## Article

# Algorithms and Trustworthiness in Tax Administration

**Ignacio González García**

Dr. Ignacio Gonzalez Garcia has over 35 years of professional experience in Tax Administration, serving in roles as Inspector, Deputy Director of Customs, and Director of the IT department at AEAT. Currently, he is a senior officer at the National Office for the Investigation of Fraud (ONIF). His academic qualifications include master's degrees in civil engineering and business administration, along with Ph.Ds. in Philosophy, Psychology, Engineering and Mathematics, specializing in Artificial Intelligence. Dr. Gonzalez Garcia has contributed his expertise to the International Monetary Fund (IMF), the Inter-American Development Bank (IDB), CIAT, and the OECD.Email: igmigonzalezgarcia@gmail.com

**Salvador Duart Crespo**

Mr. Duart is a freelance consultant specializing in customs modernization, risk management, and trade facilitation, with over 25 years of experience across public, private, and multilateral sectors. He has led reform initiatives across Europe, Asia, the Caribbean, and South America. He has advised institutions such as the Inter-American Development Bank, the Asian Development Bank, and European Commission on the integration of AI, data-driven decision-making, and regulatory innovation in Customs. His work supports the development of digital ecosystems aligned with EU standards and WCO guidelines. Mr. Duart holds degrees in Finance and Computer Science from George Mason University. Email: salvadord@bdctec.com

Received 03 December 2023, Accepted 01 October 2024

KEYWORDS:

Artificial Intelligence, Machine Learning, Tax Administration, Algorithmic Decision-Making, Bias and Fairness, Public Trust, Risk Analysis, Explainability, Governance, Transparency

ABSTRACT:

This article examines the integration of algorithms into the work of Tax Administrations. It argues that these tools are not neutral instruments but reflect historical biases and institutional choices. While they offer opportunities for greater efficiency and consistency in areas such as fraud detection and service provision, their use also raises complex legal, ethical, and governance challenges.

The authors explore how algorithms, particularly those based on machine learning and artificial intelligence, reshape decision-making processes. They show that although these technologies can implement efficient and trusted methods, their use as digital civil servants requires careful oversight. Through practical examples, such as VAT fraud detection and AI-assisted taxpayer guidance, the paper highlights both the potential and the risks involved.

Tax professionals, the authors argue, must play a central role in defining the objectives, assessing the limitations, and ensuring the ethical use of algorithmic tools. Algorithms should serve as support systems, not replacements for legal reasoning or institutional judgment.

The paper concludes that the challenge is not whether to adopt algorithms, but how to govern their use responsibly, balancing innovation with the duty to uphold public trust, fairness, and accountability.

PALABRAS CLAVES:

Inteligencia artificial, aprendizaje automático, administración tributaria, toma de decisiones algorítmica, sesgo y equidad, confianza pública, análisis de riesgos, explicabilidad, gobernanza, transparencia

RESUMEN:

Este artículo examina la integración de los algoritmos en el trabajo de las administraciones tributarias. Sostiene que estas herramientas no son instrumentos neutros, sino que reflejan sesgos históricos y decisiones institucionales. Aunque ofrecen oportunidades para mejorar la eficiencia y la coherencia en áreas como la detección del fraude y la prestación de servicios, su uso también plantea desafíos legales, éticos y de gobernanza complejos.

Los autores exploran cómo los algoritmos, especialmente aquellos basados en el aprendizaje automático y la inteligencia artificial, están transformando los procesos de toma de decisiones. Señalan que, aunque estas tecnologías pueden aplicar métodos eficaces y confiables, su uso como "funcionarios digitales" requiere una supervisión cuidadosa. A través de ejemplos prácticos, como la detección del fraude en el IVA y la asistencia automatizada a los contribuyentes, el artículo destaca tanto el potencial como los riesgos implicados. Los autores argumentan que los profesionales tributarios deben desempeñar un papel central en la definición de objetivos, la evaluación de limitaciones y la garantía de un uso ético de las herramientas algorítmicas. Los algoritmos deben funcionar como sistemas de apoyo, no como sustitutos del razonamiento jurídico ni del juicio institucional.

El artículo concluye que el desafío no es si adoptar o no los algoritmos, sino cómo gobernar su uso de manera responsable, equilibrando la innovación con el deber de preservar la confianza pública, la equidad y la rendición de cuentas

Mots Clés:

Intelligence artificielle, apprentissage automatique, administration fiscale, prise de décision algorithmique, biais et équité, confiance du public, analyse des risques, explicabilité, gouvernance, transparence

Résumé :

Cet article examine l'intégration des algorithmes dans le fonctionnement des administrations fiscales. Il soutient que ces outils ne sont pas des instruments neutres, mais qu'ils reflètent des biais historiques et des choix institutionnels. Bien qu'ils offrent des possibilités d'amélioration de l'efficacité et de la cohérence, notamment dans la détection de la fraude et la prestation de services, leur utilisation soulève également des enjeux juridiques, éthiques et de gouvernance complexes.

Les auteurs explorent la manière dont les algorithmes, en particulier ceux fondés sur l'apprentissage automatique et l'intelligence artificielle, transforment les processus décisionnels. Ils montrent que, bien que ces technologies puissent mettre en œuvre des méthodes efficaces et fiables, leur utilisation en tant que « fonctionnaires numériques » exige une surveillance rigoureuse. À travers des exemples concrets, comme la détection de la fraude à la TVA et l'assistance automatisée aux contribuables, l'article met en lumière à la fois le potentiel et les risques associés. Les auteurs affirment que les professionnels de la fiscalité doivent jouer un rôle central dans la définition des objectifs, l'évaluation des limites et la garantie d'un usage éthique de ces outils algorithmiques. Les algorithmes doivent servir de systèmes de soutien, et non se substituer au raisonnement juridique ni au jugement institutionnel.

L'article conclut que le véritable enjeu n'est pas d'adopter ou non les algorithmes, mais de savoir comment encadrer leur usage de manière responsable, en conciliant innovation, confiance publique, équité et redevabilité.

Contents:



## 1 INTRODUCTION

Tax professionals must develop a new kind of relationship, previously unnecessary, with algorithms. This relationship is not simply a passive one of knowledge, use, or acceptance. It should be active, performative and one of its attributes must be trustworthiness. The purpose of this paper is to define and describe these algorithms and their relationships.

Various branches of mathematics are applied in the tax field. To cite basic examples: calculus is used to compute the tax amount based on the base and rate; statistics, as the science of the state, is used to collect and analyze data; and algebra helps calculate interest on payment deferrals. In recent decades, advanced analytics have become widespread in selecting taxpayers, organizing audit plans, and helping to better plan the provision of services to taxpayers. With the advent of machine learning and artificial intelligence (AI), new tools have emerged to solve complex problems, such as responding to taxpayer queries without human intervention.

Algorithms, understood as sets of instructions organized to perform a task, are used in all of these activities. Some are simple; others are highly complex. When algorithms are used to replace, wholly or partially, tasks traditionally performed by humans, it becomes necessary to examine how they affect taxpayer rights and the responsibilities of tax administrations. Before we can do this, however, we must first understand what algorithms are, their nature, and how tax professionals interact with them.

In Section 2, we explore the concept of algorithms, the concerns surrounding their use in taxation, and the criteria for evaluating and accepting them. Section 3 addresses the practical use of algorithms in taxation through two case studies, identifying their challenges and key implementation phases. Section 4 focuses on the evolving relationship between tax professionals and algorithms.

## 2 ALGORITHMS

### 2.1 THE CONCEPT OF AN ALGORITHM

The word "algorithm" originates from the Arabic term al-Khwarizmi, later translated as Al-Juarismi and Latinized as *Algorithmi*. It was part of the name of the mathematician Abu Abdallah Muḥammad Ibn Mūsā al-Khwārizmī. "Khwarizmi" refers to the region of Khwarazm (present-day Uzbekistan), where he was born in 780 A.D.

When he was invited to work in the House of Wisdom in Baghdad, he authored, among other works, *Kitab al-Mukhtasar fi Hisab al-Jabr wa-l-Muqabala* (The Compendious Book on Calculation by Completion and Balancing). The term "al-jabr" is the root of the modern word "algebra."

In tribute to al-Khwarizmi, mathematics uses the term "algorithm" to describe an unambiguous procedure for solving a class of problems by following a sequence of instructions. A cooking recipe, for example, is also an algorithm. But to fully grasp the use and implications of the term in taxation, we must move beyond this simple illustration, which, while helpful, is ultimately too reductive.

### 2.2 SOLVING PROBLEMS WITH ALGORITHMS

Broadly speaking, there are three ways to solve a mathematical problem: through a flash of insight—or a "happy idea"—, through brute force and using a good method, an algorithm.

Let's consider an ancient problem, not for its practical importance, but to illustrate the concept in an engaging way. As early as 3100 BCE, Egypt systematically collected crop taxes. In the Rhind Papyrus, the scribe Ahmes who, like other scribes, was exempt from taxes, posed problems and offered solutions to assist those responsible for collection. In this sense, their role was like that of today's data engineers.

Problem 24 reads: "An amount and one-seventh of it gives a total of 19. What is the amount?" [Kline, 1990]

Someone could guess that the answer is 7 and see that it is too low:

$$7 + 7/7 = 8$$

He could try to solve it by *brute force*, by trial and error but the crude calculator should think better, because for the problem "What is the quantity whose triple is seven?", if he tries to do it, he would be calculating until the Day of Judgment, since the result is a periodic fraction.

Ahmes looks for a method and creates an algorithm. He approaches the problem by systematically splitting the difference of results in parts—perhaps inspired by the task of dividing land after the Nile floods receded, since tax was calculated based on land area and flood levels. He has a happy idea, to look for a multiplier.

He creates a table with pairs (multiplier of and result of the guess) beginning with (1; 8) Too low. He doubles the result (multiplier is 2). Still too low. He halves the original result (1/2;4). The total of the two parts is 20—too much. So, he tries a smaller fraction—one-quarter of 8, which is 2—Too low and builds a table

| Multiplier | Result | Decision |
|---|---|---|
| 1 | 8 | Too low |
| 2 | 16 | Too low |
| 2+ ½(8) | 16 +4 | Too high |
| 2+ ¼(8) | 16+2 | Too high |
| 2+ ⅛(8) | 16+1 | Too low |
| 2+1/4+1/8 | 16+2+1 =19 | Done |

He sees that 19 is 16+2+1. The corresponding multiplier of 8 should be 2 + ¼ + ⅛ = 2.375 and the number that satisfies the equation is: $7 \times 2.375 = 16.625$

This solves the problem and establishes a method—a specific algorithm for solving *this* problem. Centuries later, al-Khwārizmī would generalize this by expressing it algebraically:

$$x + x/7 = 19$$

Here the happy idea is to mix letters and numbers. The Arabic term for "the unknown" was al-shalan, which had no Spanish equivalent. Mathematicians in the Iberian Peninsula began using "x" to represent it—meaning "the thing". Solving the equation, that is using an algorithm that isolates "x", gives:

$$x = 133/8 = 16.625$$

In Don Quixote's time, an algebraist was someone who knew how to put disjointed or broken bones in their place, like the algebraist in mathematics, who is the one who puts the *x* (the thing) in its place in each different case. The student who is taught to solve systems of two equations with two unknowns is taught to apply an algorithm.

Algebra made it possible to use one method to solve many problems, with more general algorithms, rather than a separate algorithm for each result.

In summary, a mathematical algorithm is a set of instructions, many times found by a happy idea, that uses rules from a branch of mathematics to express what to do and when to stop, because the result achieved is sufficiently close to the desired result. Many of them should be known by computer scientists. They use textbooks, like the classical of T. H. Cormen "Introduction to Algorithms", that has 1,312 pages, with these mathematical recipes, like chefs use "*Le guide culinaire*" as a reference with useful ideas.

### 2.3 ALGORITHMS IN THE TAX FIELD

In the application of tax regulations, mathematics has traditionally been used prudently but without the need for sophisticated techniques. After all, taxation is not "rocket science". In the second half of the 20th century, however, the idea began to take hold that advanced analytics, like those used in *Business Intelligence*, could improve tax control. With the development of artificial intelligence in the 21st century, the notion emerged that these technologies could also be applied to the tax field, both for facilitation and control.

The use of mathematical algorithms to calculate in the tax field worried no one. But their use in ADS (*Algorithmic Decision Systems*) raised concerns for many, because it was well known that it can cause visible and claimable damage, such as the unfair denial of a mortgage loan. Problems arise, therefore, when algorithms are used not just to compute, but to decide.

Some authors foresaw insidious damage, such as the improper categorization of taxpayers by fiscal risk, which in turn can lead to more harmful categorizations and to legal and ethical problems.

### 2.4 MERCATOR'S MAPS RELOADED

The concept of map conceived as a projection is a powerful one. For mathematicians, functions are understood as maps.

Gerardus Mercator, in 1569, created a method for projecting points from a sphere onto a cylinder, allowing for the creation of highly useful maps. It is well known that this model distorts reality, modifying distances, especially near the poles, but it is valuable because it allows users to chart paths from one point to the another by following rhumb lines (loxodromes), which are lines that cross meridian at a constant angle. On the map, they appear as straight lines. *Matrix Reloaded* (2003) was a successful sequel to *The Matrix* introduced the intriguing concept of "reloading", which allow us to imagine projections that might make us live in a distorted reality.

Tax Agencies use conceptual maps to chart paths from their current state to the goals they have identified. The activity of these modern Tax Agencies can be improved by the art of numbers and reloaded through the art for vectors.

Tax Agency managers must both enforce tax rules and facilitate taxpayers' compliance. They are tasked with reconciliating efficiency with fairness, using rhumb lines, such as risk analysis, in their plans of control. By applying criteria like "risk analysis," they not only optimize their course of action, but also justify it, shielding themselves from accusations of arbitrariness or bias.

It is well known that Tax Administrations' data warehouses contain partial, and sometimes distorted, versions of reality. Risk analysis often amounts to a quantification of the voice of the experts. Yet it remains the best approach to find an equilibrium in a world overwhelmed with data, infected of "*dataism.*"

In the applied sciences, there are those who practice the art of the line, such as architects, among them Gaudi, of whom we have photos hanging weights from ropes to define the anti-funicular loading of the vaults of the *Sagrada Familia*. There are those who practice the art of the number, like Luca Pacioli. A new type of practitioners has emerged: those who practice the art of the vector, like A.I. engineers.

Data scientists provide Tax Agencies with experts in the art of numbers when they use Machine Learning to classify, estimate the probability of fraud or predict variables, and in the art of vectors when generating answers. As navigators once understood the distortions of their maps, it is equally useful for us to understand the limitations and distortions in these models.

### 2.4.1 Machine leaning and the art of numbers

Machine learning is a branch of Artificial Intelligence that uses algorithms and statistical models to achieve a goal without relying on specific, case-by-case instructions.

Ahmes, in the Rhind papyrus, created 87 algorithms. For 3600 years mathematicians created algorithms to solve specific problems.

In a seminal paper "Behavior, Purpose and Teleology", Rosenbluth and Wiener (1943) opened the field of "cybernetics", the study of circular causal process as feedback and recursion, where the effects of an action (its outputs) return as input to that system to influence subsequent actions. One simple implementation of the idea is a thermostat.

Arthur Samuel (1959) coined the term *machine leaning*. He developed a program to play checkers using an algorithm to score positions by estimating the chances of winning. The program applied the minimax algorithm, which combined stored evaluations of past positions with current scores to determine the next move.

A few years later, the idea emerged of combining cybernetics with algorithmic methods to create a system capable of winning any game. The idea was to create a type of generic "map" (algorithms) with feedback that could recursively use its outputs as inputs until reach a desired value (like a thermostat).

The idea is to create a map, place the destination point in the floor, and our current position above it, then let a marble fall. In AI and ML, *gradient descent* is the algorithm that implements Ahmes' idea: to reduce, step by step, the difference between the current state and the objective. Its strength lies in its generality, though the underlying idea is simple.

Tax Agencies have used machine learning in the past for two main purposes, classification and correlation. Classification comes in various forms: binary, when taxpayers are divided into two groups (e.g., those included in an inspection plan and those exempted); multi-class,

when categorized into several groups (e.g., risk levels in Customs); and continuous, when assigning a probability of risk to each taxpayer on a scale (e.g., 1 to 100). Correlation involves discovering and quantifying the relationship between features (such as declaration data) and an attribute (such as the presence of fraud).

These processes have a statistical base (e.g., probability). The precision and the efficiency of decisions adopted using machine learning is great, but it has a natural limit. We don't necessarily obtain better results when estimating a regression using machine learning compared to the formula we learned in school. However, if we need to use regression to identify a subset of 100 taxpayers to be controlled in a universe of 100,000, machine learning is much faster.

When we try to classify something, such as pets, into groups, we face two opposing goals: simplicity and precision. If we create only two categories, for example cats and dogs, we fail to account for other types of pets, like turtles. The same happens in the tax domain: if we classify taxpayers only as fraudsters or honest, we ignore involuntary errors and other nuanced situations. Even if we decide to simplify and focus only on cats and dogs, using a single variable like weight will not be enough. Nor will weight and color together necessarily improve accuracy. The important point is that Machine Learning allows us to classify faster and, in some cases, more accurately, but in this context, the improvement over traditional methods is often only marginal.

When we map reality, such as taxpayers, into a field like fraud or not fraud, we are projecting many dimensions onto just two. We can only see shadows, easily created by AI, but still just shadows.

If we consider the aspect of correlation, for instance between declared and input data in tax returns, there is a limit in the number of variables that can be used. Once this threshold is surpassed, a phenomenon known as *overfitting,* occurs. In such cases, any increase in precision comes at the cost of generality, meaning the conclusions drawn from the training data can no longer be reliably applied to other cases.

Tax Agencies can now apply statistical tools to billions of data points and obtain marginal gains through advanced models. The real breakthrough lies in the ability to use complete datasets instead of samples, applying these tools to all taxpayers rather than just a few. However, we do not anticipate a radical transformation in the management of existing taxes using Machine Learning. What this technology may enable is the creation of new and more efficient forms of taxation.

By contrast, the art of vectors will bring about a true sea change.

## 2.5 ARTIFICIAL INTELLIGENCE AND THE ART OF VECTORS

### 2.5.1 From words to vectors

In the 1960s, Terry Winograd and others developed rule-based methods for machine translation, working alongside linguists who manually crafted rules for computers to process language. This approach produced unnatural translations that were easily recognizable as machine-generated, as translations were done sentence by sentence and the connections between sentences were often awkward.

In the 1980s, these systems shift towards statistical methods. Natural language processing (NLP) algorithms begin to learn (art of data) from actual language data. Computational linguists had the idea of making groups of two words (*n*-grams) and counting their frequency in thousands of texts (brute force). The idea was that in the phrase "Can I have a glass of...?" it far more likely to end with the word "water" than with the word "stone." Their idea was to use statistics to make decisions instead of relying on fixed rules. The introduction of large

text corpora and the rise of the internet in the 1990s, with resources such as the Penn Treebank, provided an unprecedented volume of data for training NLP systems. However, this abundance of data required a new approach. Researchers quickly realized that when moving from pairs of words to full phrases, the number of possible combinations became unmanageable.

Yoshua Bengio and his team set a new precedent for language processing in 2001 by using models based on feed-forward neural networks. Some early proponents of A.I. believed it was a good idea to create a mathematical object, an artificial mathematical neuron, and they called the set of such neurons a neural network.

Let us consider a toy example of a mathematical neuron that tries to distinguish motorcycles from horses using data. The available variables are weight, color, and tail. Suppose we have the following input: weight (200 kg), color (black = 1), and tail (no = 0). The system must decide whether the object is a horse or a motorcycle. To do this, it must assign a level of importance to each factor—its parameters.

We need a map that takes the input (200, 1, 0), representing a 200-kilogram, black, tailless object, and maps it to 0 if it is a motorcycle, or to 1 if it is a horse. If we assign three parameters such as a = 0, b = 0, and c = 1, this toy neural network will classify correctly, because anything with a tail (c = 1) is a horse, and anything without one is not.

Some mathematicians specialized in the "theory of representations", experts in relating fields of mathematics that, in principle, no one suspected were connected, came up with the idea of converting the words into lists of numbers. The publication of the Word2Vec paper (2013) introduced a groundbreaking algorithm.

We can think of a vector as an arrow, with a length, a beginning and an end. (e.g.: in Cartesian axis from point (0,0) to the point (1,1). The mathematician knows how to say things about those arrows by calculating from the list of numbers: the sizes, the angle between the vectors, and the distance between the points. He operates with them by adding them, doing various types of products such as the scalar and the vector product. NVIDIA's success has been because it manufactures chips that do these operations very efficiently—just as Intel's success came from doing things by brute force with numbers.

The problem was to take a sentence, like "I'm going to have some ---- with the gin and tonic," and find the missing word without any consideration of the meaning or syntax.

They represented words in the realm of vectors:

- Capital of France is Paris $\rightarrow \vec{v}_1 + \vec{v}_2 = \vec{v}_3$
- Capital of Italy is Rome $\rightarrow \vec{v}_1 + \vec{v}_4 = \vec{v}_6$
- Capital of Spain is… $\rightarrow \vec{v}_1 + \vec{v}_7 = ?$

Capital = $\vec{v}_1$, France = $\vec{v}_2$, Paris = $\vec{v}_3$ and so on.

They took all the words from a dictionary, created a vector for each word, inventing the coordinates at random. Afterwards, they "trained" the system by reading many texts. If they found the word *milk* in a text "near" of coffee or cheese, they moved the point of the arrow in the middle, changing the previous coordinates. Based on many tests, the words eventually found their place in that context, because certain words tend to be used together. In a data space with hundreds of dimensions instead of three, the point of the arrow can find the right position.

They had another *happy idea*. Doing something like $\vec{v}_1 = \vec{v}_3 - \vec{v}_2$, the system can find the vector of the term that represents a concept.

These methods treated words in isolation. The attention mechanism, that was introduced in 2017 in the paper, assigned weights to each word based on its *relevance to the current task*. In 2018, the concept of LLM (Large Language Model) appeared, which applies this same idea using much more brute force. In fast succession, the Sequence-to-Sequence Modelling (2014) and the paradigm of Neural Models in Translation Services (2017) provided by Google Translate marked a pivotal move away from statistical models, offering a more nuanced and accurate translation.

In 2018, OpenAI introduced the first Generative Pre-trained Transformer (GPT), known as GPT-1, a neural network with 117 million parameters. Many millions instead of the 3 in our toy example, but the idea is similar. It was the first model trained in a "generative" mode by masking *portions of input text from left to right*. Actual version tunes 1 trillion of parameters (with an insane consumption of electricity).

In 2025, there are many large language models (LLMs) available. A programmer can now build one with relative ease (Raschka, 2024). However, only large companies have the resources to train these models, given the challenges of accessing the necessary data and the financial cost of using the vast number of processors required to compute trillions of parameters. As a result, Tax Administrations must rely on tools that have been trained in other contexts.

### 2.5.2 Usefulness of LLM and RGA for the tax expert

The reader is no doubt familiar with ChatGPT and other similar products. Due to the way they are trained it responds reasonably well to questions related to some contexts (things that appear on the Internet), and poorly on other topics. They can generate answers with *meaning*. This is different from remembering the truth obtained by experience. We find often inaccuracies in citations, the references offered are often not correct, it creates them instead of retrieving them, like bad students in an exam. We call these responses *hallucinations*.

To improve the truthfulness of the answers, companies have developed techniques known as RAG (Retrieval-Augmented Generator). The idea is that, after "understanding the question" that is, subtracting vectors, the system queries another system that contains specialized data not available on the internet, such as the Tax Administration's databases. This allows it to retrieve the specific information needed and generate a more accurate response, avoiding generalities or redirecting the user elsewhere. RAGs function as "master courses" for clever but naïve generalists like LLMs.

In this context, several questions arise for Tax Administrations:

- Should they use LLM tools? If so, in which areas?
- How much effort and resources can they invest in building and training a RAG?
- How should they interpret and apply the results?

It is claimed that OpenAI alone has spent approximately $7 billion on training and running large language models (LLMs), including up to $1.5 billion on staffing. Other analysts estimate that operating ChatGPT costs around $700,000 per day. Many companies are investing, or burning, billions of dollars in this technology. For now, the practical advantage of using LLMs remains uncertain, yet numerous companies, led by very intelligent managers, are already promoting their adoption. It is also true that many of these same leaders were recently promoting the multiverse.

It will be necessary to decide when to begin using them and at what pace. For this, we must differentiate, given the current state of the art, between two tasks. The first is being able to interact with the taxpayer in their own natural language. This problem has been solved.

The second task is to generate a proper response once the question is understood. This requires a deep understanding of legislation, not only to create a phrase with meaning. For example, if someone ask: "What and how should I pay VAT for selling Peruvian products in Madrid?" we should consider internal and external VAT and many other circumstances.

As LLMs and RAGs continue to advance, their answers will increasingly resemble those of a human, more fluent, more coherent, and more contextually appropriate. However, this apparent intelligence does not guarantee greater accuracy, nor does it make these systems suitable for making decisions. In the end, they may answer better than humans, but they will still not be capable of deciding.

## 2.6 EVALUATION OF AN ALGORITHM

We have seen that an algorithm *implements a method in a context* after it has been trained on data and directed toward a defined purpose.

### 2.6.1 Implementation and the virtue of Efficiency

We use the example of sorting, where both the sorting of a list of numbers and the ranking of taxpayer risk are relevant cases. There are many ways to sort numbers. The Wikipedia entry on "sorting algorithms" contains a comprehensive list of methods and their characteristics, such as whether they are efficient or stable. To choose the optimal algorithm, the technician must consider factors such as whether the dataset is large or small, whether it is already partially sorted, and whether there is sufficient space to store intermediate results. In most cases, they simply choose a popular and reasonably well-rated method.

When we ask an Excel spreadsheet to sort the values in a column, we don't worry about which algorithm it uses. The tax professional doesn't need to worry about this either. Data engineers will handle the technical details, but managers should have a general understanding of the underlying costs involved.

### 2.6.2 Context and the Virtue of Construct Validity

We use the example of sorting, where both the sorting of a list of numbers and the ranking of taxpayer risk are relevant cases. There are many ways to sort numbers. The Wikipedia entry on "sorting algorithms" contains a comprehensive list of methods and their characteristics, such as whether they are efficient or stable. To choose the optimal algorithm, the technician must consider factors such as whether the dataset is large or small, whether it is already partially sorted, and whether there is sufficient space to store intermediate results. In most cases, they simply choose a popular and reasonably well-rated method.

When we ask an Excel spreadsheet to sort the values in a column, we don't worry about which algorithm it uses. The tax professional doesn't need to worry about this either. Data engineers will handle the technical details, but managers should have a general understanding of the underlying costs involved.

## 2.1.2. Context and The Virtue of Construct Validity

We should evaluate internal and external validity.

The decision-maker faces a very common problem in the social sciences. If we give an intelligence test to several children, does it really reflect intelligence? If we use two versions of the test, do they yield the same score? If two children get the same score but have very different personalities and strengths, what does the number mean? In our field, similar issues

arise associated with the concept of risk. Is there any point in measuring tax risk with these techniques?

The social sciences work with concepts that are not directly observable, called constructs, such as neuroticism in psychology, quality in economics and risk in taxation. When analyzing taxpayer data, we may find omissions or inconsistencies: undeclared income reported by third parties, anomalous values such as extremely low revenue from companies with many employees, and so on. These indicators may allow us to quantify "tax risk," in much the same way that a doctor might conclude there is a health risk based on blood test results, or a dictator might see "social risk" in not joining a demonstration, or a psychologist may diagnose neuroticism based on various responses.

#### 2.6.3 Use of Algorithms and The Virtue of Prudence

We have seen the limits of AI algorithms. They do not possess intelligence or feelings. So, what are the limits of their use? We will examine this question through two practical cases that illustrate why prudence should be the guiding principle.

## 3 USE OF ALGORITHMS IN TAX PRACTICE

We will analyze two use cases: one for tax control (Detection of fake invoices) and another for tax facilitation (VAT Assistant). This is a case of machine learning, where a set of algorithms is classified by risk invoices or taxpayers.

### 3.1 CASE 1: DETECTION OF FAKE INVOICES

Tax experts know that some companies can generate false invoices, enabling taxpayers to deduct undue VAT amounts. One of their goals is to detect them and design an algorithm that is more efficient and effective than manual methods detecting fake invoices or more precisely, identifying both issuers and recipients of such invoices, seems possible. They must consider the modalities and objectives of each type of fraud.

A table like the one below could be created to organize them:

| | Issuer | |
|---|---|---|
| Receiver | Fictitious / Insolvent | Real |
| Fictitious / Insolvent | Creators of false invoices and users.<br>Fraud scheme: Creator of invoices and of false structures and comparators (Issuer and Receiver)<br>Identity theft (Issuer) | Improve financial statements to obtain financing (Issuer)<br>Bank discounting of false invoices (Issuer) |
| Real | Reduce VAT payable (Receiver)<br>Obtain refund (Receiver) | Mask actual payments (Receiver)<br>Justify subsidies (Receiver)<br>Disloyal employee (Receiver) |

*Table 1 - Causes for issuing false invoices*

#### 3.1.1 Selecting data

Tax organizations have collected data from many sources: imports, sales, records, censuses, sanctions, management; that can be processed and cross-referenced. They include indicators, such as whether past fraudsters were previously or later debtors, or defaulted;

whether the company was recently created (perhaps with the purpose of defrauding and almost no capital); and calculated magnitudes (e.g., the percentage of mismatches relative to turnover, or tax due relative to equity), along with other contextual data (sector, goods, deductions, etc.).

Ernesto Cardenal memorably tells how, at a time when he lived in a Carthusian monastery, Gethsemani Abbey, he volunteered to participate in studies that led to the identification of cholesterol as a risk variable by cardiologists. The researchers compared cholesterol levels in two controlled populations: the Benedictines, who ate meat and eggs, and the Carthusians, who were vegetarians. These researchers began with observations (descriptive statistics in epidemiology) some communities had more heart attacks. They then developed a causal model (*let's see if diet is the cause*), gathered more descriptive statistics (*some groups eat more fat and eggs),* modeled the problem, considered variables like blood lipids, and conducted tests (confirmatory analysis). It was necessary to have a population to study and two types of specialists.

Likewise, in our field, we need both data engineers and tax experts. The first step is to decide which taxpayers will be included in the study, always based on historical data, and which variables, among the thousands available to the tax administration, will be analyzed. The reader will encounter two schools of thought: one claims that with Big Data, the more data is used, the better. Its defenders argue that even variables like fat intake, zodiac sign, or mother's name should be included, because once you choose to trust the system, you shouldn't restrict it. The second theory holds that it is better to work from a model. Let's choose the second.

We could, in fact, include variables like the taxpayer's age (more fraudsters may be middle-aged than elderly or children), or whether they attend religious services (there may be fewer among priests, as they do not typically run businesses), or their zodiac sign. These ideas might slightly improve test results, but when the inspector goes to audit the company, neither age nor zodiac sign will help prove that the invoice is false, just as it wouldn't help a doctor to know the patient's birthplace, birth month, or mother's name (even if those correlate with health outcomes, due to dietary differences across social groups or the presence or absence of healthy habits). This process is known as feature selection and is critical in machine learning.

A good hospital uses AI to support healing, just as a good Tax Administration uses it to detect fraud. In both cases, the goal is not merely to predict the probability of death or fraud, but to fulfill their broader purpose. We should use AI to identify the best course of action, and then decide, as doctors do.

### 3.1.2 Analyze the relevance of the historical data with which you are going to train the algorithm.

Data scientists must train the model in our case with chosen historical data, i.e.: cases in which VAT fraud was found or not found. But how do we know whether false invoices were involved in those past fraud cases, and if so, who issued them? If we don't separate out the cases with irregularities that were triggered by false invoices, we will detect something else. In we have not, in our historical databases the detail needed to validate the model, as doctors didn´t have cholesterol data of past generations, how is it possible to validate our models? We need to review the historical data. This task can be done manually, or not. To train a system properly, we need thousands of data points, which means thousands of audit reports.

There are three options:

- Doing it wrong, seeing what happens, and hoping it is better than doing nothing.
- Doing it right, which requires hundreds of hours of work by tax specialists.

- Doing it with the help of artificial intelligence—in which case, you now have two problems instead of one.

In short, training an algorithm to answer a broad question like "Has this taxpayer committed fraud?" is easy. However, narrowing it down: "Have you committed VAT fraud as a result of participating in carousel fraud by creating front companies and selling mobile phones at a reduced price?", makes it much harder. Often, the challenge is not finding the right algorithm but having the right data to train it.

When tax administrations process electronic invoices, the amount of data is overwhelming, hundreds of millions of invoices provided by millions of taxpayers classified in hundreds of categories. To train a system to detect fraud in a category we need hundreds of cases for each category and thousands for the population model. If our historical database contains only dozens of cases for categories with the needed detail, the effort will fail.

In this case, we can use algorithms for other purposes, such as classifying taxpayers by groups and this will provide insight to auditors. This is the approach of the tool developed by the Inter-American Center of Tax Administrations (*Centro Interamericano de Administraciones Tributarias CIAT*).

### 3.1.3 Training the predictive model

Now suppose that, after a great deal of effort, a model has been identified, a reasonable objective selected, and enough historical data assembled, and the IT department has created the files needed for the analysis. Data engineers should perform.

a) *Descriptive and exploratory analysis of variables.* Data scientists will study each variable (e.g.: volume of sales) that will be analyzed separately to determine whether it relates to fraud (correlation). If not, it is a waste of time. And if several variables correlate with the same thing, the model should be simplified.

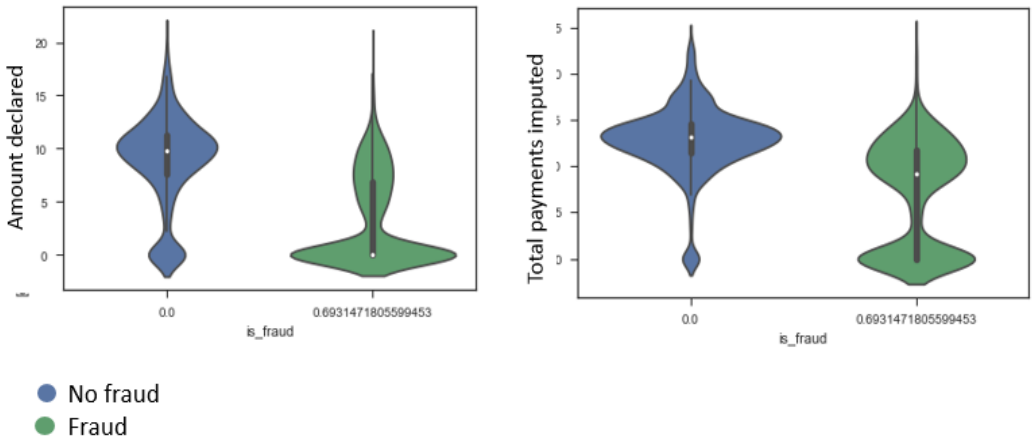


*Illustration 1. Description of features*

b) *Segmentation or clustering of cases and variables.* This involves grouping past fraud cases, like how doctors realized that dividing people into "those who eat eggs" and "those who do not" was not particularly helpful. Statistics can support us in defining meaningful clusters.

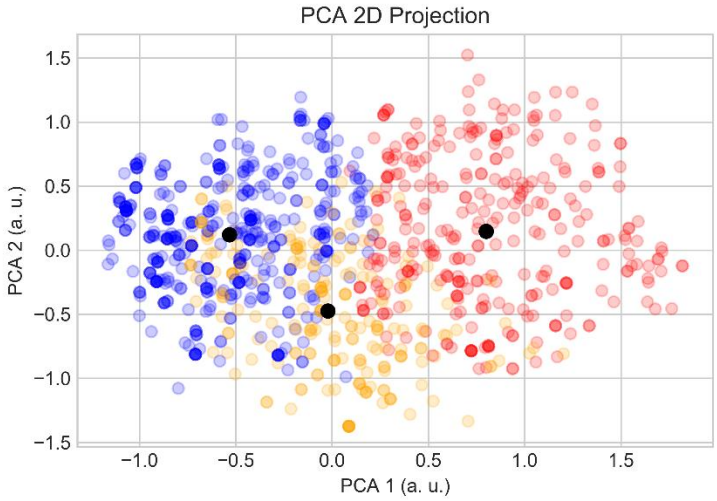


*Illustration 2. Example of segmentation in groups of taxpayers*

c) *Obtaining a representative sample of non-fraudulent cases*. To train a binary classification model, we need a set of taxpayers who did not commit fraud, across all economic sectors and business sizes, proportional to real-world distributions.

We want to transmit that, in the discovery of the importance of cholesterol, doctors need to study statistics and that tax experts, if they want to use these tools, will need to do the same.

After this, data engineers, using the dataset provided by the computer scientist, will apply various ML or AI techniques. These include different algorithms, aiming to predict whether the case involves fraud based on the variables provided. This is done based on experience, just as an engineer, when designing a bridge, knows that depending on the span and support options, certain solutions (arch, cable-stayed, suspension) are typically more efficient or cost-effective. Similarly, the data engineer may choose among neural networks, logistic regression, support vector classifiers (SVC), decision trees, and others. Sometimes, the chosen solution isn't the most efficient, but the most impressive is selected to showcase technical ability or to push a technological boundary.

### 3.1.4 Training predictive models

A tax expert must decide on the quality and soundness of the results. Let's assume a classifier with two outcomes: YES or NO. In a perfect world, the algorithm would say YES when the case is indeed fraud and NO otherwise. Accuracy would be 100%—but this never happens. To understand the nature of the errors, we use a confusion matrix.

Suppose we are trying to detect fraud—or cancer. To evaluate an algorithm, we look at:

- **True positives** (says YES, and it is yes)
- **True negatives** (says NO, and it is no)
- **False positives** (says YES, but it is no)
- **False negatives** (says NO, but it is yes)

In the case of cancer screening, many **false positives** may cause unnecessary fear for patients, but at least the condition is flagged. Many **false negatives**, however, could result in patients being diagnosed too late.

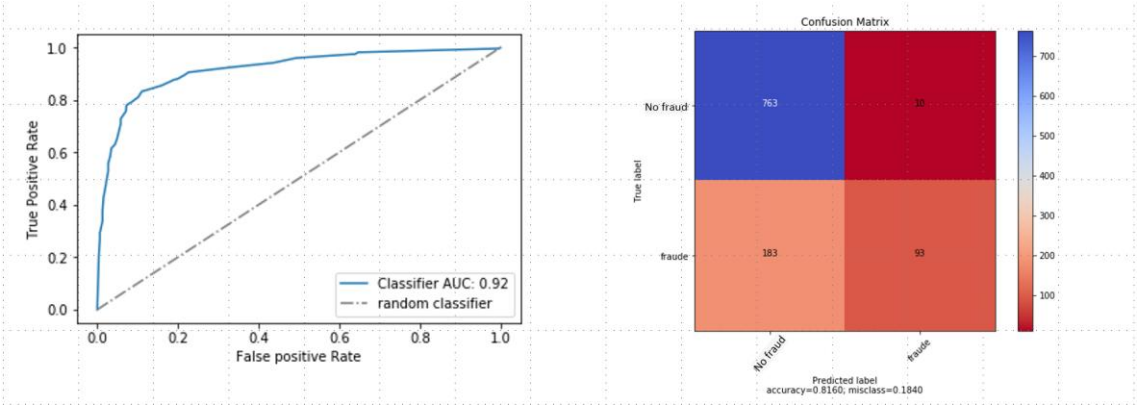


*Illustration 3. A practical case of the use of ROC in AEAT (Spain)*

In the case presented in Illustration 3 from the Spanish Tax and Customs Administration (AEAT), there were 763 true positives and 93 true negatives. In the remaining cases, the model's predictions failed. Specifically, for 10 taxpayers, fraud was predicted but not confirmed by inspectors (false positives), and in 183 cases, fraud was detected by inspectors but not predicted by the model (false negatives). The total accuracy of the model was 0.816.

#### 3.1.5 Usefulness to the expert.

The expert must assess whether what the algorithm says is useful. A magical algorithm that perfectly predicts whether a young person will die before age 50 may be great for an insurance company—but useless for a doctor. The death might come from a car accident, and if it is inevitable, the doctor cannot intervene. An algorithm is only useful to the tax professional if it explains how the fraud was committed, and which variables triggered the alert (a "white box" approach). It is far less helpful to hear that Bank A has a 2% higher score than Bank B in a model that predicts Corporate Income Tax fraud—even if that statement is true.

There are two main families of techniques:

- Those that offer *interpretable* solutions (e.g., decision trees), and
- Those that produce *"black box"* results (e.g., neural networks).

Each administration has its own rules for selecting taxpayers for inspection. In general, available resources limit intensive inspections.

Suppose a system is built to predict the existence of false invoices. Out of a billion invoices, it selects one in every 100,000—10,000 in total. Are there enough resources to check those 10,000? Would this replace the cases that, until now, were recommended by expert judgment?

Are the experts being told to conduct a "statistical experiment" instead of choosing companies they believe—with reason—have committed fraud?

The use of algorithms must be framed within a structured tax control plan.

### 3.2 CASE 2: VAT ASSISTANT

#### 3.2.1 Difficulties

Now we look at the second scenario. The problem in this case is very different. Instead of a problem of classification, with a solution based in the use of statistics using machine-learning

tools, the problem now consists in the use of a NLP tool (Natural Language Processing) that will be used to "understand" questions of the taxpayers with the aim of generating an answer.

Tax administrations need in this case three tools:

a) A tool that can "understand" the text of the question proposed by the taxpayer. This is solved with an A.I. product that interprets natural language, such ChatGPT:
b) A model to interact with the user to determine precisely what the question is. For example, if someone asks: "What should I pay for selling Peruvian products in Madrid?" the system must request the necessary data from the taxpayer.

The answer should distinguish between import duties and internal taxes. For imports, the system should ideally begin by asking for the value of the merchandise. Is it below the exemption threshold? Ultimately, the system would determine the TARIC code by interacting with the taxpayer, retrieve the applicable tariff rate, estimate the tax due, and provide an appropriate response. Afterwards, the system could assist the taxpayer with the technical aspects of domestic VAT.

In the example shown in Illustration 4, we demonstrate how experts construct a decision tree outlining the specific questions that must be presented to the user in each case, to clarify their query. Once the system, through a series of interactions, has identified the question, it can provide an answer. This answer may be pre-stored (e.g., A1, A2, etc.) or generated dynamically using an LLM-RAG system.

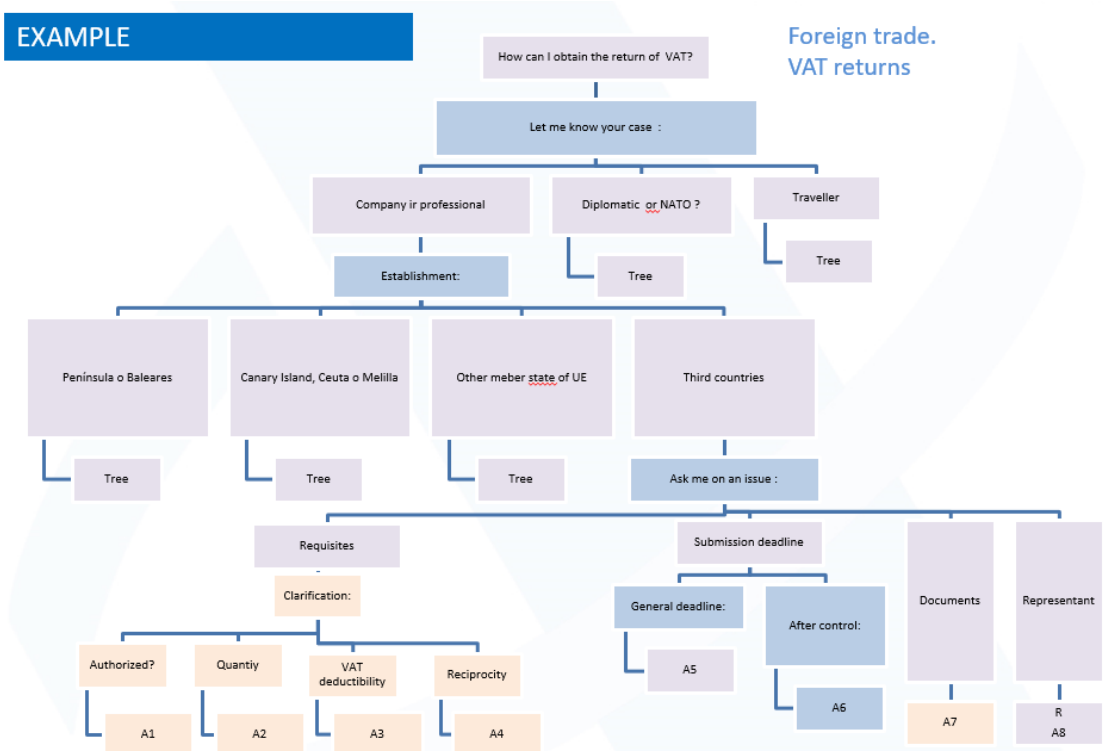


*Illustration 4. Tree aimed to clarify a question*

In short, we should need:

a) a generic tool trained on the Internet is used to understand the taxpayer's language.
b) a knowledge database with questions that we need to know before the generation of a response that must be made by experts; and
c) a generator of answers.

These answers could be either pre-stored or generated dynamically using a tool similar to ChatGPT.

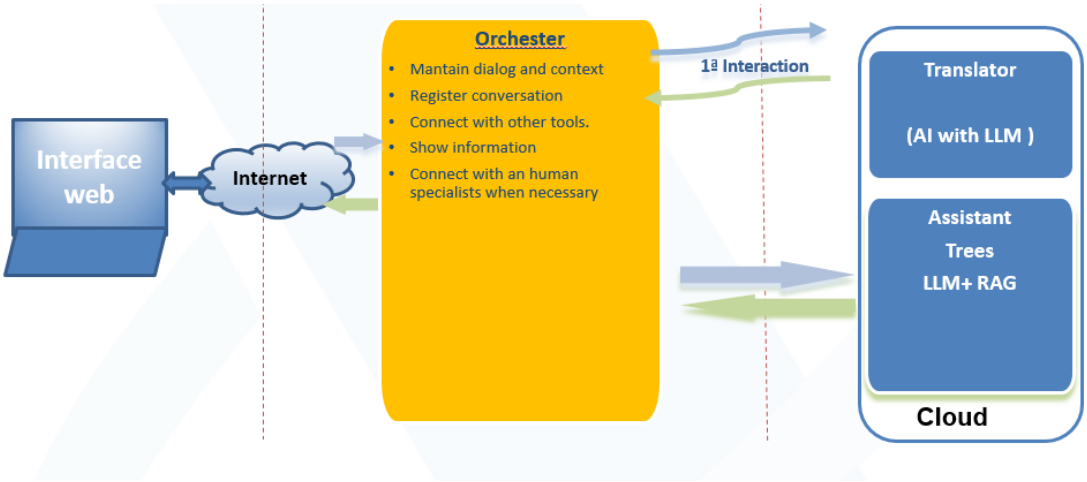


*Illustration 5. Architecture of a tool to facilitate VAT information*

## 4 RELATIONSHIP OF THE TAX EXPERT WITH THE ALGORITHMS

Ross King (2008) in his beautiful book "Brunelleschi's Dome: How a Renaissance Genius Reinvented Architecture" describes how, when the people of Florence decided to build their extraordinary cathedral, they summoned Brunelleschi, among others. He proposed a solution and succeeded in building it, but it was the citizens who chose to take that risk, financed the project, and rejected alternative proposals.

Data engineers and AI companies can propose systems of algorithms and build them, but the one who must decide whether to fund and then use them is the tax expert. Their relationship with the algorithm must be oriented, critical, and trustworthiness

### 4.1 ORIENTED

To orient oneself, to know where "east" is, we need to know the position of the sun. But that alone is not enough when using a map, and algorithms are, in a sense, maps. We also need to know approximately where we are. Ultimately, all of that is only useful if we also know where we want to go.

Tax experts must know where they are and what they want. French philosopher Gilles Deleuze said that it is not difficult to achieve the things we desire—the real difficulty lies in desiring well. For example, in the case discussed earlier, tax experts must be clear on whether they are looking for those who use false invoices or those who issue them. They must also ask themselves whether they are turning to AI to detect VAT fraud because traditional approaches have failed and they are hoping for a technological miracle, or because they have heavily invested in electronic invoicing and now see AI as a fashionable solution to justify those past investments.

A hospital can use AI to predict death, estimate the cost of treatment, or help heal patients. Likewise, AI in tax control can be used simply to compare the relative probability of fraud in corporate tax between two banks, or to generate insights for inspectors who must carry out the audits. The second use is more useful and more difficult.

### 4.2 CRITICAL

Kant acknowledged that Hume had awakened him from his *dogmatic slumber*. This led him to develop his *Critique* and to establish limits on the use of reason. It made him ask: What can I know?

We must ask ourselves the same question: What can I know based on my representations, my models, and the data I can verify?

Francis Bacon, in Novum Organum identified "idols that besets men's minds" and corrupt the understanding. Nietzsche in *Twilight of the Idols* wrote about ideas that are admired but are false. We claim here against some idols.

*AI can detect defrauders*. Algorithms are useful because they exploit regularities. Some of them are abstract, for example, an algorithm could detect triangles in pictures with precision. Events that are produced by a set of common causes manifest sets of regularities that are known as patterns. Some features of patterns can be easily quantified and are studied by statisticians using concepts as distributions. Machine learning algorithms are very effective in the fields of *pattern discovery (*e.g.: what are the regularities that are common to cases of carousel fraud) and in *pattern recovery* (e.g.: what are the companies that are elements of structures that share a given pattern of fraud). AI can identify patterns, correlations, and structures, but not individual decisions. Algorithms can classify, and we should not use them as "fraud-tellers".

*AI can give answers.* Large Language Models use statistical techniques to generate statements with meaning. The relation with the user is *antiphonal*. Phrases prompted by the user are followed by phrases generated with a *mimesis* of the past, mixing texts found on the Internet. Companies seem to play the role of God in Jeremias 33:3 "Call unto me, and I will answer thee, and shew thee great and mighty things, which thou knowest not" but this claim is not true. AI can recover old answers and blend them in statements with meaning, but without an "observer", that could act as a guardian of the truth.

*AI algorithms can learn to detect fraud*. A madman is said to have tried to always use the same match, because the first time he used it, it burned extraordinarily bright. Using data from old audit reports might help detect past frauds, but not new ones, like those related to cryptocurrency trafficking, plastics, or novel strategies devised by fraudsters. AI systems need good models trained with many well-labeled data. Believing that a good algorithm can be trained with very little data is misguided.

If we want to distinguish dogs from cats of any breed and age, we need thousands of photos, covering various sizes, ages, coat colors, and breeds. Similarly, if we want to find false invoices after receiving hundreds of millions of documents over years, historical data from a few hundred invoices or companies will not suffice. No matter how impressive the algorithm seems, it will learn poorly if not trained with enough data. Otherwise, time is wasted until sufficient data becomes available. But even in cases trained on large datasets, AI can detect regularities, yet it cannot distinguish the truly new from mere noise. Philosophers like Whitehead, Badiou, and Deleuze have studied the concept of the *event*—the moment or space where something genuinely new emerges. This is a domain beyond pattern recognition, where discernment, context, and creative thought are essential—capacities AI does not possess.

### 4.3 TRUSTWORTHINESS

Trustworthy AI has three attributes: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective. [EC, 2019].

Tax Administrations must face novelty in values, philosophical ideas and errors.

### 4.3.1 New shoots of old trees. Values

New technologies have forced us to consider new aspects of old values, like privacy, which in turn has shaped the boundaries of legitimate technological use. It is now accepted that not everything effective for preventing a crime is allowed: the police cannot listen to a phone conversation in a booth without judicial authorization, either by placing an ear to the wall or by intercepting mobile communications. There are limits on what questions may be asked and what methods may be used to uncover the truth and in the use of data collected for specific purposes.

With the rise of artificial intelligence, new values emerge because of new social values.

Thomas Piketty and Michael Sandel in "Equality: what it means and why it matters" maintain a compelling dialog thinking on the value of *equality*. Modern societies, interconnected, have been permeated by the ideas of the philosophers of *difference.* Tax experts must deal with taxpayers that honor a Janus revived, an idol with two faces.

We will explain this idea with a well-known example; is the case in which the Dutch tax administration was challenged in court in The Hague. The administration had used a system approved by law, for legitimate purposes, with anonymized data and positive outcomes. It relied on the detection of risk patterns using historical data from 17 sources that had previously been siloed. The court's ruling led to the fall of the government.

The reader is likely familiar with the court's arguments. The case was initially framed as a privacy issue (ironically, given the data was anonymized). This is an Aristotelian concept that was shaped in 1891 by the American lawyers Samuel Warren and Louis Brandeis that described *the right to privacy* in a famous article: it is the right to be let alone. The ruling ignored this approach and focused instead on *transparency* (the risk-profile algorithm was a black box) and *opportunity* (the system was used exclusively to detect fraud presumably committed mostly by racial minorities). We want to focus not on the legal dimension, but on the *philosophical* one.

The Hague court held that it is a *necessary condition* that both the taxpayer and Tax Administration must *understand* the pattern by which someone is selected for control.

This generates a problem. When the first AI systems were trained to play chess, it quickly became clear that coaching them with rules was ineffective. A machine taught by a human could never play better than its coach—or put less generously, it would be just as bad. By abandoning that method, we have reached systems like AlphaZero, which learn by playing against itself. No one understands how it arrives at its conclusions, but it plays better than any human.

The Hague court's prudent decision sets a clear boundary: the use of algorithms must be understandable not only to taxpayers but ultimately to the court's least skilled contributors. The guiding principle is this: it is legitimate to decide freely, as the law permits discretion, but once a specific criterion is used, it must be authorized by law.

In any case, this is not a major obstacle for Tax Administrations, as many machine learning tools can meet this requirement. Moreover, a dedicated field, *Explainable Artificial Intelligence (XAI),* is actively developing solutions to ensure transparency and interpretability. Therefore, the challenge in this area is largely reduced to the legal limits that govern the public disclosure of administrative actions.

Explainability is a necessary *but not sufficient* condition.

### 4.3.2 New modes of old ideas. Decisional realism

There is a debate so old as philosophy between nominalists, that believe that there are only particulars, this or that horse and realist that think that the Idea (with more ontological density that a concept) of "horseness" does exist with performative effects.

A group of individuals can be grouped by many criteria, age, sex, gender, weight, race, wealth and each of them can consider him or herself as being part of some group by a psychological mechanism of "identification". Tax Administration face a society (universal) that is seen for each of their members as a mix of "minorities".

Taxpayers feel that there are individuals and members of decisional minorities. This is the realm where AI must be used because a new idol, fairness, has many adepts.

### 4.3.3 New modes of error

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism algorithm is a tool used in the US criminal justice system to assess the risk of an individual recidivating: reoffending after being previously arrested. ProPublica, a non-profit investigative journalism newsroom, performed a study of the Risk scores assigned to 7000 people arrested in Broward County, Florida in 2013 and 2014[1] From their findings they claimed that the COMPAS algorithm was biased against Black arrestees. [Patalay, 2023]. Many others questioned this analysis. [2] [Christian, 2021]

A "universe" is composed of individuals (for example, twelve). If these individuals are not identical, each subgroup will have members with different characteristics. Suppose there are four men and eight women, including two foreigners and ten nationals. If we want to assign six grants and choose sex as the criterion for equity, we might decide to allocate two grants to men and four to women.

However, it could happen that all selected recipients are nationals. In this case, one of the foreigners might claim to have been discriminated against. When we divide a population into groups, it is mathematically impossible to ensure equity across every criterion simultaneously.

No single cause, tool, or decision can guarantee fairness across all conceivable criteria in societies that rely on grouped decision-making. To idolize such an ideal would render decision-making impossible.

When the poor performance of facial recognition algorithms in identifying certain ethnic *minorities* was detected, Google issued apologies recognizing the *bias*. At the same time a strange but well-intentioned cynicism arose since these systems are widely used by U.S. police to identify suspects, and because convictions disproportionately affect some minorities, it might be better that the system fails. But the case is that every cause has a bias, but not all bias is denounced by a minority. In the tax field it would be very easy to build artificial minorities to reject controls.

Our point is that:

a) There is *always* a bias when we apply the *one* to a group of particulars.

[1] https://medium.com/@lamdaa/compas-unfair-algorithm-812702ed6a6a

[2] William Dieterich, Ph.D. Christina Mendoza, M.S. Tim Brennan, Ph.D. Performance of the COMPAS Risk Scales in Broward County (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.

b) That it is impossible to reach statistical fairness in societies embedded for decisional realism.

c) That Tax Administration should not invest effort in avoiding bias. Their aims should provide a legal base and transparent knowledge of the precision of their tools and a legal base to their limits (tolerance).

### 4.4 NEW LIMITS TO THE GOVERNMENTAL ACTIONS

Charles Reich was a professor at Yale Law School that wrote an article: "*The new Property*". He argued that administrative agencies should not have more right to deprive a citizen of a privilege. It was a seminal article that aimed at an actual problem.

The Hague court sustained the concept of *opportunity* (the system was used exclusively to detect fraud presumably committed mostly by racial minorities). Was the court evaluating the opportunity or rejecting a form of "bullying"?

This criterion could be understood in three ways.

*Specificity.* We believe that this criterion does not imply that an AI tool cannot be used for a specific purpose, such as controlling the evasion of taxes on private swimming pools.

*Transitivity*. It may be that the Court considered the possibility that the Tax Administration was targeting a specific attribute of a particular minority. In such cases, the Tax Administration must ensure that, when addressing relationships such as "fraud by this minority," its tools and strategies are designed to detect all forms of the broader category of fraud (for example, income omission, which can occur across all groups), rather than focusing on a mode of fraud that is characteristic of a specific group (such as obtaining a subsidy intended for that group).

If we accept this reasoning, it would have dramatic implications for the control of spending policies. It would be inconsistent to design policies specifically tailored to a minority group yet prohibit the use of those same data for oversight and control.

*Arm's Length.* In the past, students were trained and encouraged to respond independently to the hardships of life and to endure initiatory rites. Today, the mood has shifted, and many behaviors that were once tolerated are now recognized and classified as bullying.

Some analysts are beginning to view the use of AI on low-income taxpayers as a form of institutional bullying. Implicitly, they are raising a pointed question: if technology is so powerful, why not apply it to aggressive tax planning or multinational corporations, instead of targeting retirees and minority groups?

### 4.5 THE USE OF EPEXEGETICAL GENITIVE

When we read the phrase *"the Spirit who is truth,"* the genitive *"of truth"* is epexegetical. It does not indicate that the Spirit possesses truth but rather explains or defines what the Spirit is—He is truth. Similarly, in Spanish, *"el tonto de mi marido"* does not mean that my husband possesses someone foolish, but that he is foolish.

In contrast, when we use the phrase *"the bias of this algorithm,"* we are doing the opposite. We imply that the algorithm is biased—always—and this is not accurate. Let us return to Ahmes: an algorithm is either correct or incorrect. Bias appears only when the algorithm is trained on biased data or evaluated using criteria that are shaped by subjective or self-identified decisions. In such cases, the data are affected, not the algorithm itself.

Consider an AI algorithm designed to calculate the correct dosage of medication needed to reduce blood pressure by three points. Among four available medications, it correctly

identifies the best one and its optimal dose. No one objects, even if the training data did not include variables like coffee or salt intake, because the output serves a clear and individual medical objective.

But if the same AI algorithm is used to determine which individuals should be admitted to a program based on their grades, the situation changes. Salt intake will not demand the right to be considered as a medicine, nor claim that it was unfairly rejected in the past. But a minority group, even one defined through administrative or political decisions, can raise such a claim.

In short, when we speak of bias, we are not truly referring to the algorithm itself, but to the training data—and to the social reality that data reflects. The use of an algorithm forces us to confront uncomfortable truths about that reality.

Tax Administrations do not exist merely to maximize fraud detection—not even within legal limits. Their role is more complex, just as scholarship committees today are no longer solely focused on rewarding the most talented or highest-achieving students. The cold clarity of mathematics demands precision, but that clarity is often politically inconvenient.

There are cases in which Tax Administrations themselves have acknowledged that, in the past, inspections disproportionately targeted political opponents or vulnerable, low-income taxpayers, while ignoring large fortunes or those who benefited from unfair and disproportionate exemptions.

Should such an administration avoid using AI because its historical data are biased?

The short answer is:

- No, when it comes to facilitation.
- Yes, when it comes to control, unless it can clearly distinguish whether its past data are genuinely biased or simply selected based on valid risk analysis criteria.

## 4.6 SCYLLA AND CHARYBDIS

We identify possible consequences of these new values that are highly undesirable.

### 4.6.1 Paralysis and lack of alignment

It may happen that Tax Administrations, which were initially active in the use of these technologies, withdrew out of caution and, to avoid complications, revert to using risk indicators with methods that only slightly differ from traditional techniques, stepping away from true machine learning.

If this happens, taxpayers will encounter an ossified administration, one that fails to offer them the qualitative leap that the previous generation provided through digitalization.

### 4.6.2 Constant litigation

On the other hand, if the administration promotes these technologies to:

- offer binding responses,
- assign probabilities of success to each legal position in disputes with taxpayers,
- build arguments adapted to the profile of the taxpayer and the courts involved,
- rely on historical data and employ "nudges" or behavioral economics strategies,

In that case, the administration will face increasing litigation, including strategic lawsuits, fueled by suspicions of *manipulative or illegitimate* use of technology, promoting the idea of

an "arm's length" that could oblige Tax Administration to avoid the control of minorities by the fact of its numeric condition.

## 5 CONCLUSIONS

The growing use of algorithms in the work of Tax Administrations is not simply a matter of technological modernization and a consequence of digitalization. Their use in decision-making and in generating responses represents a deeper transformation, one that touches on law, ethics (González, 2024), governance (González and Duart, 2024), and public trust.

Throughout this article, we have shown that while algorithms offer opportunities for greater efficiency in tax administration and can implement efficient and trusted methods, their use as "digital civil servants" faces many problems.

Tax professionals should be the architects of the new solutions. Like the Florentines who evaluated Brunelleschi's daring dome design, they must be capable of making informed decisions, orienting themselves toward clear goals, aligning technology with institutional values, exercising critical judgment, and upholding ethical standards. Ultimately, they must decide whether to accept or reject the proposals of data engineers, fully understanding and assuming the associated costs.

Avoiding both the Scylla of institutional paralysis and the Charybdis of unchecked automation requires a measured, deliberate approach. The legitimacy of algorithmic decisions depends not only on their accuracy, but also on their transparency, fairness, and accountability. Ultimately, it is a question of trustworthiness.

To use algorithms wisely, we must recognize both what they can do and what they cannot, what is appropriate and what is excessive. In our pursuit of optimized outcomes, we are also influencing human experiences, perceptions of justice, and the distribution of power. At the time of the Council of Trent, a cardinal was questioned about the use of haruspicy. He replied, "I don't know if it works, but it is resorting to excessive means to know the truth."

We should remember that. The use of AI algorithms is a valuable way to improve the productivity of Tax Administrations, but it is not the best path to discovering the truth.

## 6 REFERENCES


Christian, B. (2021). The alignment problem: Machine learning and human values. Atlantic Books.

European Commission. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

González, I. M. (2024). The ethical need for a new type of tax norms in the world of artificial intelligence. *Review of International and European Economic Law, 3*(6), a2.1–a2.17. https://www.rieel.com/index.php/rieel/article/view/102

González, I. M., & Duart Crespo, S. (2024). AI, the unexpected attractor for tax and customs administrations: Taming the loop. *Review of International and European Economic Law, 3*(5). https://rieel.com/index.php/rieel/article/view/82

Klein, M. (1990). *Mathematical thought from ancient to modern times* (Vol. 1). Oxford University Press.

Piketty, T., & Sandel, M. (2025). *Equality: What it means and why it matters*. Polity.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science, 10*(1), 18–24. http://www.jstor.org/stable/184878

Ross, K. (2008). Brunelleschi's dome: The story of the great cathedral in Florence. Vintage Books.