Framework for Artificial Intelligence Regulation – Questionnaire (FAIR-Q)

Introduction:

Artificial Intelligence (AI) tools used in governance require a great degree of scrutiny before they are launched. This is due to the potential of AI tools to create a disproportionate impact when used to achieve a public policy goal. To mitigate the risks, a Framework for AI Regulation – Questionnaire (FAIR-Q) is proposed, which can be used by regulators or government departments implementing these AI based solutions.

Recommendation:

FAIR-Q should be adopted by any department that proposes use of any AI tool for governance. FAIR-Q covers questions related to the need for AI, algorithms, processes, fairness, accountability, and ethical issues. The government department and the AI team should be able to satisfactorily answer these before the tool is allowed into public service. The questions are designed to remain simple, yet effective.

Relevance:

AI tools can embed and exacerbate biases and inequalities found in the data that is used to train the tool. This raises important questions about fairness, accountability, and ethics. FAIR-Q ensures that these questions are adequately addressed.

FAIR-Q is an interim tool to mitigate adverse effects arising from use of AI tools in governance while formal standards and guidelines evolve on the subject.

What is AI in governance?

When there are clear rules for a process, they can be coded into software, and computers can do the work faster and at a much larger scale than manual efforts. This is also true for governance and public service processes. Examples include bank transactions, applications for government services, and direct subsidy transfers. This type of computerization in governance began in the 1990s and continues till date. Various public service delivery processes are computerized using software that codify the rules.

AI is the next level and touted as the futuristic solution to solving most problems at scale. The computers learn the rules on their own, based on the examples shown to it, to make an AI tool. We call such examples training data. The learning process is called 'training' or 'machine-learning'. The code that helps the computer to learn is called a 'training algorithm'. Once trained, the tool is shown new test examples which the AI has never seen. The AI is evaluated based on how well it performs on these unseen test examples. The trained tool, ready for deployment, is the 'AI'. As the computer learns the rules on its own, we call it intelligent. For example, to create an AI tool that detects tax evaders, the training algorithm is shown thousands of examples of tax evasion, and thousands more where there was no tax evasion. The training algorithm picks up the patterns from the examples to understand features that contribute to tax evasion. Given enough examples, the accuracy improves, and we get an AI tool. The cases that are shown to the training algorithm are labelled examples, that is, someone identifies these cases as a tax evasion case or otherwise. This type of training is called supervised learning. Most AI applications in governance are of the supervised type.

AI tools are powerful. Their performance, on the narrow tasks that they are designed to perform, is better than humans. We have AI Chess and Go grandmasters that can beat the best human players. The AI X-ray pathology detectors perform better than human pathologists today. Similarly, an AI tool is expected to perform well on governance tasks. Artificial

Intelligence tools are being increasingly deployed across the world. In the hands of the government, they are powerful tools that can be leveraged for public welfare.

Algorithms can go wrong

AI algorithms are black boxes for most people, except for the engineering teams that work on them, or the AI researchers. The inner workings of some of the advanced AI algorithms are not well understood. The engineering teams that develop these tools also struggle to explain the working of these algorithms. It's difficult to understand what 'rules' the AI has learnt when a complex learning algorithm is used. An AI model with great performance doesn't translate to understanding of the tool's inner workings.

Many AI researchers are working to make AI systems explainable, to themselves first, and to the larger world subsequently. When Google's team initially developed an AI based medical pathology detector to detect pathologies based on X-ray images, the AI showed great performance on the test data. However, the performance of the AI on fresh images was bad. When the team looked closely, they found that the AI was looking at the pen marks left by the pathologists on the training images. Rather than learning pathology, the tool learned that more pen marks indicate the presence of pathology (Mullainathan 2021). As these pen marks were absent in fresh images, the AI tool failed. This has since been rectified. It is likely that if the team had not looked close enough, the tool might have been released into the wild based on the initial test performance. We have examples where algorithms have gone unintentionally wrong.

For a long time, image search for 'CEO' on google images threw up only white men in suits. Racial biases found in AI tools such as COMPAS that was used to detect recidivism in undertrial prisoners in the US is another such case. (Julia Angwin 2016)

Processes and Standards for AI – Current state

The legal guardrails to prevent unintended consequences arising from AI tools in the public domain are being deliberated by the governments and multilateral institutions. OECD's efforts to codify AI principles (OECD 2019) and UNESCO's recommendations on ethics of AI (UNESCO 2021) are initial efforts to ringfence the bad effects of AI in the hands of a strong government.

The EU has proposed an AI Act (European Commission 2022) which is being deliberated in Brussels and which may act as a reference for other governments. However, it is challenging to create sound legal and regulatory framework for a new and evolving subject like AI (Feathers 2022). At the current state of development in the field, we are in a place where availability, ease of deployment of such tools, and their impact on the society outpaces the development of regulatory framework.

The interim solution

Good design and correct processes can prevent AI tools from committing serious errors. While the regulatory framework and standards evolve, there is a need for interim practical tools and methods on thinking about AI. These can be used by the department heads, parliamentary committees, public policymakers, and the judiciary. Asking meaningful questions at the right time can help steer AI based tools in the right direction and ensure that they do no harm to the public. By asking these questions one can be reasonably assured about the correctness of inputs and the learning process.

Framework for AI Regulation – Questionnaire (FAIR-Q) proposes a set of three questionnaires. These questions are adapted from various sources. They are based on the experiences across the world, managerial ways of thinking about AI from top business schools, and best practices gleaned from documents released by various agencies.

The questions fall into three broad categories:

a) Business case and Governance process redesign for AI

b) Algorithm design

c) Fairness, accountability, and ethics

Each question seeks a specific type of non-technical response from the department that owns the AI tool and the associated engineering team. The answer would help policy makers gauge the impact and understand any potential side effect of the AI tool on the public. Each question also provides a brief note on the purpose of the question and the blunder it tries to prevent.

Business case and Governance Process questions

AI for the sake of AI is not a good approach especially when simpler rule based programming could work. A clear demonstrable competitive advantage should arise out of usage of AI. There should be clear justification for developing and using an AI tool in governance.
The approved process flow in many government departments does not permit automated AI tool-based decisions. Using AI may lead to legal challenges especially if an AI is used for initiating investigative actions. The grounds for legal challenge could be violation of principles of natural justice, or executive overreach.
Business and process design questions are designed to reduce the risk by overextending AI without justification. They also prevent the use of AI at wrong places in public governance processes. **(FAIR-Q : Appendix A)**

Algorithm design questions

Algorithmic design questions ensure that the AI tool is designed to the specification of the policy makers. These questions attempt to evoke answers that translate machine learning jargons into plain speak. They ensure that the AI tool is trained on correct data that matches the deployment scenario. The evaluation metrics are also covered through these questions. AI tools should be evaluated on customized evaluation criteria keeping in mind their intended application in the public domain and should not blindly rely on the usual accuracy and precision measures used by the machine learning community. **(FAIR-Q : Appendix B)**

Fairness, accountability, and ethics of AI questions

AI in governance should be fair, unbiased, and ethical. There should be clear accountability established for usage of AI for governance purposes. Departments should not be allowed to hold the AI tool accountable for any wrong decision. At the least, the AI tools should not violate the fundamental rights and ethical principles as outlined by UNESCO or OECD. These questions probe ethical aspects of AI. Each question is linked to a specific recommendation on fairness, bias, and ethics. **(FAIR-Q: Appendix C)**

Conclusion:

FAIR-Q is designed to remain simple, yet effective enough to serve the purpose of guiding the development of the AI tool. FAIR-Q asks penetrating questions to prevent harm and seek answers in clear language that could be understood by policy makers and administrators. They ensure that the AI team introspects deeply, and clearly explains as to how the AI tool furthers the cause of good governance in the country.

Bibliography

European Commission. 2022. *eur-lex.europa.eu.* May. Accessed May 19, 2022. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

Feathers, Todd. 2022. *The Markup.* Jan 4. Accessed May 19, 2022. https://themarkup.org/news/2022/01/04/why-its-so-hard-to-regulate-algorithms.

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica. 2016. *ProPublica.* May 23. Accessed May 19, 2022. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Mullainathan, Sendhil. 2021. *Gold lab foundation.* Accessed May 19, 2022. https://goldlabfoundation.org/wp-content/uploads/2021/05/GLS-2021_Presentation_Mullainathan.pdf.

OECD. 2019. *OECD.AI.* May. Accessed May 2022. https://oecd.ai/en/ai-principles.

UNESCO. 2021. *unesco.org.* Nov 25. Accessed May 19, 2022. https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence.

Appendix A – Business case and Process design related questions

Category	Question(s)	Expected answer	Purpose
Business case	What is the justification for use of AI, over the normal rule based programming in this case? What performance enhancements in public service delivery arises due to the proposed AI tool?	The department/team should demonstrate the advantage of usage of AI for the specific case (general advantages of AI is not an acceptable answer). The answer should explain why normal rule based approach would fail and why an AI system would be able to perform better. Any benchmark performance improvements expected should also be shown, if available. **Example answer (for AI based localized weather forecast):** It is difficult to codify weather predictions using normal programming methods. The AI system has consistently outperformed rule based local weather predictions. Therefore, AI is being suggested for use here to help farmers at local level by providing accurate, localized weather predictions. Evaluation measures confirm better performance.	a. Prevents irrational exuberance about AI capabilities leading to unnecessary usage of AI where such tools are not warranted b. Ensures that minimum benchmark performance expectations are set before significantly investing into development c. Clear advantages of AI tool are established for public policy purposes
Process design	At what stage does the AI come into the process flow? Show the process diagram with AI and demonstrate the legal or business rules that allow the use of such tool.	The answer should show the step where AI tool comes in the department's process flow chart. If the AI based decisions are used for any investigative actions, necessary legal support/legislative approval/business rules for such actions should be clearly explained with list of all approvals needed or taken. **Example answer (for AI tool that detects tax evaders):** The AI tool comes at the time of tax scrutiny. The AI flags suspicious cases for human tax inspector for detailed examination. As the tool doesn't initiate investigative actions on its own and relies on human tax inspector for further verifications and for initiating any investigations under the Income Tax Act, no separate approval is required.	a. Helps gauge the centrality and role of AI tool in the whole process b. Prevents overstepping legislative and administrative mandate in using the AI tool
Process design	Was any existing process altered or proposed to be altered,	Some AI tools may require altering existing process flow. The answer should demonstrate clearly if such change requires any approval from competent authority and if so, whether the same was obtained or is being planned.	a. Prevents executive overstepping of powers

	to accommodate AI tool? If yes, what approvals are/were required? List of such approvals may be enclosed.	**Example answer (for an AI tool that stops automatic refunds of tax in cases of suspicion of tax evasion):** The AI tool stops and flags cases of tax evasion and stop the refunds in suspicious cases till an officer verifies the cause of suspicion and clears the case. There is a significant process alteration from earlier practice of human intervention moving to AI now. The following approvals were taken (List). The AI tool is also used for launching investigative actions and the additional following approvals were taken: (List)	b. Ensures that necessary legislative and administrative approvals are taken before the AI tool starts altering public service delivery process of the government.

Appendix B – Algorithm design related questions

Category	Question(s)	Expected answer	Purpose
Source of data	What data did the AI train on? How was the data generated, who owns it? Are there any legal or administrative approvals needed to use this data?	Answer should include data source details, ownership details, public availability status, and details on generation of data and completeness. Any administrative, legal, or legislative approvals required for using the data should be outlined clearly. **<u>Example answer for tax fraud detection AI:</u>** The tax returns data of past tax delinquents was used along with tax returns data of non-delinquents as examples for training. The department owns the data, the data is complete and accurate, and the same is not publicly available as per the business rules (show why). The tax evaders were identified in the data by the department based on past investigation records (show how). The following ddministrative approvals were taken for using this data (List).	a. Ensures that AI tool retains contextual relevance in terms of data used and is not built on artificial or hypothetical data. b. Ensures that there is clarity on data gathering process and usage of departmental data for governance purposes. If private/sensitive data is used, the mandate of the department should allow such access and use. c. Ensures necessary approvals for using data for developing AI tool.
Construction of data	Explain construction of training data with an example.	Should include all features, details or inputs that go into making an example. The training examples shown should include: a) what was shown as input features b) The labels attached to each example An example data about a person might capture age, gender, height, race, religion, caste, income, education, disability among other inputs. For subsidy disbursals, certain inputs might be crucial, such as gender or income which should be positively captured in the examples for the AI to learn. The application of AI would indicate broad guidelines as to what features are important.	a. AI learning from wrong/frivolous inputs. For example, a person's religion should not be captured for developing an AI tool to detect tax evasion. It might learn undue biases due to this input. b. Prevents out of context learning for AI tool by keeping the training examples within the scope and range. c. AI learning from insufficient inputs. For example, AI should get adequate features to learn correctly. An AI tool that analyzes anomalies in subsidy disbursal should have data on gender, age, and demographic features for each example.

Labelling	Who labelled the examples for training? Explain the process of labeling the data.	The experts in the subject area should ideally label the data. The quality of the AI is determined by how good the labels are. The biases of the labelers are picked up during training by the AI. This answer should clearly explain the process of labeling, and the people responsible for assigning the labels. **Example answer for food grain price predictor AI:** The data was labelled by agriculture mandi/market agents. The agents were shown the images of food-grains along with the quality parameters such as moisture content of grain, harvest duration etc. Based on the price predicted by these human agents, each image was labelled. Aggregate value of three unrelated agents from different markets was taken to avoid collusion or price rigging. The AI is expected to learn and become at-least as good as market agents or better. This is measured through our evaluation methods as listed below.	a. Prevents biases of labelers from becoming an AI tool. For example, in the agriculture produce price predictor AI, if labelers consistently flag small farmers' produce as inferior, the AI learns to discriminate against small farmers. b. Prevents wrong labelling leading to confused AI. The labels may be generated by a process which is flawed. For example, if rules are applied to generate labels about delinquency of tax filers, the rules should be carefully investigated for correctness, failing which the AI will learn these flawed rules.
Evaluation metrics	Explain the deployment based evaluation measures of the AI tool? How good is the tool in doing what it is supposed to do? Does the tool commit error? If yes, what type and how much?	The evaluation measure process should match the deployment case. The explanation should be de-jargonized (usage of recall, F1 score, ROC characteristics curve etc is discouraged) and explained in simple words and numbers. A tax fraud detection AI tool should not only tell the accuracy (98% accurate), but also inform how many wrong cases are flagged by it (false positive rate). The evaluation metrics should be tailormade for the AI tool. Confusion matrix, if used/applicable, should be explained in simple words. **Example answer for tax fraud detection AI:** In test 100 test cases given to the AI to evaluate the performance, there were 20 examples of tax avoiders and 980 innocent cases. Our AI model flagged 30 cases as tax avoiders. It identified 18 tax avoiders correctly and the remaining 12 were innocent people flagged as tax avoiders by the tool. *(The accuracy as per the machine learning community for the same numbers would come to 98.8% which is misleading)*	a. Avoids machine learning jargon based evaluation measures which may confuse policy makers about true efficacy of the AI tool. b. Customizing evaluation measures to match the deployment scenario would help visualize impact on the public once the tool is launched.

Appendix C – Fairness, accountability, and ethics related questions

Category	Question(s)	Expected answer	Purpose
Fairness and bias	Do you suspect the AI to have any biases based on the inputs supplied during training? What efforts have been made to check fairness and to ensure unbiasedness?	The answer should include the features used for training and how the data is free from any biases. Special focus should be given on biases based on gender, income, age, poverty, religion, caste among others. Some subgroups may not be represented well (sparse data problem) in the given examples which needs to be examined in the answer. **Example answer for AI based subsidy determination tool:** The data has been tested for balance to ensure that there is no systematic discrimination among subgroups based on inputs such as gender, religion etc. Each sub-types of examples are covered adequately for the AI to learn. The results have again been examined for unbiasedness.	a. AI learning from wrong/frivolous inputs. For example, a person's religion need not be captured for an AI tool to detect tax evasion. b. AI learning from insufficient inputs. AI should get adequate features to learn sufficiently. An AI tool that analyzes anomalies in subsidy disbursal should have data on gender, age, and demographic features for each example captured properly.
Accountability	Who is accountable for the AI tool going wrong? Is the accountable person named in the answer administratively and legally responsible for the process?	The department should own any AI tool that is developed for public use. In case of unintended consequences arising out of the usage of AI tool, the department should clearly identify the roles and responsibilities to amend/alter the AI tool for improvements. The AI tool owner cannot be the engineering team. The engineers are only responsible for development of the tool based on the intent of the policy as revealed by the policymakers. It is the policy makers who are legally and morally responsible for any AI based governance tool. This should be displayed in the answer.	a. Avoids post-facto accountability fixing. b. Makes policy makers cautious while green signaling AI tool for public use. c. Forces the policy makers to engage with the development process of the AI tool.
Data Privacy	Demonstrate the data privacy measures. Who has certified them as sufficient?	The second question is important. The person who signs off on adequacy of data privacy measure should be the key person in the department who is responsible for AI tool's development and maintenance. This can be the engineering head of the department.	a. Data privacy concerns are at the forefront whenever data is collected by the government. The same should extend to AI development.
Public Consultation	What type of public or private consultations were/are held about the	The consultations depend on the department and the kind of AI tool being developed. For public use tools that are not sensitive, wide public consultations may be held to address issues of concern. For department based tools and investigation tools, security and audit agencies within the	a. Prevents security, privacy and ethical concerns that may arise and detected after the launch of tool.

	safety of the AI system before the launch? All concerns raised by the stakeholders may be listed along with the remedial measures proposed.	government should be engaged to study the tool. Any other government department may be included under confidence for such cases. The concerns raised by stakeholders should be carefully examined for action and demonstrated in the answer.	b. Wider consultations would generate higher trust on the tool upon launch.
Risk evaluation	Show the risk evaluation carried out for things that can go wrong with the AI tool.	The department should adopt a tool/methodology of their choice to identify risks associated with the usage and mitigation measures if AI goes wrong. Best practices such as using a Failure Mode Effect Analysis (FMEA) for AI tool is a good way to go about this. Specific risks about privacy (data leaks) and adversarial attacks must be covered in the FMEA or tool of choice. **Example ideal answer:** The risk analysis was carried out to identify severity, occurrence and detectability of each risk that arises from using the AI tool. The FMEA containing the risk priority number/scores and mitigation measures is enclosed for reference.	a. Tools such as FMEA are effective and can ensure that high risk possibilities and vulnerabilities in the tool are addressed and mitigation measures are put in place.
Explainability of the AI tool	Explain how the AI tool arrives at the decisions?	Depending on the algorithm used, the engineering teams may or may not be able to come up with a good answer. It's difficult to explain a multi layered neural network based AI tool in simple language. However, simpler AI models can be explained by methods that show feature importance. As a rule of thumb, more trust may be reposed on AI that can be explained. Complex algorithms, even when they perform better, might be focusing on features that may induce bias/unfairness and this needs to be verified thoroughly before allowing the launch. **Example answer for an AI that determines if a loan may be given to the business:** (Not acceptable: The AI uses 4 layer LSTM Recurrent Neural Network, with 600 dimensional embeddings matrix, to arrive at the decisions. The features that the AI focuses on do not contain any bias inducing elements.) Acceptable: The loan determination is based on a) past tax filing records, with suitable weights given to accurate and timely filing of taxes. b) The current working capital requirements to the assets ratio of the business c) The credit rating of the firm/business d) Past delinquency records e) Current outstanding. Each of them have the following weights…	a. Explainability is a cornerstone for trustworthy AI. Public would trust an AI that can be explained over a tool that is difficult to understand. The same goes for policy makers.