SUPPORT POOL
OF EXPERTS PROGRAMME

AI Auditing

Proposal for Algo-scores

by Dr. Gemma GALDON CLAVELL

As part of the SPE programme, the EDPB may commission contractors to provide reports and tools on specific topics.

The views expressed in the deliverables are those of their authors and they do not necessarily reflect the official position of the EDPB. The EDPB does not guarantee the accuracy of the information included in the deliverables. Neither the EDPB nor any person acting on the EDPB’s behalf may be held responsible for any use that may be made of the information contained in the deliverables.

Some excerpts may be redacted or removed from the deliverables as their publication would undermine the protection of legitimate interests, including, inter alia, the privacy and integrity of an individual regarding the protection of personal data in accordance with Regulation (EU) 2018/1725 and/or the commercial interests of a natural or legal person.

Table of Contents

Document initially submitted in January 2023, updated in June 2024

1. Introduction

As AI systems proliferate and impact on peoples' life-chances, there is a need to facilitate public awareness and understanding of how algorithmic systems work. The proposal below draws from two existing methodologies of significant societal impact: the Nutriscore and A+++.

1.1. Nutriscore

NUTRI-SCORE
A
B
C
D
E

Nutriscore relies on the computation of a nutrient profiling system derived from the United Kingdom Food Standards Agency nutrient profiling system (FSA score). The calculation process consists of three steps. In the first step, the nutritional score of the food product is assessed. The next step is the determination of Rayner's score, calculated in the same way for all food products, with the exception of cheese, vegetable and animal fats and oils, and drinks. The two scores are then used to classify the food product on the five-level Nutri-Score scale.

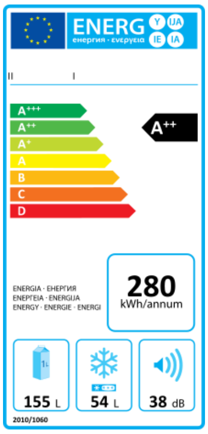
total N score - total P score = Nutritional score

Product ingredients negatively (**N**) affecting the Nutri-Score	Product ingredients positively (**P**) affecting the Nutri-Score
❖ High energy density per 100 g or per 100 ml ❖ High sugar content ❖ High content of saturated fatty acids ❖ High salt content	❖ Content of fruits, vegetables, nuts and legumes ❖ Fiber content ❖ Protein content ❖ Rapeseed, walnut and olive oil content

Based on this classification, the system awards 0 to 10 points for energy value and ingredients that should be limited in the diet (saturated fatty acids, sugar and salt, etc.) and 0 to 5 points for beneficial ingredients whose consumption should be promoted (fiber, protein, fruits, vegetables, legumes, nuts, and rapeseed oil). To determine the value of the label of a given product, i.e. the letter A, B, C, D or E, the sum of points awarded for the beneficial ingredients must be subtracted from the sum of points given for the ingredients that need to be limited. The product is classified in one of five value classes (A to E) based on the final score, which may vary from -15 to +40. The lower the score, the better the nutritional value of the product.

1.2. A+++

In 1992, the Council of the European Communities published EU Directive 92/75/EC "to enable the harmonization of national measures on the publication, particularly by means of labelling and of product information, of information on the consumption of energy and of other essential resources, and additional information concerning certain types of household appliances, thereby allowing consumers to choose more energy-efficient appliances." Since 2010, a new type of label exists that makes use of pictograms rather than words, to allow manufacturers to use a single label for products sold in different countries. Updated labelling requirements entered into force in 2021 (EU 221/340).



The energy labels are separated into at least four categories: 1) The appliance's details: according to each appliance, specific details, of the model and its materials, 2) Energy class: a color code associated to a letter (from A to G) that gives an idea of the appliance's electrical consumption, 3) Consumption, efficiency, capacity, etc.: this section gives information according to appliance type, and 4) Noise: the noise emitted by the appliance is described in decibels.

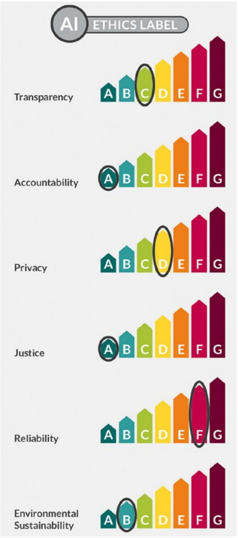
For instance, lamps are classified into class A if $\leq 0.240.\sqrt{\Phi} + 0.0103.\Phi$, where Φ is the luminous flux in lm and *P* is the power consumption of the lamp in W. Calculations and values are different for home appliances such as ovens, fridges and air conditioners, bulbs, cars or tyres.

2. Towards an AI algo-score

The above examples show how using visual cues to inform citizens/buyers of their choices is a method that has been tried and tested in other domains, where the regulator has played a significant role in establishing the methodology and setting up follow-up boards that continuously review the calculation algorithms used and propose updates to it.

2.1. Existing initiatives

2.1.1. *AI ethics label*[1]



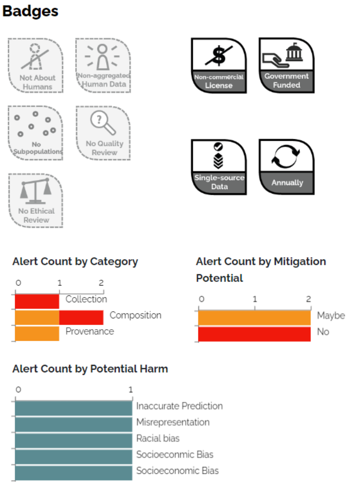
The possibility of an Algo-score being developed to guide citizens and buyers of AI systems has already been explored. In 2019, the AI Ethics Impact Group at Bertelsmann Stiftung proposed an AI ethics label. They reviewed 100 AI ethics guidelines and proposed six key values to be rated: transparency, accountability, privacy, justice, reliability and environmental sustainability. The scoring is obtained from the VCIO model (values, criteria, indicators and observables) that is used to define the requirements needed to achieve a certain rating.

The label includes one rating for each of the values captured with the VCIO-approach. In order to determine the level of compliance (a, B, C, D, E, F, G), the system aggregates observables on the basis of minimum requirements, thus moving away from the Nutriscore approach where bad practices can be offset by good practices.

Observables are then converted into indicators. Those indicators are in their turn converted into criteria, and these criteria are finally turned into values, and viceversa.

2.1.2. *Dataset Nutrition Label*

The Data Nutrition Project takes inspiration from nutritional labels used in the food industry. It enhances context, contents, and legibility of datasets by highlighting the 'ingredients' of a dataset to help shed light on whether the dataset is "healthy" for a particular use case. Its format is closer to a medical leaflet than a label such as the A+++ in that it provides a structured and visual representation of dataset characteristics (including description, date, author, content, source and foreseen use cases) and a series of "badges" (to visualise whether a system has gone through ethics review, if it handles personal information, whether the system is for-profit or non-for-profit, etc.). It also includes an "alert" scale that is designed to highlight issues, restrictions, and other relevant information about the data that might not be obvious to someone unfamiliar with the dataset. These alerts include a colour code that shows whether a specific risk/alert can be mitigated.



[1] From principles to practice: How can we make AI ethics measurable? (bertelsmann-stiftung.de)

2.1.3. Single focus labels

It is also worth mentioning two examples of single-focus labels: the **Open Ethics label**,[2] which proposes a voluntary scheme focused on the openness of the tools and results of an AI system and so assigns value to using open and not proprietary training data, open source code, and unrestricted decision spaces. On the other hand, **Le label Numérique Responsable**, developed by the Institute for Sustainable IT in partnership with the French Ministry of Ecological Transition, ADEME, and the WWF, is focused on issues of sustainability and digital environmental impact. It has been designed to be used by both communities and digital services businesses and it is based on 4 themes and 14 broad principles which cover diverse commitments in areas such as governance, participation, proactivity, inclusivity and software management.

2.2. Proposal for an algo-score

The analysis above shows the possibilities and limitations of existing labels, and point to the issues that need to be taken into account to develop a label for AI-related technical systems.

A scoring methodology based on the Nutriscore model would not work when ranking technological developments as it implies that bad practices can be "compensated" by good practices. This would probably be unacceptable for many stakeholders. A system based on the A+++ model is also not an immediately useful reference as it is based on shared and agreed on values of impact. Namely, electricity consumption and noise indicators.

As for the methods developed specifically for AI systems, which have not yet been implemented, they pose several challenges. The Dataset Nutrition Label, on the one hand, is more of a transparency tool than a method to compare between systems. As it offers multiple possibilities for graphically representing technical systems, and as those systems may be very diverse, it is highly likely that systems would use non-comparable indicators. This does not mean that the method proposed does not have value, but its usefulness deviates from a system used to facilitate consumer choice.

The AI ethics label is by far the most developed method to assess compliance. However, it does not look like it has managed to get traction since it was released in 2019. Some of its shortcomings are probably linked to the incentives to adopt such a model, but also to the fact that it is somewhat removed from legal compliance requirements -even though these could be emphasized.

Taking the lessons from these previous efforts, and the name and image recognition of existing systems, the proposal below begins to sketch the methodology that could help push forward a practical proposal for a visual system designed to promote transparency, accountability and consumer choice.

The proposal takes as a reference the A+++ method, and specifically 3 of its attributes:

- It is immediately recognizable
- It allows for a main measure of specific variables, but also other icons that are deemed relevant
- It is a method that generates incentives to improve in the alphabetical scale

Additionally, in its use in appliances, the A+++ label has proven to be able to incorporate different measures for different appliance types (from fridges to tyres). This opens the possibility for different technical systems to use slightly different methods, and thus the use of the same label and recognizable image for recommender systems, social media platforms, LLMs and biometrics systems.

[2] Open Ethics Label – Open Ethics Initiative

Based on the above, we propose a label that captures three "levels" of analysis:

- AI governance: appointment of the relevant governance positions (DPO, AI officer, etc), compliance with relevant standards (ISO/IEC 42001, 23894, 23053, 24027, etc), documentation of key risk assessment documents (DPIA, conformity assessment, environmental impact assessment, etc.), transparency documents (AI leaflet, etc.), data provenance certifications, and links to open data/sources.
- Model fairness and performance: Datasheets for Datasets, Model card/s, key metrics (fairness, robustness, security, accuracy… , etc), key attributes, experimental testing (red teaming, etc.)
- Post-market monitoring and auditing: regular bias audits, incident reporting, effective recourse and redress mechanisms and metrics

An issue that is not easy to export form the A+++ methodology to AI is the use of an agreed-upon measure of calculation such as watts. Therefore, the proposal is for the system to assess the existence/compliance of different exercises and supporting documentation. Having completed specific tasks would allow those subjected to the label assessment to scale up on the alphabetical scale.

Moreover, for those developers making additional efforts to ensure compliance, accountability and consumer choice, one to three "+" can be added to each letter (and not just the A). Thus, algorithmic systems could get positive "points" for complying in proactive, accountable ways, but also for going beyond their obligations to test and shape better practices in the AI industry. System developers and implementors could qualify for additional "+" if the information provided had been verified by an independent auditor, if the source documents mentioned were accessible by the user in an open repository, if the developer made the data available to third-party auditors and researchers, etc.

This approach is designed to promote and recognize best practices in AI trust and safety at a time when both AI solutions and regulatory requirements are in full bloom but not yet mature, and therefore it is still unclear both which AI solutions will have a broad, lasting impact, and how regulations will be translated into technical specifications.

The algo-score approach creates incentives for the AI industry to design their own trust and safety solutions and technical specifications and share them with regulators so that they in turn, request these best practices in upcoming regulations and guidelines.

edpb

European Data Protection Board