MAPFRE

FROM Insurance Perspective **IMPORT** Responsible AI Challenges and Opportunities

César Ortega Quintero
Expert Data Scientist, MAPFRE

CONTENT

01. INTRO

02. RESPONSIBLE AI TOOLKIT 🧰

1.1. How not to get lost in translation

1.2. Mostly harmful AI risks?

1.3. Regulatory landscape

03. LESSONS FROM MY JOURNEY 🤓

2.1. (AI)pocalypse now!

2.2. We’re at an organizational crossroads: are we up to it?

2.3. RAI by design

2.4. Embracing AInsurance

04. HOW I LEARNED TO (ALMOST) STOP WORRYING AND LOVE AI🥰

05. GLOSSARY

Intro

"Insurance is one of the key enablers of our new AI society," is the phrase I keep remembering from my first meeting with the Global Head of Disruptive Innovation when I joined MAPFRE back in the summer of 2022. At that time, I was assigned to a recently created workstream of Responsible Artificial Intelligence (RAI).

One year on and I'm now coming to the end of a scouting process in which we've conducted several Proof of Concepts to compare tier 1 RAI providers from Europe and the US by using some of our core models as guinea pigs, in areas such as Underwriting, Customer Lifetime Value (CLTV), and Fraud Detection, among others.

So what's AI? And what's RAI? Why is insurance the centerpiece of the new AI society? And what exactly is this new AI society? These are some of the important questions I'll try to answer here, but I also want to share some of the key lessons I've learned during my journey.

Without further ado, let's dive in...

Responsible AI toolkit 🧰

HOW NOT TO GET LOST IN TRANSLATION

An exhaustive list of concepts required to master RAI could be introduced (here's my personal selection), but let's focus on the most fundamental definition necessary to navigate this debate. What is AI? The reality is that we still don't have consensus on this. The definition of AI is a topic of discussion for several reasons, and here's my take on three main ones:

- **Complexity**: The concept integrates a broad spectrum of areas of knowledge, from rule-based systems, natural language processing (NLP), robotics, deep learning, among others, and each has its own assumptions, theories, and nuances.

- **Evolving capabilities**: The rapid development of AI means that capabilities are constantly expanding and outgrowing previous definitions. What was once considered AI, for example Optical Character Recognition (OCR), can now be seen as a non-AI commodity because our expectations of AI have evolved.

- **Emerging regulation**: Governments and international bodies are actively working to establish standards and legal frameworks for AI, reflecting the growing awareness of its impact on society in recent years. However, these regulatory efforts vary significantly from one region to another, and even within the same country, leading to diverse perspectives[1]. Normally, different legislations emphasize distinct aspects in terms of **which systems can be classified as AI**, **their autonomy, the role of humans, the nature of the outputs** (e.g., prediction scores, recommendation, unstructured content, etc.), **ethical considerations, accountability mechanism**, etc. In the article titled "Lost in Transl(A)t(I)on: Differing Definitions of AI"[2] you can find some of the key differences among AI definitions based on different legislation.

But you're probably still wondering what exactly AI is. Here's what I consider it as… (Remember this is still an ongoing debate. Please be kind 😊.)

> Artificial Intelligence (AI) is a multidisciplinary area of knowledge that combines theoretical principles of subjects such as mathematics, statistics, physics, computer science, graph structures etc. Its purpose is to develop physical and virtual machines

[1] https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation
[2]https://www.holisticai.com/blog/comparing-definitions-of-ai?utm_source=hai_linkedin&utm_medium=organic_post&utm_campaign=ai_definitions

capable of mimicking and improving human cognitive processes to perform all manner of tasks "intelligently", from prediction and optimizing decisions to content generation.

Note that this definition emphasized three distinct points: (1) the **multidisciplinary nature of AI**, (2) the pursual of it to **mimic and improve human cognitive processes**, and last but not least, (3) **to perform all manner of tasks "intelligently"**, with the quotation marks around the last word because I personally believe that the definition of intelligence itself is an even more complex debate.

MOSTLY HARMFUL AI RISKS?

Responsible AI (RAI) is a recently coined expression that refers to the **ethical**, **transparent**, and **accountable** development, deployment, and use of AI. It establishes a framework of principles to ensure that AI maximizes social well-being while minimizing potential harms derived from the fiery AI evolution we have experienced in the last decades.

Depending on the legislation (as discussed previously), the industry and the theoretical approach, risks can be named and grouped differently. For example: Performance or Efficacy, Robustness, Fairness & Biases, Transparency (Explainability & Interpretability), Security & Privacy, Reputation, Financial, Compliance - the list could go on much longer. For the purposes of this article, let's refer to the concise yet comprehensive list of **7 requirements suggested by the EU Guidelines for Trustworthy AI**[3]. I'll do my best to summarize them:

1. **Human agency and oversight**, or what I like to refer to as a "human-centered design approach". AI systems should respect human agency and fundamental rights by facilitating user autonomy, risk assessment, and human oversight. Users must make informed choices, and mechanisms for intervention and monitoring are crucial in protecting human autonomy.

2. **Technical robustness and safety**: Security measures are needed to protect against adversarial attacks and vulnerabilities of the system. A fallback plan, accuracy, reliability, and reproducibility further contribute to safety, transparency, and accountability in AI development and deployment.

3. **Privacy and data governance**: Privacy, a fundamental right, entails protecting data governance, ensuring privacy and data protection, data quality and integrity, and implementing controlled access protocols to prevent discrimination.

[3] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

4. **Transparency**: AI transparency, traceability and documentation of data, processes, and system decisions. Interpretability involves making AI systems understandable to humans and explainability refers to clarifying the reasons why the model arrived at a certain outcome. In addition, users need to be clearly informed when they are interacting with AI systems.

5. **Diversity, non-discrimination, and fairness**: Promoting inclusion and diversity, by involving stakeholders and ensuring equal access and treatment. Addressing the avoidance of unfair bias in data and system development.

6. **Societal and environmental well-being**: The impact of AI should extend to the broader society and the environment. Efforts should prioritize sustainable and environmentally friendly AI development and assess social and democratic implications.

7. **Accountability**: establishing mechanisms for responsibility and oversight of AI systems, both pre- and post-development. This includes enabling auditability, minimizing, and reporting negative impacts, addressing trade-offs, and providing accessible mechanisms for individuals, especially for vulnerable groups.

REGULATORY LANDSCAPE

Now that we've established the basics of RAI, its principles, risks, and requirements, let's quickly review the fundamental approaches of the regulation from a geographical perspective:

- **BRAZIL**: With a similar approach to the EU AI Act principles, Brazil has established that AI providers must inform users about AI interactions and offer explanations for AI decisions. Users can challenge AI decisions and request human intervention. Developers must conduct risk assessments, and the law classifies AI systems into high-risk and prohibited categories, with public disclosure of risk assessments. All AI developers are held liable for damages, with high-risk product developers facing rigorous liability standards[4].

- **CHINA**: China has just released a draft regulation for Gen AI[5], emphasizing adherence to "socialist core values". The regulations assign responsibility to developers for AI output and impose legal liability for the misuse of training data. These rules are being built on existing legislation related to deepfakes[6], recommendation algorithms[7], and data security.

[4] https://www.washingtonpost.com/world/2023/09/03/ai-regulation-law-china-israel-eu/
[5] https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/
[6] https://www.wsj.com/articles/china-drafts-rules-for-facial-recognition-use-4953506e?mod=world_lead_pos3
[7] https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-

- **EUROPEAN UNION** (risk-based approach): The EU AI Act is a comprehensive and transversal regulation aiming to ensure the responsible development, deployment, and use of AI technology. It categorizes AI systems into different risk levels. **Unacceptable risk** refers to those systems promoting harmful behavior or conducting social scoring, face bans, etc. **High-risk** AI, affecting safety and fundamental rights, will be subject to strict oversight and registration, covering areas such as critical infrastructure and law enforcement. **Limited-risk** AI must comply with transparency requirements, allowing users to make informed decisions. Last December the 8th, the EU Parliament finally approved the AI Law that should come into full effect by late 2026, with certain provisions, such as the ban on prohibited AI systems will be enforced within six months, and requirements for generative AI systems and models will be in place within 12 months[8].

  "Spanish Lab": Spain, in collaboration with the European Commission, is expected to launch the first AI Sandbox[9] in 2024 to experiment with future AI regulation requirements, including risk management, data governance, transparency, and cybersecurity for high-risk AI systems.

- **JAPAN**: Japan is taking a moderate approach, aiming to maximize AI's positive societal impact rather than suppressing it due to overestimated risks. The country is focusing on a risk-based, agile, and multistakeholder process, and sees AI as a stimulant for economic growth. Recently, the G7 has reached an agreement on regulating AI despite differing views, with Japan's 'Hiroshima Process on AI' proposal acting as a middle ground between the EU and US approach. The agreement involves creating a "Code of Conduct for AI" that developers must follow[10].

- **USA**: US legislation relies on a comprehensive ecosystem of state and local efforts focusing mostly on specific use cases. Some of the most notable include the Colorado Privacy Act (CPA), NYC Biometric Identifier Information Act[11], and Rhode Island Insurance Law, among others. Moreover, the federal government has recently introduced a new executive order on standards for AI Safety and Security[12].

management-provisions-effective-march-1-2022/

[8] https://www.consilium.europa.eu/es/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

[9] https://portal.mineco.gob.es/en-us/comunicacion/Pages/231002_sandbox_ia.aspx

[10] https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems

[11] https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3704369&GUID=070402C0-43F0-47AE-AA6E-DEF06CDF702A

[12] https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Does this evolving regulatory framework have a significant repercussion on insurers?

-> The short answer is Yes, it most certainly does!

Historically, insurers have being one of the most heavily regulated industries due to the impact of their activities on social well-being; Furthermore, it's worth remembering that insurers have capitalized for decades on using data and algorithms (actuarial modeling). Moreover, the emerging regulation not only directly affects its **core operations** (pricing, underwriting, claim handling, fraud detection, etc.), but the whole ecosystem of stakeholders including **supporting areas** (legal, innovation, recruitment) and **strategic partners** (automotive, biotech, insurtech). Based on this, and due to the significant uncertainty around potential hazards deriving from the latest AI advances, regulatory pressure has turned into an increasingly dynamic process with special focus on insurance.

Lessons from my journey 🤓

(AI)pocalypse now!

If we consider the lack of consensus around what AI is, the complex and ambiguous interpretation of its risks, and the rapid evolution of regulation in recent years, you could argue that we're in the middle of a perfect storm that could result in a complete dissociation between theoretical regulatory requirements and real-life solution.

Then, it becomes fundamental to find actionable solutions to control and mitigate AI risk and **potential harm to individuals** (e.g., physical damage, mental health, economic losses, digital security, privacy, dignity), **organizations** (e.g., financial and operational losses, regulatory sanctions, legal conflicts, reputational damage) and **society** as a whole (e.g., national security, economic instability, environmental impact, infrastructure integrity). It's worth taking a look at this interesting analysis on the subject: "'Dark Star' and the debate on AI"[13].

Indeed, we all should be rightfully aware of the historic moment we're experiencing in terms of **one of the greatest leaps forward for humanity and its potential impact on a social and economic paradigm shift**. But don't just focus on doomsayers' messages, as I just did, but strictly in the interests of raising awareness 😉 - let's take a more constructive approach and join me to disentangle some of the main RAI challenges and potential solutions.

We're at an organizational crossroads: are we up to it?

Adopting an honest Responsible AI philosophy in a multinational insurance enterprise with thousands of employees and hundreds of models is not an insubstantial challenge.

When we consider the multidisciplinary nature of AI and the need for experts in multiple fields, **the assignment of roles and responsibilities** may be one of the most relevant organizational challenges to deal with. It's crucial to establish clear lines of accountability to ensure that the necessary standards are upheld across the organization, especially nowadays where every employee is a potential AI user thanks to its democratization that Generative AI offers.

[13] https://www.mapfream.com/en/dark-star-and-the-artificial-intelligence-debate-2/

Promoting a culture that values ethical considerations in AI development and usage is even more important. This implies not only internal communication campaigns and RAI training for all employees, but also **truly integrating these principles into the company's core values and culture**.

RAI by design

The open-source mindset embraced by most of the AI community, from big tech companies to pre-seed startups, is the only path to go down when facing the technical challenges around RAI. For example, you can refer to the Responsible AI practices published by Google[14], with recommended strategists to design and develop AI from a responsible perspective, in other words, RAI by design. Here are my favorite recommended practices: "human-centered design approach", "understand the limitations of your dataset and model" (I would add "and use case" here), "Test, Test, Test", and "Continue to monitor and update the system after deployment".

As we discussed previously, AI encompasses various dimensions of risk that we must analyze separately if we are to find the right balance across them to ensure that AI systems not only accomplish their intended objective but also meet the expected ethical principles and regulatory requirements. However, finding this right balance could be a cumbersome task and leads me to raise the following question:

Is risk mitigation a zero-sum game?

To some degree, it is, in that improving one dimension of risk might lead to the deterioration of another. For instance, increasing transparency might compromise privacy and security. Similarly, focusing on robustness might affect performance. The inconvenient truth is that normally, a good amount of time and resources is needed to find that sweet spot where all risk dimensions meet desirable thresholds, and the model can still match the business's expected performance. This is why it's critical to triage the general level of risk of all the models so that the limited time of experts (data scientists, data engineers, business specialists, etc.) can be optimally assigned to improve truly sensitive models. There is an expression that I'm going to borrow from a workmate: **"Zero risk has infinite cost"**.

Let's share some key tips around the main dimension of risks:

- **Performance**: Academia and practitioners have dedicated extensive resources to developing concepts such as data drift and concept drift and establishing platforms to continuously monitor and update the systems. This realm is relatively mature in terms of methodological approaches, but if I were to give you just one piece of advice, it'd be to always evaluate your performance metrics contextually. There is no silver bullet or rule of the thumb that can unambiguously and

[14] https://ai.google/responsibility/responsible-ai-practices/

unequivocally be applied to all models. Instead, compare the results with the best alternative available (benchmark model) making sure to explore several metrics and always incorporating users' feedback.

- **Robustness** can be especially challenging given the unpredictable nature of real-world data and scenarios (adversarial fashion?[15]). It can be improved by incorporating techniques like data augmentation, adversarial training, or capsule networks to design systems resilient to unexpected inputs or deliberate manipulation (e.g., data poisoning).

- **Security and Privacy**: Proactively preventing data and AI models from suffering breaches is a vital task that is closely related to model robustness. Challenges arise in accounting for all possible scenarios and balancing proactive safety limitations with flexibility. Unfortunately, this is an active area of research that had introduced techniques (adversarial training, federated learning,) and protocols (ethical hacking, red teaming), due to the lack of practical defense system that can be smoothly integrated into production environments[16].

- **Transparency**: As AI systems become more complex, understanding their decision-making process becomes more difficult. The lack of interpretability and explainability is a significant challenge in many applications where understanding the reasoning behind a decision is crucial, especially in sensitive areas such as healthcare, justice systems, law enforcement, finance, or insurance. Happily, the open-source community has made significant progress in recent years by introducing local feature importance techniques (e.g., SHAP, Grad-Cam, LIME, DeepLift) or developing mechanistic interpretability techniques (e.g., representation inversions or Anthropic's work on decomposing Language Models[17]). However, the most important tip I can share for the development of a model is to follow the parsimony principle (Occam's razor). In other words, if a simpler model does the trick (e.g., decision trees or regression models), stick to it.

- **Fairness**: AI systems could inadvertently perpetuate or amplify existing biases present in the training data toward sensitive groups systematically. This is what I personally refer to as the "paradox of (un)conscious algorithm discrimination". Let me explain…

  Highly sensitive, protected attributes (e.g., ethnicity, citizenship status, race, religion belief, veteran status) are normally not available and never included in the models, except for gender and age, which are now excluded from new models following on from the recommendation of our legal teams. However, even if we try to meticulously build a model following RAI-by-design principles and best practices, there is a significant risk of incurring discrimination because, normally, models have less data to learn from minority classes such as immigrants and, even if

[15] https://www.theguardian.com/world/2019/aug/13/the-fashion-line-designed-to-trick-surveillance-cameras
[16] https://www.youtube.com/watch?v=Zd9kYgUjgSU
[17] https://transformer-circuits.pub/2023/monosemantic-features/index.html

these attributes are not included, their effects could be leaked by a proxy variable that we may not be aware of, such as a person's zip code.

Here lies the gist of the "paradox of (un)conscious algorithm discrimination" - not having access to these protected attributes and excluding them from the models is not only not solving the problem, but also masking a potentially harmful situation.

Ultimately, models learn to generalize by minimizing the **prediction error and we must ensure that prediction error across minority groups is not significantly different from the majority classes** based on the equality of opportunity approach. However, depending on the context of the use case, we may be interested in following the equality of outcome approach in which the likelihood of a desirable outcome is the same for members of each group, regardless of the prediction error.

In this context, we should conduct modeling bias analysis to identify and compensate any divergence in performance with respect to relevant bias-error metrics between protected and non-protected groups by following these steps:

a) Prepare data including sensitive attributes and data needed for performance metrics. However, as protected attributes are not recorded, we must opt for alternative approaches:

    a. If location data is available (zip codes, neighborhood, county), it's possible to infer protected attributes at an aggregated level (zip code) using official census demographic data.

    b. Alternatively, the protected groups at an observation level can be inferred by using specific algorithms (e.g., BISG) that uses full names and/or geolocation as input data.

b) Determine the maximum bias metric value or threshold for each sensitive attribute to consider the system free of bias and compute the bias metric values by comparing all groups against the most favored group for each sensitive attribute.

c) Finally, enhance performance within protected groups by identifying the source of the bias and introducing appropriate mitigation actions. **Pre-processing** (reweighting, disparate Impact remover, optimized pre-processing, learning data representations), **in-processing** (adversarial debiasing, exponentiated gradient reduction, grid search reduction or prejudice remover), and **post-processing** (equalized odds post-processing, calibrated equalized odds postprocessing, reject option classification)[18].

[18] https://holisticai.gitbook.io/roadmaps-for-risk-mitigation/mitigation-roadmaps/mitigating-bias-and-discrimination/step-2-mitigating-bias

The harsh reality is that a universally accepted concept of fairness doesn't exist – not in life and certainly not in AI. Identifying the right criteria for an AI system involves considering a wide range of factors, including user experience, cultural, social, historical, political, legal, and ethical background, some of which may be contradictory. Even in seemingly straightforward scenarios, individuals may hold differing opinions on what constitutes fairness, and determining whose perspective should guide AI policy can be a complex and ambiguous matter.

Embracing AInsurance

We're currently in the middle of a journey to understand AI risks and incorporate best practices and methodological approaches that will allow us not just to mitigate existing risks, but to proactively prevent them, i.e., RAI by design. Insurers must be top of the RAI class so they can eventually start supporting their clients and partners and ultimately offer the compensatory functions of an AI insurance product.

In this context, some insurers are already introducing policies that offer financial protection against AI incidents under very specific circumstances[19]. These policies address a range of potential risks associated with generative AI, including cybersecurity issues, copyright infringement, biased outputs, misinformation, and proprietary data leakage[20].

However, the main barrier to developing AI insurance policies is the absence of curated historical data for assessing the use and performance of AI models. To properly price an insurance product, it is necessary to quantify the frequency and severity of the incident in question, and even though we've seen some major effort being put into this, with the OECD AI Incidents Monitor[21] being a notable case in point, they're fairly recent. Furthermore, generative AI models are evolving rapidly, which means insurers need to develop dynamic risk assessment methods.

While AI insurance is becoming more prevalent, the role of insurers must start earlier in the prevention and implementation of robust AI guardrails, as insurance should serve solely as a last resort or safety net rather than the primary method for managing AI risks.

[19] https://www.armilla.insure/ or https://www.munichre.com/en/solutions/for-industry-clients/insure-ai/ai-self.html
[20] https://www.wsj.com/articles/is-your-ai-model-going-off-the-rails-there-may-be-an-insurance-policy-for-that-adf012d7
[21] https://oecd.ai/en/incidents-methodology

How I learned to (almost) stop worrying and love AI 🥲

As the adoption of AI models continues to grow at breakneck speed, general awareness of the inherent risks can be regarded as one of the central issues for society. We're experiencing a nuanced relationship with AI, acknowledging both the incredible potential it offers and the concerns surrounding it.

In an era when AI has become an omnipresent force, questions about the ethical implications of AI, its repercussions on economy, and its potential to increase inequalities have all contributed to a sense of unease. However, as we navigate these unfamiliar waters, we're also discovering the remarkable ways in which AI can improve our lives. For insurers embracing RAI, the journey from curiosity to understanding has just started, and our industry has the social commitment to be one of the standard bearers driving the responsible development and usage of AI.

OK so, that's it from me. If you've made it this far, I appreciate you tagging along for the ride. To sum up, let me leave you with the five essential takeaways from my journey:

- Zero risk has infinite cost.
- RAI has to be part of the core values and culture of any organization.
- The open-source community is our true ally in overcoming technical challenges.
- Having robust RAI guardrails in place must be the goal, and AI insurance should be our last resource.
- As with gradient descent, RAI is less about finding the optimal solution and more about taking small steps interactively in the right direction.

GLOSSARY

1. **Algorithm**: A set of computational rules to be followed to solve a mathematical problem.
2. **Artificial Intelligence**: The applications of computational tools to address tasks traditionally requiring human analysis.
3. **Average odd difference**: A statistical measure that quantifies the disparity in odds between two diverse groups or populations, often used to assess fairness in algorithms and decision-making.
4. **Back-testing**: A form of outcome analysis that involves the comparison of actual outcomes with a modeled forecast during a development sample period (in-sample back-testing) and during a sample period not used in model development (out-of-time-back-testing), and at an observation frequency that matches the forecast horizon or performance window of the model.
5. **Benchmarking**: An alternative prediction or approach used to compare a model's inputs and outputs to estimate from alternative internal or external data or model.
6. **Data proxies**: Data that are closely related to and served up in place of data that are either unobservable or immeasurable.
7. **Deep learning**: A subset of machine learning that utilizes neural networks with multiple layers (deep neural networks) to automatically learn and represent complex patterns from data, commonly used in tasks like generative AI, image and speech recognition.
8. **Bias**: is the differential treatment that results in favored or unfavored treatment of a person, group, or attribute
9. **Big data**: are datasets that are characterized by, at a minimum, their volume (i.e., size) velocity (i.e., speed of transmission), and variety (i.e., internal, external, including third-party data) that requires scalable computer architecture to analyze and model.
10. **Ethical hacking**: A practice in which cybersecurity experts, known as ethical hackers, attempt to identify vulnerabilities in computer systems, networks, or software to help organizations improve their security.
11. **Equality of opportunity**: A principle of fairness and justice that aims to ensure that individuals have the same opportunities and access to resources, irrespective of their background or circumstances.
12. **Equality of outcome**: A concept emphasizing equal distribution of resources, wealth, or opportunities to achieve a specific result or outcome, regardless of individual effort or merit.

13. **Explainability**: The extent to which AI decisioning processes and outcomes are reasonably understood.

14. **Human-in-the-loop (HITL)**: Refers to a system or process where a human is an integral part of the decision-making or operational loop. This means that AI systems work in conjunction with human oversight, and human intervention is available when needed.

15. **Interpretability**: Ability to understand and make sense of the internal workings of an AI model or algorithm, often in terms of the relationships between input data and the model's output.

16. **Machine Learning**: Method of designing a sequence of actions to solve a problem that optimizes automatically through experience and with limited or no human intervention.

17. **Model**: A quantitative method that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into output. Input data can be quantitative and/or qualitative in nature and the output can be quantitative or qualitative.

18. **Model overlay**: Judgmental or qualitative adjustments to model inputs or outputs to compensate for model, data, or other known limitations. A model overlay is a type of override.

19. **Ongoing monitoring**: One of the core elements of model validation. Ongoing monitoring confirms that a model is appropriately implemented and is performing and being used as intended.

20. **Outcomes analysis**: The comparison of model estimates and outputs to actual outcomes to help evaluate model performance by establishing expected ranges for actual outcomes in relation to intended objectives and assessing the reasons for observed variation between the two.

21. **Override**: Model output or input that is ignored, altered, rejected, or reversed.

22. **Parsimony principle**: the simplest explanation or hypothesis is usually the best when faced with multiple options. It encourages choosing the solution with the fewest assumptions or entities, promoting clarity and efficiency in problem-solving and scientific reasoning.

23. **Performance threshold**: A particular value or range of values of a performance measure or diagnostic that determines the acceptance or rejection of a model's performance.

24. **Poisson data**: In the context of machine learning, it refers to the manipulation of training data with malicious or deceptive inputs to compromise the performance and reliability of machine learning models.

25. **Protected group**: Refers to a category or demographic of individuals who are safeguarded by anti-discrimination laws and principles. These groups typically include those protected from

discrimination based on characteristics such as race, gender, religion, ethnicity, disability, age, sexual orientation, and more.

26. **Proxy variable**: Feature or characteristic that is used as a substitute or stand-in for a more sensitive or protected attribute when making decisions or predictions. AI systems often rely on proxy variables when the direct use of certain attributes, such as race or gender, would be discriminatory or raise ethical concerns.

27. **Red teaming**: A cybersecurity practice in which an independent team simulates attacks on a system, network, or organization to identify vulnerabilities, evaluate defenses, and improve security measures.

28. **Responsible Artificial Intelligence**: Refers to the ethical, transparent, and accountable development, deployment, and use of AI systems and technologies. It encompasses a set of principles, practices, and guidelines that aim to ensure that AI benefits individuals and society while minimizing potential harm and risks. Responsible AI involves considerations of fairness, transparency, accountability, bias mitigation, data privacy, and ethical decision-making throughout the AI lifecycle. It seeks to strike a balance between technological advancement and the well-being and rights of individuals, addressing issues like bias, discrimination, safety, and the impact of AI on society.

29. **Root Mean Square Error** (RMSE): A statistical metric that quantifies the average difference between observed and predicted values, often used to assess the accuracy of predictive models and algorithms.

30. **Social well-being**: Multidimensional concept that assesses the quality of life and overall welfare of individuals and communities. It encompasses factors such as economic prosperity, physical and mental health, social relationships, access to education, safety, and the cultural and environmental aspects of a society, aiming to measure the level of happiness and satisfaction experienced by its members.