The Use of Explainable AI(XAI) Methodologies for Transfer Pricing Documentation

Abeer Kayser

Abstract

Black-box artificial intelligence technologies have been readily integrated into transfer pricing adjustments and enforcement, leading to an incompatibility between the resulting efficiency and the requirements for legal admissibility. US Treas. Reg. § 1.482 requires a human-readable functional analysis as a legal mandate for "reliability"; however, black-box models inherently fail to meet this requirement. This paper evaluates the *Daubert* standard for expert testimony and the Best Method Rule in accordance with US regulatory requirements to identify interpretability gaps that currently leave AI-driven transfer pricing adjustments vulnerable to legal inadmissibility. This supports the proposal of a standard for a transfer pricing algorithm, which charts qualitative legal criteria to quantitative XAI methodologies to meet a criterion for explainability. A Standardized Disclosure Protocol for AI-driven adjustments would apply the game-theoretic approach of SHAP(Shapley Additive exPlanations) to legal criteria, particularly "auditability", and convert black-box outputs into standardized contributions, meeting the "reasoned basis" required by the IRS. In doing so, it establishes the admissibility of algorithmic transfer pricing contingent upon the ability to add a layer of human-readable explainability. The paper concludes its investigation by evaluating standardization as a viable compliance instrument in reconciling human legal oversight and AI-driven transfer pricing.

Introduction

Transfer pricing encapsulates the accounting practices that enable related entities housed within a company or parent organization to conduct transactions with each other. In particular, transfer pricing concerns the prices set in these inter-company transactions to ensure that they reflect those of external-market dealings. To note, this practice does not just pertain to physical goods; rather, it applies to any inter-company transactions, including but not limited to the transfer or exchange of intellectual property and loans. While the principle is to ensure that prices are reflective of market-based pricing, in terms of accounting treatment, these are internal ledger entries and influence the profitability recorded by one subsidiary in conjunction with another. Multinational enterprises (MNEs) seek to use transfer pricing to minimize their global tax liability by shifting revenues to subsidiaries in low-tax jurisdictions, or tax havens, while inflating expenses in higher-tax jurisdictions. Ireland, for instance, is often considered a favorable jurisdiction for various MNEs operating in the US due to its taxation and economic policies. As a motivating example for this paper, consider a scenario where a US-based parent company transfers the rights of its proprietary software algorithm to its Irish subsidiary. The Irish subsidiary then licenses this intellectual property back to the US parent, charging a significant royalty fee. This is an internal ledger entry, and the US parent company records an expense, subsequently reducing its taxable income in the US, while the Irish subsidiary records the income which is taxed at Ireland's lower corporate rate.

Nonetheless, transfer pricing practices are heavily regulated in an evolving landscape due to their macroeconomic impact. Base Erosion and Profits Shifting (BEPS) specifically refers to a direct consequence of this practice in eroding the tax bases of sovereign nations, leading to

billions of dollars in lost tax revenue. Globally this costs countries between $100 billion and $240 billion annually, representing 4% to 10% of global corporate income tax. Thus, via international tax audits, stringent and substantial documentation requirements for transfer pricing reports, tax authorities seek to protect national revenues. There are a myriad of transfer pricing methods in use; some of the most common ones include the Comparable Uncontrolled Price (CUP) method, which compares prices charged in a controlled (related) transaction to a similar uncontrolled (independent) transaction, the Cost Plus Method, which adds an appropriate gross profit mark-up to supplier's costs, and the resale price method, which subtracts an appropriate gross margin from the resale price to an independent party.

Regulatory Baseline

The Arm's Length Principle (ALP) is the internationally accepted standard dictating regulatory practices surrounding transfer pricing. At its core, it requires intercompany transactions to be priced as if they were being conducted between two independent entities negotiating in an open, competitive market. The IRS provides guidelines for companies operating in the US for implementing and remaining in compliance with the ALP. 26 CFR 1.482-0 provides the authority that the IRS uses to enforce this. A precursor to compliance with this principle is proving "comparability," which requires large-scale scraping of databases to find transactions between independent entities that are similarly priced. This is one area where AI has substantial potential in revolutionizing transfer pricing practices. In automating the analysis of vast datasets of comparables, however, the arising complexities give way to the "black-box" issue.

To prove compliance with the Arm's Length Principle for the software transfer in our motivating example, the US parent must scrape massive economic databases to find independent tech companies licensing similarly complex algorithms. Finding a comparable uncontrolled transaction for a unique piece of intellectual property is characteristic of the type of historic bottleneck that drives MNEs toward algorithmic automation.

Use of AI in Transfer Pricing

AI is used across various realms of transfer pricing, including but not limited to three main functions. Benchmarking involves using machine learning to analyze complex datasets to instantaneously rank thousands of potential comparables, reducing the possibility of manual error and increasing efficiency. Natural Language Processing (NLP) capabilities can be used to automate document creation through translating raw financial data into reports to ensure compliance across the varying standards in different jurisdictions. AI can also provide continuous, real-time monitoring of intercompany transactions to allow for the transition from retroactive to proactive pricing adjustments. However, as AI transforms transfer pricing with the scaling of these tools, how can we ensure human-mandated legal compliance? Is a standardized protocol the answer?

Instead of relying on manual searches, the US parent company, in our illustrative example, deploys AI to instantly analyze global tech market datasets. The algorithm benchmarks the proprietary software against thousands of potential comparables and continuously monitors the intercompany IP transactions in real-time to proactively adjust the royalty rate in response to market volatility.

Use of "Black-Box" Models and Why They Fail

The adoption of certain machine learning models within tax law and tax administration currently remains suppressed due to the standards of accountability required for legal admissibility. The lack of accountability is not a feature of all machine learning models, however, and it is imperative to understand the distinction. Predictive models, the most common outputs of machine learning algorithms, depend on the use of historical data, including observation and outcome datasets, to be able to make predictions and respond to new data. Predictive models can be broken down further into their "white-box" and "black-box" subcomponents depending on the variability of transparency.

Decision trees and linear models offer predictions transparently, as they would enable you to see what data contributed most to a certain outcome. These "white-box" predictive models are easier for humans to understand and offer transparency at the cost of some level of accuracy. Neural networks and ensembles are more accurate, complex, "black-box" types of predictive models. The tradeoff is the understandability, or lack thereof, that these types offer. The outcomes generated are not intuitive, and it is difficult to understand how the model transformed the inputs into the outputs or outcomes, that is, how or why the model produced a specific result. Generally, transfer pricing methods and the determination of a transfer price that is in compliance with the ALP are dependent upon economists and tax specialists to scrape large, complex datasets. This process, when done manually, runs a higher risk of errors, lacks time efficiency, and can be reliant on historical data, making it a retroactive process. In using AI, these issues are largely mitigated.

The architecture behind black-box models explains why a company cannot explain the model or the path it took to produce the results. Black-box models operate in high dimensionality with their capabilities eclipsing human cognitive capacity. These models evaluate tens of thousands of variables simultaneously, and as the data moves from the input to the output, or the final transfer price, it passes through multiple "hidden layers." The algorithm applies non-linear transformations to find subtle correlations. The resulting transfer price is optimal in accuracy, but it is not possible for human cognitive capacity to mathematically reverse engineer the path the model took to produce the output from the original inputs.

In designating black-box model outputs or results to be more "accurate," it is due to their mathematical optimization of the transfer price. This is due to their capability for high dimensionality, that is, their capacity to evaluate thousands of variables concurrently in practices that currently evade cognitive capacity. The reasoning behind the production of the transfer price is not contained to a single rule, rather it is distributed across millions of mathematical weights wherein the crux of the issue lies. Since they do not solely apply linear transformations, the complexity of mathematical functions warps the data and results in obscurity in the path from input to output.

Returning to our motivating example, the US parent company's neural network evaluates thousands of variables simultaneously, assessing the software's code complexity, Irish market volatility, and global tech trends all at once. It optimizes an arm's length royalty rate of 13%. While this is the most mathematically accurate price, the US parent company cannot explain the algorithm's internal, hidden logic to an auditor.

It is vital to recognize that tax compliance is not contingent solely on accuracy and requires traceability, or some form of verifiable audit trail. While black-box machine learning models may produce the most accurate transfer price, they critically fail at meeting the requirements for legal admissibility as they are unable to explain the internal logic as to why they made the decisions they did to reach a certain outcome. To elaborate on why "black-box" models fail in conducting transfer pricing that would be legally admissible, that is, meet the requirements set by regulatory authorities and hold up in court, it is critical to examine the legal framework surrounding transfer pricing practice.

Legal Framework

Best Method Rule

Treas. Reg. § 1.482-1(c) or the Best Method Rule states that the arm's length result of a controlled transaction must be determined under the method that, under the facts and circumstances, provides the most reliable measure of an arm's length result. Since the Best Method Rule does not explicitly give preference to one method over another, except that it most reliably produces the arm's length result, it inherently does not restrict the use of machine learning algorithms or delineate the use of automated models. When determining which method provides the most reliable measure, however, the two primary factors set forth by the Treasury regulation are the degree of comparability between the controlled transaction and any uncontrolled variables and the quality of the data and assumptions used in the analysis. The quality of the data and assumptions is dependent on three underlying factors: the completeness and accuracy of data, the reliability of assumptions, and the sensitivity of results to deficiencies in data and assumptions. The legal admissibility of black-box models is called into question here,

that is, without a method to determine or explain, in human-readable logic, the hidden weights employed by black-box models in transfer pricing analysis, the results produced would not be viable under the question of whether it meets the specific criteria for the quality and reliability of data and underlying assumptions.

Penalty Protection: Treas. Reg § 1.6662-6

Treas. Reg § 1.6662-6 holds that taxpayers must maintain "contemporaneous documentation" to establish that they reasonably concluded that their chosen method was the most reliable to be in accordance with the price being an arm's length result. To prove this, it becomes essential that a human-readable functional analysis detailing the choice of comparables is producible. Since the black-box model output lacks the narrative required by the IRS to explain the economic rationale behind the selection of comparables, it would not be legally admissible in US Tax Court. An interpretable audit trail that explains why the model weighted certain variables is required for the taxpayer to be able to prove the "reasonableness" of their method. Thus, this leaves the taxpayer fully exposed to penalties for substantial or gross valuation misstatements. Explainable AI bridges this gap by translating the algorithm's mathematical weights into the functional analysis, providing necessary documentation.

Daubert Standard, Federal Rule of Evidence 702

Federal Rule of Evidence 702 mandates that expert testimony must be based on reliable principles and methods that are applied to the facts of the case. As established by the Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993)*, judges assess factors like empirical testability, known error rates, general acceptance within the relevant scientific or economic community, and ability for cross-examination by the relevant experts. While

reproducibility of the outcome or result is possible with black-box models, the reproducibility is legally void if the underlying methodology cannot be examined. As outlined in the discussion above, black-box models do not produce an examinable internal logic, thus the underlying methodology and consequently all assumptions cannot be audited by economic experts, rendering these models to inherently fail the *Daubert* standard.

During an IRS audit of the software transfer in our illustrative example, the raw algorithmic output of 13% is a conclusion without a premise. It lacks the functional analysis mandated by Treas. Reg. § 1.6662-6 outlining exactly why that rate was chosen. Furthermore, because opposing IRS economists cannot audit the hidden network to understand how the royalty was calculated, the model inherently fails the Daubert standard for cross-examination in US Tax Court.

SHapley Additive exPlanations

Post-hoc explainability must be achieved via a translation layer added after the model's use. Created by Scott Lundberg and Su-In Lee (2017), at the University of Washington, SHAP(SHapley Additive exPlanations) explains the output of any machine learning model by employing a game theoretic approach. It must be noted that this approach can reverse engineer the output of any AI model, regardless of the underlying architecture, on the basis of how much various factors, or inputs, contribute to that result.

To understand SHAP, it is necessary to comprehend the Shapley Value, which is the cooperative game theory concept that provides its foundation. Developed by Lloyd Shapley in 1953, the Shapley Value framework addresses the question of individual contribution within a

collective outcome: when a coalition of players collaborates to produce an outcome, what is the specific contribution of each player? By evaluating the total earnings of a group and contrasting it with the payout earned in the absence of a particular player, that player's marginal contribution is isolated. Through calculating these differentials across all possible groupings of players, the Shapley Value yields a mathematically grounded measure of each player's individual worth. A direct parallel is thus established with transfer pricing; just as the Shapley Value delineates a player's contribution to aggregate earnings, SHAP isolates the specific weight of each input variable in the determination of a model's final transfer price. In doing so, it effectively translates cooperative game theory logic into a post-hoc explainability framework for machine learning.

In this framework, input features are treated as players, and the model prediction is the payout. These inputs are not arbitrarily chosen and must be provided directly by the firm whose model is being evaluated. SHAP calculates the average marginal contribution of each player across all feature coalitions. By converting non-linear algorithmic logic into quantifiable, additive feature contributions, SHAP successfully reverse-engineers a black-box model into a linear, human-readable explanation. SHAP does this via simulating the model's output both with and without specific variables across all feature combinations and subsequently isolating the exact value that each input adds to the final calculation. By applying SHAP to the software transfer in our motivating example, the 13% royalty rate becomes the "payout." SHAP calculates the marginal contribution of each variable, successfully reverse-engineering the black-box into a linear explanation: e.g., base royalty is 5%, software development hours contributed +5%, and Irish market size contributed +3%, resulting in the final 13%.

The integrality of the SHAP approach lies in its role in tax regulation. Regardless of the complexities of the black-box model, the SHAP approach evaluates the proportions of the inputs that contribute to the output. This means that for firms seeking a loophole through claiming that their models may be too sophisticated to evaluate, it can still be analyzed objectively, mitigating the occurrence of weaponized ignorance for corporate gain. SHAP provides a regulatory process that verifies the model's validity irrespective of whether a firm does not understand its pricing model or is employing ignorance to utilize more favorable pricing. This does not digress significantly from the current transfer pricing standards; the SHAP approach is simply the adoption of an approach that allows for technological change and advancement in international tax compliance by providing a realistic methodology for methods that may be too complicated for human cognition and understandability.

Nonetheless, a limitation exists here because firms supply their own input features for SHAP analysis. It follows that a firm could strategically choose features that produce favorable, stable attributions while leaving out variables that actually drove the pricing decision. This type of manipulation cannot be resolved by the SDP's internal metrics. The appropriate fix is external and regulatory and requires mandating feature categories for common transaction types, similar to how Treas. Reg. § 1.482-1 already specifies comparability factors, thus, making omitting economically material variables a documentable violation rather than a silent distortion that constitutes a legal loophole.

Development of a Standardized Disclosure Protocol

To operationalize legal admissibility, qualitative legal requirements must be mapped directly to technical metrics through a SHAP Stability Framework. To satisfy Treas. Reg. §1.6662-6, the mandated functional analysis of why comparables were selected must be provided, thus, SHAP Local Explainability is utilized. The variance of SHAP attributions across model runs must stay below a calibrated threshold, mathematically represented as $\sigma(\varphi) \leq \tau$. This is to ensure consistency by certifying that the algorithmic explanation does not arbitrarily change upon revision or re-evaluation. Furthermore, a Kendall rank correlation is employed specifically because it measures ordinal association. It is used to calculate the ratio of concordant pairs (rankings of variables where the rank order remains the same) to discordant pairs (the model flips the rank order of variables) across multiple runs of the model. The proposed Standardized Disclosure Protocol would mandate a Kendall rank correlation of $\tau \geq 0.9$ on a scale of -1 to 1. A threshold of $\tau \geq 0.9$ implies that at least 95% of pairwise comparable rankings are concordant across model runs, a standard consistent with high-reliability benchmarks in applied statistics for legal and medical decision-making contexts, where ordinal consistency is treated as a prerequisite for admissible expert judgment (Fleiss et al., 2003). In the legal context, this also meets the IRS requirement for a "reasoned basis," as an explanation that shifts upon re-evaluation cannot constitute a reliable functional analysis because it would produce different economic narratives for the same transaction depending on when or how the model was run.

To satisfy the *Daubert* standard (FRE 702), empirical testability and auditability must be ensured. SHAP-ranked comparables must maintain Ranking Comparability, represented as $\Delta R \leq \delta$. This is to ensure that the pairwise inversion rate, that is, the frequency with which the model

incorrectly flips the rank order of two given market comparables, is minimized. If comparable A functionally dominates comparable B based on the underlying economic data, A must reliably rank higher according to the algorithm as well. If the model is simply producing random, untestable results, the resultant inversion rate will be high, warranting a failure in meeting *Daubert*.

These conditions are conjunctive and must be met concurrently. For instance, if a model is consistently stable in ranking market comparables in the correct order but produces erratic SHAP values regarding why it ranked them, failing the Kendall rank correlation of $\tau \geq 0.9$, it would fail the Treasury's functional analysis requirements. In contrast, if a model produces stable SHAP values but produces biased comparable rankings by routinely ranking weaker comparables higher, it fails the *Daubert* standard for testable reliability.

To govern this, an AI-Specific Standardized Disclosure Protocol (SDP) is proposed. While the OECD currently utilizes an SDP, it does not directly address or regulate AI use. This proposed SDP serves as a mandatory "Cover Sheet" for AI-driven adjustments, incorporating three critical components:

- Model Card: Documents version control, the model's overarching architecture, and the exact training data sources utilized.
- Global Feature Importance: Outlines the overall priorities and weightings of the model across the aggregate data.
- Local Explainability: Utilizes SHAP to map exact contributions and variables for a single, controlled transaction to prove its specific arm's length validity.

Returning to our illustrative example, the US parent company utilizes the SDP as the mandatory cover sheet for its tax return. It submits a Model Card detailing the tech data sources, and utilizes SHAP to map the specific variable contributions directly to the single Irish software transaction. By guaranteeing conjunctive compliance with the aforementioned conditions, the firm proves the model's stability, reliability, and testability, which formally bridges the interpretability gap and allows the IRS to cross-examine the methodology.

Conclusion and Implications

The Standardized Disclosure Protocol proposed in this paper represents a concrete and actionable path toward reconciling algorithmic efficiency with the human-readable accountability demanded by US tax law. This provides a definitive path forward. However, it is critical to note that the use of XAI and standardization does not eliminate the motivations of the Base Erosion and Profit Shifting or neutralize tax avoidance. Base Erosion and Profit Shifting is fundamentally an incentive problem, and as long as meaningful corporate tax differentials exist between jurisdictions, MNEs will seek to exploit them.

Finally, to ensure that it fulfills its purpose, the SDP must be a living document that adapts alongside algorithmic advancements to ensure that black-box adjustments are appropriately subject to human oversight and auditability. The OECD's existing transfer pricing guidelines have historically lagged behind market practice, oftentimes by years, leading to enforcement gaps that sophisticated MNEs have exploited. A governance mechanism should be established to ensure the SDP's technical thresholds and methodological requirements are

updated in step with advances in both AI capabilities and XAI research. Without this, standardization risks becoming obsolete precisely when it is needed most.

References

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint*. https://arxiv.org/abs/1806.08049

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review, 71*(1), 1–56.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley.

Internal Revenue Service. (2024). *Allocation of income and deductions among taxpayers* (26 C.F.R. § 1.482-1). https://www.law.cornell.edu/cfr/text/26/1.482-1

Internal Revenue Service. (2024). *Penalties for transfer pricing misstatements* (26 C.F.R. § 1.6662-6). https://www.law.cornell.edu/cfr/text/26/1.6662-6

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1705.07874

OECD. (2022). *OECD transfer pricing guidelines for multinational enterprises and tax administrations*. OECD Publishing. https://doi.org/10.1787/0e655865-en

Federal Rules of Evidence, Rule 702, 28 U.S.C. (2023). https://www.law.cornell.edu/rules/fre/rule_702