taxnotes® international

Volume 95, Number 12 ■ September 16, 2019

Ruled by Algorithms: The Use of 'Black Box' Models in Tax Law

by Aleksandra Bal

Reprinted from *Tax Notes International*, September 16, 2019, p. 1159

Ruled by Algorithms: The Use of 'Black Box' Models in Tax Law

by Aleksandra Bal

Aleksandra Bal

Aleksandra Bal is a senior product manager, EU indirect tax reporting solutions, at Vertex Inc. Email: aleksandra.bal@vertexinc.com

The opinions in this article are those of the author and do not necessarily reflect the views of any organizations with which the author is affiliated.

In this article, the author evaluates the use of "black box" models as part of EU data protection and human rights legislation, focusing on a new Polish algorithm-supported system to detect VAT fraud.

The use of automated decision-making systems is on the rise. Algorithms already control, or at least affect, large parts of our lives. They make decisions about recruitment, credit scoring, and job promotion. In the foreseeable future, they will also be driving our cars. And we are fine with it as long as we more or less understand what the algorithms are doing. If their decisions depart significantly from our perception of what is right and proper,[1] we immediately become concerned about ceding control to artificial intelligence.

AI provides massive opportunities to do things better, more efficiently, and more cheaply for both tax administrations and taxpayers. Predictive analytics allows tax administrations to identify taxpayers that are most likely to be noncompliant. Administrators can allocate their resources more efficiently by focusing on high-risk cases, which leads to fewer and better-targeted audits. By using analytics, administrators can also deliver better-targeted services based on a deeper understanding of taxpayers' needs and circumstances. AI tools can help communicate differently across groups of taxpayers for maximum impact. They can also be used to explain tax consequences of some situations in simple language (tax chatbots). In the business sector, AI is commonly used to scan invoices to identify opportunities for VAT recovery or to detect anomalies in transaction data.

Several AI algorithms operate in a "black box" manner, meaning that it is difficult to understand how the system has arrived at a decision. A black box model will not explain itself or give the logic used to produce results. The increasing use of black box models has sparked a debate about algorithmic accountability and has led to calls for increased transparency in algorithmic decision-making, including in the forms of both explaining individual decisions and of audits that enable expert third-party oversight.

This article investigates the acceptability and legality of the use of black box models in tax law. Is it lawful to cede decision-making powers about taxpayers to those models? Should explicability and transparency be paramount in designing a model, or is accuracy the overriding consideration?

Basic AI Concepts

Almost every AI model in use relies heavily on machine learning — that is, the use of algorithms and statistical models to analyze data.

[1] Algorithmic decisions could produce discriminatory results. Because algorithms learn from observation data, if those data are biased, the algorithm will pick that up. For example, a recruiting tool developed by one large company tended to discriminate against women for technical jobs. The company's hiring tool used AI to score job candidates from 1 to 5. Because most resumes came from men, the system taught itself that male candidates were preferable. It penalized resumes that included the word "women's."

The most common outputs produced by machine learning algorithms are predictive models, which are constructed based on development samples that consist of observation and outcome data. The algorithms capture correlations between the observation and outcome data sets in the form of a model that can be used to predict events. In other words, they learn from data to respond intelligently to new data.

The most popular types of predictive models are linear models, decision trees, neural networks, and ensembles.[2] Linear models and decision trees are relatively easy to understand because they make predictions in a transparent way. They are "white box" in nature because it is easy to see which data contributed most to the outcome and which did not. Neural networks and ensembles tend to be more complex and black-box in nature. They generally deliver more accurate predictions, but it is difficult to understand why and how they produced a particular result, and the outcomes they generate are not intuitive.[3]

No type of predictive models can be said to be generally better than others. Data scientists frequently build various models and compare them to determine which is the most optimal for solving a particular problem.

Predictive models do not display deliberate bias, tend to be more accurate than humans, and are fast and cost-efficient because they can evaluate many cases in seconds. However, they can also get things wrong. A model's quality is only as good as that of the data used to construct the model. If the data is biased or incomplete, the model's results will be flawed as well. The same applies if the model developers make incorrect assumptions about how the model will operate. A common mistake in building a predictive model is to include every piece of available information in the machine learning process. Data that is outdated, unrepresentative of the target population, or unstable — that is, it will be unavailable when the model is applied — should be excluded. A model is built on past data but will be used in the future. Therefore, if some data will be unavailable, it should be excluded from the development sample. Also, models age. The relationships between data could change, and that could lead to a decrease in predictive accuracy. If the model monitoring shows that accuracy is decreasing, it is time for a new model to be developed.

There are two main types of machine learning: supervised and unsupervised. Supervised learning is machine learning that applies to development samples in which each observation has an associated outcome that one wants to predict. The algorithms learn how to map from input (observation data) to output (outcome data) using data with "correct" values already assigned to them. The initial phase of supervised learning creates a predictive model that will be used for making predictions.

If outcome data is unavailable, unsupervised learning can be applied. The goal of unsupervised learning is to discover interesting patterns in the data or identify groups of objects based on similarities between them. Unsupervised learning does not generate predictions. The most common type of unsupervised learning is clustering, or grouping similar cases together. This process can be used by tax administrations to identify outliers or unusual cases: Taxpayers are grouped into clusters, and if their return data deviate from that of their peers, they are flagged for further investigations.

STIR: An Algorithm to Detect VAT Fraud

VAT is one of Poland's biggest revenue sources, and the country was losing a lot of revenue because of VAT fraud. According to reports by the European Commission, the Polish VAT gap grew sharply between 2006 and 2011, rising from 0.4 percent of GDP to 1.5 percent. In 2012 it reached PLN 43.1 billion (approximately

[2] In linear models, the outcome is calculated by multiplying the value of each factor by its relevant weight and then summing up the results. Examples include logistic and linear regressions. A decision tree is created by recursively segmenting a population into smaller and smaller groups. Neural networks are a set of algorithms modeled loosely after the human brain and designed to recognize patterns in data. They use deep learning, which involves feeding a lot of data through multilayered neural networks that classify the data based on the outputs from each successive layer. Ensemble models are large collections of individual models, each of which has been developed using different data or algorithms. Each model makes predictions in a slightly different way; the ensemble combines the individual predictions to arrive at a final prediction.

[3] Google has developed AlphaGo, a computer system powered by deep learning, to play the board game Go. Although AlphaGo made several rational moves, its reasoning for others has been described as "inhuman" because no human could comprehend its rationale.

$11.2 billion). Because of the prevalence of VAT fraud and its impact on the country's financial stability, Poland started implementing a comprehensive plan to strengthen its VAT system. The plan included broadening the catalogue of goods and services subject to the reverse-charge mechanism, criminal sanctions, the introduction of split payments, and the standard audit file for tax (or SAF-T) reporting obligation.

In 2017 Poland adopted the System Teleinformatyczny Izby Rozliczeniowej (STIR), an innovative anti-fraud measure meant to reduce the VAT gap and detect carousel fraud.[4] STIR allows risk analysis and information exchange among the financial sector, the National Revenue Administration (NRA), and the Central Register of Tax Data.

Under STIR, banks and credit unions must report daily to the clearinghouse information on bank accounts and all transactions carried out by entrepreneurs (including the identities of parties to those transactions). The clearinghouse establishes a risk indicator for each entrepreneur, which is calculated using secret clearinghouse algorithms based on criteria used by the financial sector to combat tax fraud. Those criteria include customer residence, complex ownership structure, and unusual transactional circumstances. Taxpayers may not know how the risk indicators are determined.

The clearinghouse transmits the information from the banks and the risk indicator to the NRA daily. If the head of the NRA concludes that an entrepreneur is at high risk of being involved in VAT fraud, he may impose administrative measures, including blocking a bank account for up to 72 hours. In that period, the head of the NRA is expected to examine the case to determine whether there is a probability that it concerns tax fraud. The 72 hours can be extended up to three months if there is a justified suspicion that the entrepreneur will fail to settle any tax liability over €10,000. As long as the bank account is blocked, the entrepreneur cannot make bank transfers, and no funds can be withdrawn. The head of the NRA may authorize some payments to be made from a blocked bank account — for example, tax liabilities, maintenance payments, and employee remuneration.

Another administrative measure that may be imposed on entrepreneurs at risk of carrying out fraudulent activities is the refusal or cancellation of their VAT registrations. That measure is meant to protect honest taxpayers from entering into transactions with potential fraudsters by providing a publicly available list of persons whose registrations have been canceled or refused. The law does not specify what procedural rules apply to the decision to refuse or cancel a VAT registration, and the NRA is not obligated to inform an entrepreneur that her VAT registration has been canceled or refused.

STIR has proven a useful tool in the fight against VAT fraud because it allows the NRA to monitor bank accounts and transactions in nearly real time. The administration is immediately informed if a fraudster opens a new bank account to carry out a large transaction or transfer the funds abroad. In today's fast-paced business environment, speed is a key consideration in preventing carousel fraud. Before STIR was implemented, the tax administration was able to detect fraudulent carousel schemes only after two months of activity.

According to information from the Polish Finance Ministry, in 2018 STIR used information from 619 banks to monitor 11.6 million bank accounts of 3.4 million entrepreneurs.[5] Almost 30,000 entrepreneurs received a high-risk indicator. Only 23 had their bank accounts blocked, but all blockages were extended beyond the 72-hour period. The total amount of funds accumulated in the blocked accounts was PLN 10.3 million.

From the taxpayer's perspective, STIR is a black box model: The algorithms used to determine the risk indicator are not disclosed. A high-risk indicator plays a fundamental role in the NRA's risk assessment and decision to apply measures, such as blocking a bank account or canceling a VAT registration. Wrongly receiving a high-risk indicator could have disastrous consequences for entrepreneurs: A blocked bank

[4] Act of November 24, 2017, on Preventing the Use of the Financial Sector for VAT Fraud.

[5] Polish Finance Ministry, "2018 Annual Report on Preventing the Use of Banks and Credit Unions for VAT Fraud Purposes" (June 2019).

account could lead to insolvency and bankruptcy, and the cancellation of VAT registration and publication of that fact in a special register could seriously disrupt business activity. Further, the lack of clarity in the procedures for acting as a result of receiving a high-risk indicator makes it difficult for entrepreneurs to challenge the NRA's decisions.

Data Protection Legislation

General Characteristics

The most comprehensive data protection legislation ever enacted came into force on May 25, 2018, in the EU. The General Data Protection Regulation (GDPR) imposes numerous obligations on organizations regarding how they manage, collect, and process individuals' personal data — that is, any information regarding an identified or identifiable natural person.

Under the GDPR, the individual may determine who can collect his data and how it will be used. To store and process an individual's data without his permission is illegal.[6] The individual must give his consent, which he has the right to withdraw at any time and require the company to erase all his data.

An important characteristic of the GDPR is its extraterritorial application: It applies to all organizations processing personal data of individuals residing in the EU, regardless of their location. The regulation is binding on non-EU businesses that offer goods or services to, or monitor the activity of, EU individuals. Companies that are found in breach of the GDPR can be fined up to 4 percent of their annual global turnover or €20 million (whichever is greater).

Automated Decision-Making

The GDPR contains four articles that explicitly address algorithmic decision-making. Because those articles impose strict obligations on the developers of AI models, the media has suggested that the GDPR will slow the development and use of AI in Europe by holding developers to a standard that is often infeasible.[7]

Article 22 of the GDPR addresses automated individual decision-making, including profiling. It gives an individual the right to opt out of automatic processing:

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

That means that governments may not make decisions about people using an automated process unless people give their consent for automated decision-making to be used. A company applying automated decision-making tools must implement "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests," which must include "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision." In other words, individuals who are affected by decisions based on automated processing have the right to challenge that decision and have it reviewed by a human.

Right to Explanation

GDPR articles 13, 14, and 15 establish the right to explanation by requiring organizations handling personal data of EU citizens to explain how an automated decision was reached by providing meaningful information about the logic involved, as well as the consequences of that decision.

The European Commission has noted that "complexity is no excuse for failing to provide information."[8] The organization must mention "factors taken into account for the decision-making process" and "their respective 'weight' in

[6] Other circumstances in which personal data processing is lawful without an individual's consent include processing necessary for the performance of a contract to which the individual is party or to take steps at the request of the individual before entering into a contract, or processing necessary for the performance of a task carried out in the public interest.

[7] Nick Wallace, "EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence," TechZone360, Jan. 25, 2017.

[8] European Commission Article 29 Data Protection Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679" (Feb. 6, 2018).

an aggregate level." It has provided examples of information that should be given to individuals:

- the categories of data that have been or will be used in the profiling or decision-making process;
- why those categories are considered pertinent;
- how any profile used in the automated decision-making process is built, including any statistics used in the analysis;
- why that profile is relevant to the automated decision-making process; and
- how the profile is used for a decision concerning the individual.

The organization does not need to provide a complex mathematical explanation about how algorithms work or disclose the algorithm itself, but the information provided must be comprehensive enough for the individual to act on it to contest a decision, correct inaccuracies, or request erasure.

Evaluation

The GDPR creates a barrier to using black box models to make decisions about individuals if a suitable explanatory mechanism does not exist. Whereas white box models can explain themselves, it is often impractical, or even impossible, to explain decisions made by more complex machine learning algorithms. Ensemble methods or neural networks pose the biggest challenge because predictions result from an aggregation or averaging procedure. The requirements for explicability and manual intervention mandated by the GDPR can have a major impact on the costs of developing and maintaining automated decision-making systems. Those costs must be included in the cost-benefit analysis undertaken before the project begins.

From a GDPR perspective, STIR does not subject individuals to purely automated decisions. Although the risk indicator is determined by secret algorithms, it is reviewed by a person (the head of the NRA). The statistics (29,000 entrepreneurs with a high-risk indicator but only 23 accounts blocked) indicate that human review is not a mere formality.

However, it might be questionable whether STIR is in line with the GDPR right to explanation. An individual whose bank account is blocked for 72 hours has no opportunity to quickly contest that decision or to provide an explanation for a high-risk indicator. He is informed about the decision to block his bank account only after the measure has taken effect. On the one hand, an anti-fraud tool would be inefficient if fraudsters were informed about potential sanctions beforehand. And if the underlying logic of the algorithm became public, fraudsters could structure their activities to avoid detection. On the other hand, a person subject to sanctions is entitled to receive an explanation for those measures. To make STIR entirely GDPR-proof, affected entrepreneurs should be told why they are suspected of VAT fraud.

Fundamental Human Rights

In Europe, the legal framework for the protection of human rights consists of many sources, including the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR), which all EU members have signed. The guarantees of the ECHR lie behind many general principles of EU law, and its provisions were used as a basis for the EU Charter.

One of the fundamental ECHR guarantees is the right to a fair trial. It includes not only the right to be present, but also the right to hear and follow the proceedings. It applies throughout the entire process, from the investigation to the final decision. Although article 6 ECHR refers to "criminal charges," it can also be invoked in the context of taxation if a measure is imposed based on a legal rule with both deterrent and punitive purposes of pressuring taxpayers to comply with their obligations.[9] The right to a fair trial includes the minimum guarantees of equality of arms, right of defense, and presumption of innocence.

Equality of arms requires that the parties be given a reasonable opportunity to present their case under conditions that do not place them at a disadvantage vis-à-vis their opponent. The European Court of Human Rights has ruled that equality of arms might be breached when the

[9] European Court of Human Rights, "Guide on Article 6 of the European Convention on Human Rights, Right to a Fair Trial (Criminal Limb)" (Apr. 30, 2019).

accused has limited access to her case file or other documents.[10] In other words, unrestricted access to the case file is an important element of a fair trial.

The right of defense includes the right to be promptly informed of the nature and cause of the accusation. The accused must be provided with sufficient information to understand the extent of the charges against him[11] and be given an opportunity to challenge the authenticity of the evidence and oppose its use. The Court of Justice of the European Union has held that the observance of the right of defense is a general principle of EU law, and that it applies if tax authorities adopt a measure that will adversely affect an individual.

Viewed as a procedural guarantee, the presumption of innocence imposes requirements for the burden of proof and legal presumptions of fact and law. The prosecution must inform the accused that a case will be made against her so that she may prepare and present her defense.[12] The presumption of innocence is violated when the burden of proof is shifted from the prosecution to the defense.[13] The ECHR requires states to keep their legal presumptions of fact and law within reasonable limits and to strike a balance between the importance of what is at stake and the rights of the defense. In other words, the means must be reasonably proportionate to the legitimate ends sought.[14]

Equality of arms and the right of defense mean that taxpayers must be placed in a position in which they can effectively convey their views about information on which the authorities base their decisions.[15] A tax administration must give reasons for its decision, and the affected individual must have proper access to his case file; a decision made solely based on a black box model will likely conflict with those fundamental rights. If the taxpayer does not know how the decision was reached, there is no fair balance between the parties. He is hindered in his ability to provide evidence because he does not understand which objective factors the algorithm used in reaching the decision. Thus, the use of black box models may be questioned from the perspective of the right to a fair trial.

Because the Polish STIR allows the head of the NRA to impose punitive and deterrent measures and is targeted at preventing a criminal offense (VAT fraud), it falls within the scope of the right to a fair trial. Because the algorithms used to determine the risk score are kept secret, the addressee of those punitive measures does not know the objective facts that triggered the application of sanctions. The entrepreneur is not provided with sufficient information to understand the extent of the charges against him, which puts him at a disadvantage vis-à-vis the tax administration, thus creating imbalance. Moreover, the entrepreneur does not have the ability to challenge the 72-hour bank account blockage.

Although STIR significantly restricts the right to a fair trial, the European Court of Human Rights has held on numerous occasions that fundamental rights may be limited if it is strictly necessary to safeguard public interests, and the measures used are reasonably proportionate to the legitimate goals they seek to achieve. STIR pursues a legitimate objective in the public interest: It seeks to combat VAT fraud and prevent revenue losses. Disclosing the algorithms would reduce its effectiveness because fraudsters would structure their transactions to avoid detection.

However, it is questionable whether the system complies with the principle of proportionality. The nondisclosure of reasons for which the punitive measures were applied and the lack of opportunity to challenge the measures are serious limitations on the right to a fair trial. A less restrictive and more proportional solution would be to explain the decision to the taxpayer when sanctions are imposed and establish procedural rules to challenge the punitive measures. Even so, Polish law does ensure that STIR sanctions are imposed only when there is a strong suspicion of VAT fraud: They can be applied only by the head of the NRA, not by

[10] *Matyjek v. Poland*, 38184/03 (ECtHR 2007); *Moiseyev v. Russia*, 62936/00 (ECtHR 2008).

[11] *Mattoccia v. Italy*, 23969/94 (ECtHR 2000).

[12] *Barberà, Messegué and Jabardo v. Spain*, 10590/83 (ECtHR 1988); and *Janosevic v. Sweden*, 34619/97 (ECtHR 2003).

[13] *Telfner v. Austria*, 33501/96 (ECtHR 2001).

[14] *Janosevic v. Sweden*, 34619/97 (ECtHR 2003); *Falk v. the Netherlands*, 66273/01 (ECtHR 2004).

[15] *WebMindLicenses Kft. v. Hungary*, C-419/14 (CJEU 2015).

ordinary tax inspectors. That could be interpreted as limiting punitive measures to what is strictly necessary for effective tax collection.

Conclusions

AI-powered algorithms can be used to make cheaper, faster, and more accurate decisions than those made by humans. However, just like humans, algorithms can make mistakes or be biased. Therefore, appropriate checks and balances are necessary to prevent misuse of decision-making systems that rely on machine learning.

In developing new models for tax administration, accuracy should not be the overriding consideration. Having an explicable model is far more important than having one that is slightly more accurate but much less understood by regulators and business users. When building a model, the transparency and explicability of the resulting solution should be considered in light of the applicable legal framework. Black box models that produce very accurate but inexplicable outcomes might not be preferable because they could conflict with legislation protecting personal data or fundamental human rights.

The GDPR has established the right to explanation when automated decision-making systems are used, and the ECHR requires an individual to be promptly provided sufficient information on the nature and cause of penalizing measures. Both legal frameworks have important legal implications for the design and deployment of automated data processing systems. It can be predicted that algorithmic auditing and transparency will become key considerations for enterprises deploying machine learning systems both inside and outside the EU.

To achieve a proper level of transparency in algorithmic decision-making, it should be ensured that any decisions produced by an automated system can be explained to the people they affect. Those explanations must be understandable by the target audience. Also, it should be clear who has the authority to review and potentially reverse algorithmic decisions. Finally, algorithms should be monitored and regularly checked to ensure they are up-to-date and socially relevant. ■