W205 Data Engineering - Final Project
A Framework for the Metamorphosis of Algorithmic Harm

By: Jonathan Hernandez
August 6, 2024

Preface

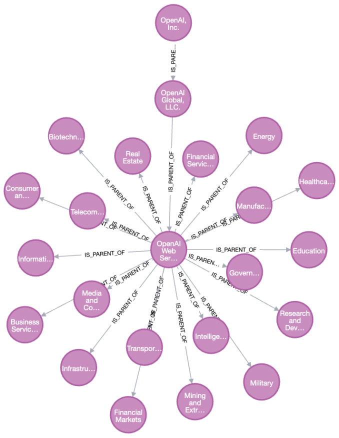
This Introduction is offered as optional. It sets a hypothetical context for our alternative business case, in the year 2034.

Introduction

The year is 2034, OpenAI runs several quadrillion parameters' worth of models each hour, serving hundreds of millions of customers. Models are served across geographies, languages, time zones, cultures, regulatory regimes, and populations. Hundreds of millions of climate refugees have migrated into new regions and countries, which have become increasingly diverse—across race, culture, language, spirituality, and wealth.

Last year, in 2033, the Institute of Electrical and Electronics Engineers (IEEE) estimated that the models running in production across the world will collectively cross fifty quadrillion parameters by 2035, and Data Centers will comprise over 17% of the U.S. energy consumption.

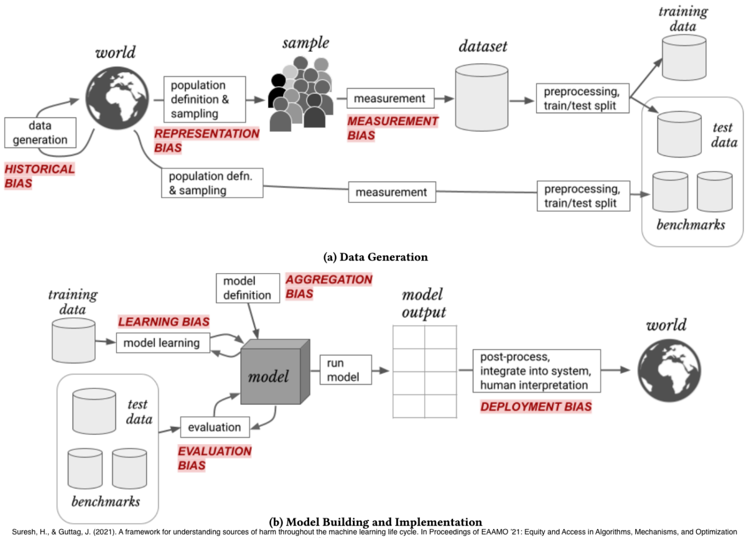
Your company, OpenAI Web Services (OWS), and its two parent companies, OpenAI Global, LLC., and OpenAI, Inc., collectively serve models to the public and private sectors, closed source, ubiquitously consumable over API endpoints. Last year, your subsidiaries were valued at $3.2 Trillion—the highest value for a non-public company in recorded history. You have fine-tuned various models to customers' business cases across Manufacturing, Energy, Mining and Extraction, Financial Services, Healthcare, Education, Real Estate, Information Technology, Biotechnology, Telecommunications, Consumer Goods and Services, Transportation and Logistics, Media and Communications, Infrastructure, Financial Markets, Business Services, Research and Development, Government, Intelligence and Military.



Executive Summary & Business Case

As a result of the 2032 Greater Pacific Coalition vs. OpenAI, 2033 EU vs. OpenAI, and 2034 U.S. Federal Trade Commission vs. OpenAI court cases, you've hired us, Applied Adaptations Consulting, to help you comply with new legal requirements. Your mandated coincident deadlines come into effect in six months.

OpenAI, Inc. has been ordered to create an independent, technologically advanced, investigative office, which we call the Bias & Fairness Compliance Technologies Office (BFCTO). Bias is defined as, "prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair". In alignment with the court orders, our mandate at the BFCTO is to deliver compliance, reporting, and active risk mitigation capabilities that identify sources and instances of bias and/or harm to the public.



(a) Data Generation

(b) Model Building and Implementation

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1-9). ACM. https://doi.org/10.1145/3465416.3483305

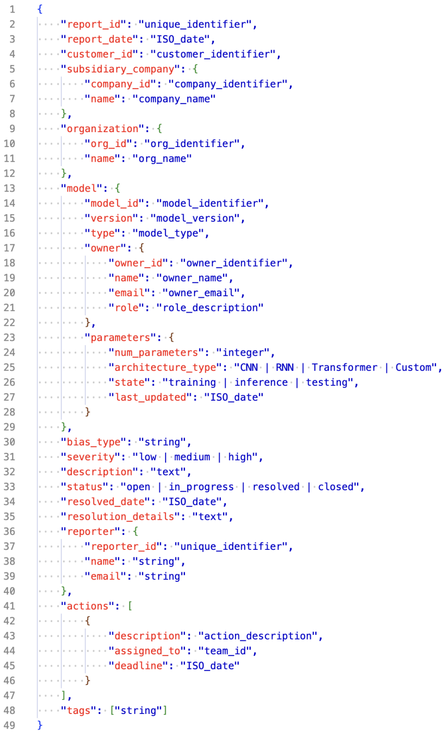
Our Fix-Firm Price contract covers the development of the BFCTO, which will enable you to succeed in all three regulatory regimes. We will work closely together so you succeed in AI & Machine Learning Compliance. Supporting workstream 42 of our contract, our team of three Berkeley-trained Data Scientists, led by our Senior Manager, Jonathan Hernandez, will propose a Solution Architecture portfolio that meets the BFCTO's needs.

We will deliver:

- **42.1:** The reporting of **bias risks** throughout your company and subsidiaries' operations;
- **42.1.1:** Statistics and disclosures of **bias reports** from your customers and/or their customers;
- **42.2:** A comprehensive **enterprise ML model catalog**;
- **42.2.1:** Resources, and protection, for your employees to report **known or suspected bias** in organizations, processes, people, or models;
- **42.3: Quick reaction capabilities** to report and prevent suspected harm;
- **42.3.1** A **real-time scorecard** of live issues that are not known in advance;
- **42.4:** Regular and **strategic analysis of the highest likelihood areas** of your operations to produce bias.

This high-level report will serve as a strategic North Star for our ensuing planning and development efforts. The following sections will focus on technologies that will serve the BFCTO's mission, better than others. During project Initiation and Discovery, we learned that you're considering various SQL solutions to enable these goals, and we want to help guide you through the decision-making process. Therefore, we will propose alternatives, and explain why they are a better fit for your business case, and the labyrinthine data and compliance requirements.

42.1 Bias Risk Reporting, 42.1.1 Relevant Statistics & Disclosures, 42.2 Enterprise ML Model Catalog - Proposed Solution: MongoDB



```
{
    "report_id": "unique_identifier",
    "report_date": "ISO_date",
    "customer_id": "customer_identifier",
    "subsidiary_company": {
        "company_id": "company_identifier",
        "name": "company_name"
    },
    "organization": {
        "org_id": "org_identifier",
        "name": "org_name"
    },
    "model": {
        "model_id": "model_identifier",
        "version": "model_version",
        "type": "model_type",
        "owner": {
            "owner_id": "owner_identifier",
            "name": "owner_name",
            "email": "owner_email",
            "role": "role_description"
        },
        "parameters": {
            "num_parameters": "integer",
            "architecture_type": "CNN | RNN | Transformer | Custom",
            "state": "training | inference | testing",
            "last_updated": "ISO_date"
        }
    },
    "bias_type": "string",
    "severity": "low | medium | high",
    "description": "text",
    "status": "open | in_progress | resolved | closed",
    "resolved_date": "ISO_date",
    "resolution_details": "text",
    "reporter": {
        "reporter_id": "unique_identifier",
        "name": "string",
        "email": "string"
    },
    "actions": [
        {
            "description": "action_description",
            "assigned_to": "team_id",
            "deadline": "ISO_date"
        }
    ],
    "tags": ["string"]
}
```

Use Case: MongoDB will support Model Catalog Management, Compliance Reporting, and act as a scalable distributed repository for your business.
Capabilities: Schema Flexibility, Horizontal Scalability (Sharding)
Data Model: BSON (Binary JSON), nested documents, arrays
Advantages Over Relational Databases:

1. **Schema Flexibility:** Documents (records) in a collection (table) don't require a predefined structure, whereas SQL databases do. MongoDB has an ability to handle diverse data structures with unstructured data or evolving schema patterns. It can aggregate information across diverse Organizations, Subsidiaries, and your ML Models. Traditional SQL databases require strong consistency, as they're optimized for ACID (Atomicity, Consistency, Isolation, and Durability).
    a. **Tradeoff:** Data and integrity are your developers' responsibility, which could potentially lead to anomalies if underlying documents change without prior time to accommodate it. In addition, your queries can become more complex, and slower if not well-optimized—our team can help with this.
2. **Scalability:** By scaling out horizontally with sharding, MongoDB provides an advantage over SQL's need to scale up with more powerful hardware. It should be noted that this adds complexity to the system.
    a. **Tradeoff:** There may be data redundancy present in your systems, due to duplication of data. Unlike a common normalized schema approach used in SQL, Third Normal Form (3NF), there is no requirement to normalize your data.
    b. While this is one of the costs of moving to a distributed system, if you store your data in Storage Area Networks (SANs) of the same vendor, you can leverage deduplication at the block store level to reduce your overall footprint, which eliminates unnecessary duplication of identical blocks of data stored on your hard drives. However, there would be other considerations, such as setting block sizes small enough so that MongoDB's frequent modification of small parts of large datasets did not hamper the effectiveness of deduplication, but which is part of a ratio that then increases the total Input/Output Read/Write Operations per second (IOPS) across storage controllers, increasing their workload. Slowdowns in MongoDB could be mitigated by employing indexing and aggregation frameworks.

42.1 Bias Risk Reporting: You have recently mandated regular reporting of bias risks across your subsidiaries. MongoDB is well-suited to consolidate this information, which could come in different forms and schemas.

42.1.1 Relevant Statistics & Disclosures: After consolidating this information, it will enable you to report internally and externally on the accumulation of bias reports from multiple lenses. Depending on the data you're capturing, MongoDB can be queried from multiple points of view, such as by Model, Organization, Leader, Subsidiary, Type of bias, and even the bias Source. We can report on Incident Rates, Severity, Resolution Metrics, Model Vulnerability, and derive other insights.

42.2 Enterprise ML Model Catalog: You will gain visibility into many aspects of your operations, and be able to quickly determine how often bias occurs, how, and then identify stakeholders who can support remediation.

As you will see in section 42.2, Strategic Analysis, we can identify high-risk areas by combining information from MongoDB with a Graph database, allowing us to run centrality and community detection algorithms.

42.2.1 Known or Suspected Bias Reporting, 42.3 Quick Reaction Capabilities, 42.3.1 Real-Time Scorecard - Proposed Solution: Redis

Use Case: Redis will support the real-time detection and reporting of elevated bias reports.
Capabilities: Low latency real-time analytics, integration to pub-sub messaging queues.
Data Model: Key/value pairs
Advantages Over Relational Databases: Because Redis operates entirely in-memory, it reduces data access time compared to disk-based storage systems. This allows us to detect and report on bias reports, fast and real-time.

1. **Speed:** You can expect consistent response times in the low milliseconds. This minimizes latency, and allows it to quickly ingest and process streams of bias reports, and related data. SQL often has higher latency, more limited scalability and performance bottlenecks when there is a high concurrency of read/write operations. It also has a more limited schema.
    a. **Tradeoff:** RAM is more expensive than disk storage, and will increase as your volume of data grows.
2. **Front End Cache for MongoDB:** For data coming directly from the "bias reporting web API", we can output it in JSON, and use a pub-sub messaging queue to send it directly to a consumer that will write it to Redis, first. Additionally, queries from your stakeholders and business users will reach Redis first (up to a specified period of retention, like 30-120 days). When the key is present, the query will be returned by Redis at high speed. When the key is not present, then MongoDB will return the data.

42.2.1 Known or Suspected Bias Reporting: We can use web servers to provide employees, customers, and members of the public the ability to report known or suspected bias. Our system will rapidly ingest and query these data.

42.3 Quick Reaction Capabilities: When a bias incident is reported, Redis can trigger alerts and updates across your systems. We can integrate it with pub-sub messaging queues to allow for immediate dissemination of information to relevant stakeholders.

42.3.1 Real-Time Scorecard: Our scorecard will show live bias incidents and their statuses, which can support similar points of view as described in 42.1.1.

42.4 Strategic Analysis of High Risk Operations - Proposed Solution: Neo4j

Use Case: Graph-based representations will effectively map your business, models, processes, dependencies, and interactions. Our advanced graph algorithms are designed to identify communities of bias, pinpoint key nodes that contribute to increasing bias, and uncover less secure paths that facilitate its spread—leading to fewer harmful models entering production.
Capabilities: Graphs, algorithms, queries, scalability, intricate relationship and weights modeling
Data Model: Nodes (vertices) with labels and properties (key/value pairs), and Relationships (edges) with type, direction, and properties (key/value pairs).

Advantages Over Relational Databases:

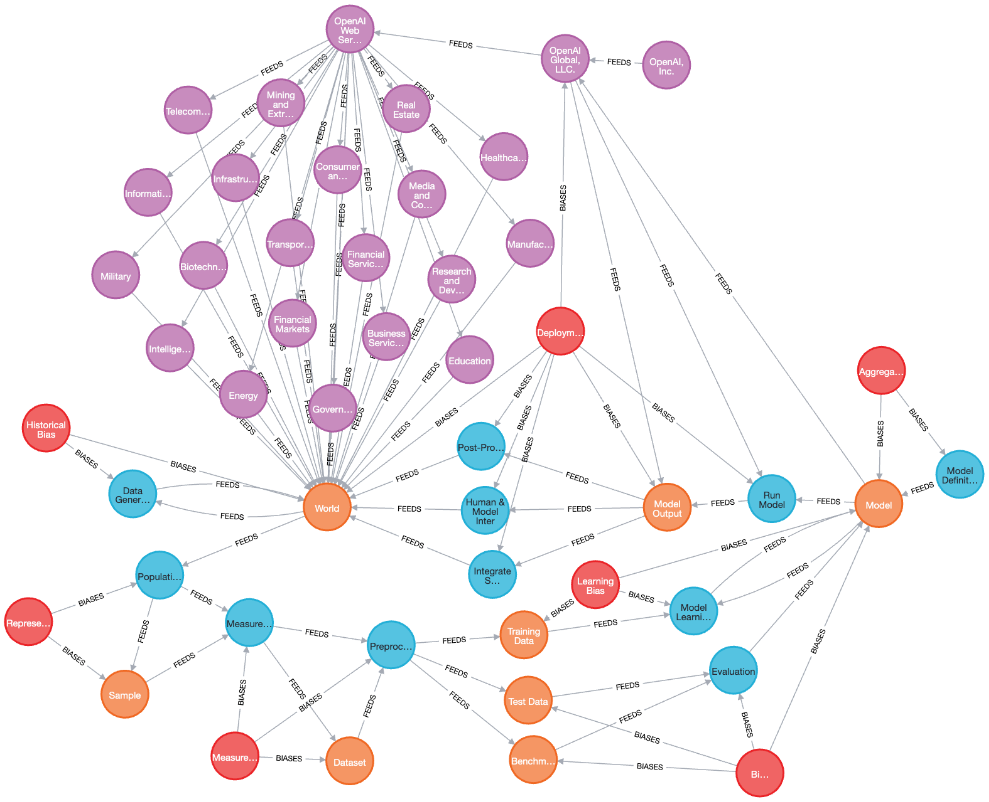
1. **Schema-less Relationship Handling & Performance:** SQL can require complex JOIN operations to handle relationships, Neo4j's flexible data model allows us to capture evolving features and relationships. As we learn and infer causal relationships producing bias from the increasing MongoDB repository, we can modify our graphs to support our current understanding of those forces in the network. Additionally, Neo4j can better support deep, recursive, relationships that would require multiple JOIN operations with SQL. This makes it better suited for network analysis.

42.4 Strategic Risk Analysis:

In our strategic analysis of high-risk operations, Neo4j will be utilized to map and analyze intricate relationships.

At first, we will focus on OpenAI's Organizational Structure and the bias throughout the Machine Learning Workflow.

This analysis will focus on identifying how biases propagate across model development, subsidiaries, and back to the world itself. We can detect areas most susceptible to bias and implement targeted mitigation strategies, using our MongoDB, Redis, and Graph solutions to measure improvements over time.



Types of Bias:

In the graph above, we show **four Node Types**, and **two Relationship Types**:

1. **Nodes:** Workflow Stages (in orange), Processes (in blue), Biases (in red), Company Organizations (in purple)
2. **Relationships:** 'BIASES' relate to nodes that are susceptible to a particular type of bias, and 'FEEDS' relate to Nodes that human or technical processes relate to.

While many types of Cognitive Biases exist, we will constrain our explanations to Bias in the Machine Learning Workflow at OpenAI, Inc. In general, Bias in Machine Learning can be categorized into **seven Categories**:

1. Historical Bias
2. Representation Bias
3. Measurement Bias
4. Learning Bias
5. Evaluation Bias
6. Aggregation Bias
7. Deployment Bias

Preliminary Risk Analysis Results:

We are excited to show you the preliminary results of our algorithms, which we've run in order to demonstrate the power of these methods to you, and the BFCTO. We will discuss **three Algorithms**, Degree Centrality, Betweenness Centrality, and Label Propagation.

Degree Centrality measures the number of direct connections a node has. When a node has more direct connections, high degree centrality can indicate how pivotal it is within the wider network.

Three areas of focus were identified:

1. OpenAI Web Services (OWS)
2. Deployment Bias
3. Evaluation Bias

Our degree centrality results highlight key nodes that are potentially influential in your network. These results will help the BFCTO use its limited resources to target the highest impact areas.

Second, **Betweenness Centrality** measures the extent to which a node lies within the shortest path between other nodes. Greater betweenness centrality can show how a node acts as a bridge within the network. 3 groups of nodes were identified as the most important:

1. The World facilitates the creation of, and receives, data. There appears to be a cyclical relationship between bias that is created in the world, which is then received by it again.
2. Population Definition & Sampling processes and actual Samples
3. Datasets, impacted through Measurement and disseminated through Preprocessing, Train/Test Splits prior to entering models through training.

These results indicate that lowering bias may come from a mix of internal process changes, and external advocacy in the world. While external advocacy is not a requirement of the mandated regulatory requirements, they may make a better world, and ultimately decrease how much bias our models capture.

Lastly, **Label Propagation** identifies clusters of nodes (or communities) within a network, based on connectivity through a particular label or relationship.

Our experiment propagated 'FEEDS' as a force throughout the network, as a proxy for understanding how communication, policy, labor, and technology flows through your company, processes, and the world.

Our results show two distinct clusters of communities: 1. OpenAI Global, LLC. and its subsidiaries like OWS, and 2. OpenAI, Inc. (the non-profit company and board overseeing the enterprise), and ML Workflow nodes and processes.

These results indicate that bias and harm reduction efforts could be most impactful by splitting efforts into two branches. First, OpenAI Global, LLC can lead operational improvements like improving communication, schema alignment, reporting cadences, and bias remediation workflows throughout OWS. Second, OpenAI, Inc. can decisively make modifications to ML Workflow policies that would subsequently be adopted by its subsidiaries.

Conclusion:

In conclusion, the BFCTO is well-placed to mitigate algorithmic bias and harm within OpenAI's operations. Not only can we adhere to new legal mandates, but also set a pioneering standard for the industry. Our comprehensive approach spans advanced data handling with MongoDB and Redis, to strategic network analysis using Neo4j. We've established a North Star for our project goals, a robust model catalog, real-time bias reporting and incident tracking, and strategic insights that catalyze continuous improvement. Our efforts reflect our dedication to equity and fairness in the world of AI, ensuring that OpenAI remains at the forefront of responsible AI development.