universal and absolute truth, making it difficult to exercise a critical perspective on the functioning of these systems. To avoid this dynamic, it may be beneficial to start from the premise and keep in mind at all times that data is a cutout of reality. This premise should help enable mechanisms that facilitate a critical perspective rather than naturalizing AI systems as infallible.

Whenever a real-world problem is considered and translated into data, it must be considered that only a sample of the phenomenon is represented. Furthermore, how the problem is modeled is also a cutout of reality. For example, the choice of which characteristics to include to describe a person as a bank customer for a customer segmentation model will define which customer characteristics are relevant to the model. Characteristics like gender, age, and credit behavior could be variables that a priori make sense to include, whereas other characteristics like height do not. However, if the objective is to develop a system to predict the stock

[10] Paz Peña y Joana Varón, "Oppressive A.I.: Feminist Categories to Understand its Political Effects", Not my A.I., 10th october, 2021, https://notmy.ai/news/oppressive-a-i-feminist-categories-to-understand-its-political-effects/

of clothes to be sold in the summer, the height of the person could be a characteristic to consider.

2.b. Problem definition

What does this part of the process consist of?

In this first step, the phenomena to be treated automatically and the type of treatment to be given to them are defined. It is decided:

1) **delimitation of the problem**: which problems are interesting enough to dedicate the effort of development, implementation and follow-up to them. The interest of a problem may lie in the fact that it makes it possible to do things that were impossible until now (review a large number of documents to see which ones have a certain word), in the amount of work it saves, in the number of people it benefits, in the amount of time or money it saves, in the seriousness of the cases it makes it possible to treat... and the feasibility of treating them satisfactorily in an automatic way.
2) **definition of the results**: the solutions to be obtained for each type of problem to be treated.
3) **choosing and creating performance metrics**, which allow us to compare the goodness of different approaches and to detect problems in the behavior of the system.

The problem definition process should include not only the technical team that will develop the system, but also domain experts, who define the problem in terms of a real-world question or a business metric to be optimized, such as the number of sales.

Take as an example the problem of choosing which post to display on a social network. At this stage the problem can be defined as optimizing different measures: for example, we may want to maximize the probability that a user interacts with the post. Another option may be to optimize the probability that a user gives a positive rating to the post (smiley face, like, etc.). These different ways of modeling the problem lead to the development of very different products, which impact user behavior in different ways. For example, optimizing the probability of interaction favors discussions that, without moderation mechanisms or very clear interaction rules, often end in tension, polarization and violent content.

In each of the definition steps, it is critical to integrate insights that can detect potential impacts on different population groups or application contexts. For this reason, it is critical that teams incorporate diversities at the design stage and not only at the evaluation stage. For example, a person with reduced mobility may

detect that an extra category is required in the results of an image recognition system to identify whether an object is an obstacle.

Possible problems

The representation of the problem to be solved, and of what makes a positive outcome (to be optimized) determines the functioning of the system. If I believe that bananas are only yellow, when defining the problem I will introduce this bias and when creating a system that detects bananas, the system will detect yellow oval things. In the world there are bananas of different colors, but as they were not in the initial conception of banana, the system will leave them out. That is to say that the initial conception of the problem does not correspond to reality. Although very often this happens unconsciously, every time the initial conception of the problem leaves out part of the reality, there can be negative impacts on certain communities that are not represented by the system. Even if they are not left out, communities may be misrepresented, resulting in harmful behaviors of the system, for example, a larger number of errors for those communities.

Recommendations

- **Include existing knowledge in the problem area** to be addressed during the ideation of the approach, especially when determining what constitutes a good outcome or a bad outcome. This is a way to integrate in the design of the system knowledge about the problem that may not be represented by the data, that has not been collected in available datasets, but is known to domain experts to be important to adequately represent the problem.
- **Make all assumptions explicit**. Ex: We assume that our data are representative.
- **Include diversity in the design team**, end-users and the social science community especially in the problem definition, to better understand the impact of different options on the problem representation and avoid the so-called "discrimination by design".
- **Prefer human intervention** whenever decisions that may be harmful are involved, preferring analytical, reporting or BI (Business Intelligence) systems over fully automated decision making systems. We can find different situations in which human intervention is desirable:
    - **Recommendation instead of automated decisions** Whenever human rights are involved, it is preferable to use AI models to display information to help people make decisions, instead of letting them make the decision. However, it must be taken into account that people tend to take recommendations by automated systems as validated and objective, thus an education of the users is necessary to allow them to keep their own judgment, possibly with mechanisms to avoid that users rely excessively on recommendations, like letting

them know that some incorrect recommendations are included randomly among correct ones.

- **Previously unseen scenarios**: In some cases, scenarios are so different from the data from which the models have been inferred, that predictions are not reliable any more. That is often the case in emergencies or crises. For example, in the 2020 COVID pandemic, logistics for retail could not detect in time the change of behavior for toilet paper. In this case, a specialist should take over the direction of the system. A similar case is that of the autoPilot where the vehicle can be asked to delegate driving to the pilot.
- **Need for human interaction**: In cases where the automatic agent does not work as the user would like, it is desirable that there is the option to do something by hand, for example, if I want to make a medical appointment with a chat program with AI and the AI does not understand me repeatedly, there should be the option to talk to a human (human handoff).

2.c. Data compilation and curation

What does this part of the process consist of?

This part of the process consists, first of all, of collecting and selecting the data sources with which the machine learning algorithm will be trained. This may include the design of the data collection device, or determining which of the existing data sets are relevant to the problem. When defining what these data sources are going to be, different factors are taken into account, mainly their availability and cost of collection, but also their usefulness to address the problem. Another important step is to curate the data, that is, eliminate or modify those parts of the data that are either inconsistent (because of errors in the collection, for example), do not reflect the problem we are trying to model, or reflect unwanted biases.

In this section, we also include preliminary data analysis or exploratory data analysis. Through this analysis, interesting properties of the data can be detected that can be exploited in the machine learning process, such as previously unknown patterns. Errors can also be detected in the collection process, in the design of the sample, etc. Finally, exploratory analysis is a good time for early detection of biases, as it avoids further developments based on biased data. We consider this stage important and it is a good practice to devote considerable time to it because it impacts the work of subsequent phases and avoids bigger problems.

Possible problems

- Errors in the acquisition process
- Inconsistent data, that is, data with values that invalidate each other.
- Biases that discriminate against historically marginalized groups
- Assuming that having a large amount of data is equivalent to having quality data, representative of the problem domain, without asking whether they take into account different perspectives, other sources. It is important to ask, which cases are represented and which cases are not.
- Lack of information on capture and labeling. The labels that are assigned to the data define the problem, as the model will infer patterns from these labels.
- Using all available data in the model, without refining correlated or sensitive variables, variables containing errors, variables irrelevant to the representation of the problem, or comparing numerical variables that are on different scales, without further interpreting their meaning, e.g. comparing years with pupil size.

Recommendations

- **Preliminary analysis of the data:** It consists of reviewing the characteristics of the data available, and carry out these 3 activities:
    - Analysis of the data: what values the data can take, what type they are, what part of the problem they represent, etc. .
    - Sample manual inspection: take a sample of the data that is available and analyze it on a case-by-case basis.
    - Descriptive statistics: Show aggregate characteristics: most frequent and rarest values, minimum and maximum values, correlations between values, analysis of variance, outliers, etc.
- **Verify that data adequately represent the populations and define evaluation metrics:** Many times it happens that the samples are unbalanced. The most common configuration is that some values or clusters of values are much more frequent than others. This happens naturally because, frequently, data have a Pareto distribution, but it needs to be taken into account for designing the evaluation phase. Also, many machine learning algorithms are susceptible to exaggerating an unbalanced distribution (with majority class, and majority values for features). At this time it is important to notice whether a group of individuals, characteristics or phenomena is underrepresented or represented in such a way that the automatic systems based on the data may have discriminatory effects on a population. To control for this, it is necessary to document the distribution of classes and define evaluation metrics that can be disaggregated by social group. These metrics will be used at the evaluation stage in order to assess whether the model may harm some particular groups.
- **Analyze if there are groups in the domain that are not represented:** For example, a sample of university students could contain much more data on

young people between 20-30 years of age than on older adults, if we use these data to predict graduation age, the results of the model may not make sense for older adults since there are no data on this population.

- **If synthetic data is used, compare it with real-world data or validate it with a domain expert:** In the event that data does not exist, or these are not sufficient for model training, artificial data generation is often used. This may cause the training data to be different from the data that will be used when the system is in production. For example, software used to predict pathologies in medical studies could be trained with high-quality images generated specifically for model training, while those used by doctors in their offices could be of lower quality or out of focus. captured by cell phone cameras. Although it is often not possible to obtain non-synthetic data before the model is put into production, we recommend taking into account the differences that might exist and evaluating the system with real data and error analysis before putting it in production.
- **Document data sources and selection and collection processes:** One temptation in the absence of data is to use data collected for another problem. The lack of context about the data production carried out leads in certain cases to misinterpreting or misusing it. Frameworks like Datasheets[11] aim to guide those who build data sources to analyze the assumptions involved, risks, and implications of their use. This methodology also allows for transparency in who captures, labels, and defines the rules to process the data.
- **Document the curation processes:** It is very common to perform transformations on the data to eliminate errors in the acquisition process, better represent some phenomenon of interest, or simply avoid known limitations and maximize the results of machine learning systems. Curation processes usually have a very important component of knowledge based on experience, applying "common sense", ad hoc for each particular case, which can only be known if it is made explicit. For this reason it is especially important to document it, so that any team that uses the data later can spot potential mismatches with their problem.
- **Create a good user interface for data collection:** On some occasions, the data entered by users or by the people hired to collect them contain contradictory information that could be avoided from the design of the interface, such as: clarifying the units of measurement that may not be unified, or letting the user choose from a list instead of using free text when it is appropriate.
- **Define the least amount of characteristics or information necessary for the problem about the user**. When collecting or compiling the data that the model will use to represent the problem, unnecessary data is collected

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for datasets. commun. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723

just in case it becomes useful in potential future applications. However, greedy collection policies pose a potential threat to users, due to the vulnerability of making data available that may be sensitive in some way. Even if security mechanisms are established, no mechanism is infallible. Also unnecessary data collection incurs in avoidable ecological cost, due to the cost of data acquisition and storage.

- **Ensure the author's agreement for the use of data with intellectual property:** It must be verified that the data license allows the use that is going to be given or obtain an exception otherwise.
- **Ensure user compliance for the use of personal data:** If the application to be developed uses personal data, it is important to be informed and updated on the legislation in terms of personal data protection. For example, the Argentine personal data protection law indicates requirements to be met when collecting and storing this type of data, such as the explicit consent of the user.
- **Deploy security mechanisms to protect sensitive data:** **"***Sensitive data is considered to be personal data that reveals racial or ethnic origin, political opinions, religious, philosophical or moral convictions, union membership and information regarding health or sexual life."* (Personal Data Protection Law 25,326). In the case of projects that include this type of data, it is advisable to store the information in databases that implement the necessary security mechanisms. But even in this way, when this sensitive information is stored in databases, there is the possibility that the security mechanisms are violated and the privacy of the data holders is exposed. It is recommended as an additional security measure to use personal data anonymization tools and that said anonymization is then manually reviewed by domain experts. It is also important to take into account that some data that is not considered sensitive in the current context may be sensitive in future contexts or secondary uses.

2.d. Model training

What does this part of the process consist of?

This part of the process, largely conducted by the technical team, involves inference, evaluation and selection of the machine learning model(s) to be used for predictions. As part of this process, transformations may be performed on the data to take full advantage of the characteristics of each machine learning method or features of the dataset or distribution of the target predictions. It is important to divide the training data into training and validation sets that do not overlap or have any information leakage.

Possible problems

- Inferring models on data that have not gone through a curation process, and may bring problems of inadequate representation of the problem and biases of different kinds.
- Failure to incorporate the findings of the results analysis into the data representation, i.e., into the data curation process, to improve modeling.
- Failure to take into account the effects of a majority class on aggregate metrics such as average accuracy.
- Reproducing or amplifying biases present in the data, especially frequent in linear models, which tend to favor the majority class.
- Using models that assume distributions over the data and not verifying that these restrictions are met, for example, using a model that assumes a Gaussian distribution of the data for geometrically distributed data.
- Applying models inferred on one population on a different population, e.g., using a model inferred on a light-skinned population for a dark-skinned population.

Recommendations

- Use techniques that **reduce the tendency to favor the majority class** of linear methods, such as regressions or support vector machines, such as class balancing techniques:
    - **oversampling** of minority classes
    - **undersampling** of the majority class.
- When applying smoothing techniques to avoid overfitting, check that the majority values or the majority class in general are not being overrepresented.
- In the case of more expressive methods, such as decision trees or neural networks, methods can be applied to **improve the representation of minority cases**, like boosting and stacking approaches which by definition focus on improving the errors they make. For example, in an XGBoost, by increasing the number of trees, we can correct the errors of the previous trees.
- Train models with **data as close as possible to their actual context of use**: It is common to train models with a set of data and apply them on contexts different from those of the data collection. This practice can lead to very large errors in predictions. For example, suppose we are developing a system that turns on an air conditioner at a certain temperature and we use the ambient temperature data at which users in northern Mexico turn on the air conditioner. When we train that model and apply it, for example, in Patagonia in Argentina, the temperatures for users in the Argentina area may be very different.
- **Correlation does not imply causation**: just because two things co-occur, it does not necessarily mean that one causes the other. For example, if we

observe that there is a correlation between eating apples and having healthy skin, we cannot conclude that eating apples causes healthy skin. It could be that the two are related by some other factor, such as having a healthy lifestyle in general. Therefore, it is important to keep this concept in mind when conducting studies and interpreting the results.

2.e Evaluation and model selection

What does this part of the process consist of?

At this stage, the results of different trained models are analyzed to select the options that best fit the needs of the problem. The main question that arises at this stage is how the model will be evaluated, and this depends on its intended use. There are different metrics for evaluating model performance, and each of these metrics assesses different qualities.

Possible problems

- One problem that can appear at this stage is to find models that have very good general characteristics, but do not perform well for my particular problem. For example, a classifier designed to predict the means of locomotion used by students to attend a school (walking, public transport, bicycle) performed very well in accuracy and recall overall, but when analyzed further, it was discovered that most of the errors were in the bicycle class: the model predicted that students who actually used bicycles would not use them. In other words, it predicted that fewer students would use bicycles. This system was to be used to estimate how many bicycles would need to be purchased for students.
- Another consideration may be the explainability of the models, i.e. many times "it gives good results" but we don't know why and in particular cases we don't know why it made certain decisions. For example, in the medical field, decision making is a critical point, because a decision can directly influence the life and health of people. Therefore, if these AI methods are used as an aid in decision making, it is necessary to know something more about how each variable affects the prediction issued by the model.

Recommendations

- Favor the use of **models whose behavior is easy to explain**: Some AI models are interpretable per se. Simple models such as regressions, which in themselves offer us the importance of each variable in the decisions taken, or decision trees, which by their own structure indicate the decision path on the different variables that lead to the prediction or final decision.
- Use **explainability methods** to understand the behavior of the model: Explainability methods help to understand the behavior of models whose internal operation cannot be accessed. It is valuable to use these techniques to detect weaknesses and possible discriminatory behavior of the model, but always keeping in mind that they do not necessarily reflect the actual reasons for a model's behavior, but just some correlations. Some existing techniques are, for example, SHAP indices and saliency maps.
- **Error analysis**: By error analysis we do not only mean evaluating the models but assessing how the error is distributed, specially across groups that may be subject to discrimination, which are the worst cases, how often they occur and which groups they affect the most.
- Use **metrics that do not suffer in contexts of class imbalance**: we need to use metrics that allow us to zoom into the details of performance across different groups or classes:
    - instead of overall accuracy, favor **F1** per class or different averages, also with the detail of precision and recall
    - ROC AUC, with independent threshold.
- Use **data ablations** to observe the behavior of the system in previously unseen circumstances. Data ablations consist in modifying data to check what would happen to predictions with such modifications. It is not necessary to wait for something to happen to know how the model will behave in those cases.

2.f. Putting the model into production

What does this part of the process consist of?

This stage consists of the deployment of the developed system for its final use. In previous stages, the model interacted with data that was previously curated and selected to optimize its performance. At this stage, the generalization capability of the model is critical, as it will interact with previously unseen data that may have differences from the training data, even unexpected differences.

Possible problems

- The main problem we may have at this stage is that the new data from the real world is very different from the training data, and therefore the model does not know how to behave with them, or behaves in unexpected ways, producing possible ethical problems. The most prominent example of this problem is known as "Data Drift," or the paradox of predicting the past: for models that predict trends, if there is a sudden change in reality, the new data takes a long time to incorporate and consolidate for model retraining.For example, the shortage of toilet paper in Spanish supermarkets at the beginning of the pandemic. The predictive models used for stock replenishment took too long to detect this behavior change, to the point that they arrived too late.
- A second problem that emerges is the secondary uses of an application, which can lead to problems that were not foreseen in the application's design. An illustrative example is the use of systems that socialize the behavior of users and share information on where and when they practice a certain sport. These systems have as their primary use the formation of exercise groups in sports, but as a secondary use, they have been used to harass people.

Recommendations

- **Avoid using feedback loops**, i.e., prevent the mode from using only its predictions as a source for the retraining process. This might cause the model to have less predictable capabilities when applied to real data. It is recommended to update the model by retraining it with feedback from the performance of the deployed system.
- **Monitor the model's metrics in production**. Metrics used in the pre-production stages reflect the model's performance on training data.As real data may be quite different from training data, it is advisable to acknowledge prediction over training data and compare them with real data, to find out when it is necessary to retrain the model. In this stage, the developers have to establish the metrics and values that indicate the need for retraining.
- **Set up a reporting channel to report models' failures (either bias or discrimination)**. It is expected that models will have failures that may impact final users, so it is necessary to include tools to let users alert the models' developers when they detect these problems. For example, social networks often include a reporting method to tell if something is spam, not interesting, has hateful content, etc. Another case is Google Translate, which allows the user to report if a translation is incorrect and also lets the user suggest another one instead. In addition, if the YouTube algorithm

removes a video because it believes the user is violating copyright laws, the user can file a disclaimer through the interface.

- **Integrate feedback adequately**. Sometimes, when developers receive a report about a malfunction, they patch the specific case instead of incorporating it into training or doing a comprehensive analysis of the failure. This procedure is not recommended. As a developer and from a pragmatic point of view, this would be like solving a bug for a particular case or forcing a result for a particular case. Instead, proper analysis and integration of the feedback into the model retraining process are recommended.
- **Models trained in different contexts must be evaluated before being put into production**. Models trained in one specific context do not necessarily serve a different context, i.e., using a movie recommendation model trained with Chilean data for recommendations in Argentina can be very different. So, before proceeding, a performance analysis must be performed.

Final words

When developers are guided by their "common sense", that is, by what seems to be the most probable or the standard situation, without checking conditions and thinking about diversities and representations, they reproduce hegemonic practices and ways of thinking. We must question our most naturalized beliefs to create an anti-hegemonic AI that is not Eurocentric, white, patriarchal, or capacitive. In this sense, what we can see throughout the document can be summarized as "not taking anything for granted."

Throughout this document, we reviewed the main stages of developing data-driven products, identified potential issues, and made recommendations to mitigate those issues. Although, from a product quality and human rights perspective, we cannot ensure the production of a data-driven system without "failures" or undesired behaviors, these practices aim to improve the quality that strongly affects artificial intelligence ethics.

We are convinced that the technical and social perspectives are not dissociated. In this sense, as technicians, an AI that discriminates should be considered a low-quality AI. Similarly, systems that avoid discriminatory behavior and other types of unwanted bias should be regarded as higher-quality AI. Because technical constraints cause the majority of biases in data-driven systems, solving this problem can be considered a technical challenge. As Friedman and Nissenbaum (1996) suggest: "freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency— according to which the quality of systems in use in society should be judged."[12]

[12] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

eticaenia@vialibre.org.ar
www.vialibre.org.ar

Made with the support of