TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

MACHINE LEARNING AND TAX ENFORCEMENT

Janet Holtzblatt and Alex Engler

June 22, 2022

"As we strive to operate more efficiently, provide superior service to taxpayers and their representatives and ensure successful implementation of changes in tax laws, we're embracing and integrating data into our culture. Using analytics, we can continuously improve all facets of our operations—taxpayer service, enforcement efforts and a range of internal operations—maximizing our learning from tests and data.…The IRS must respond to other changes (e.g., process robotics, blockchain and artificial intelligence) and integrate technologies that enable more efficient mission delivery." (IRS, 2018b)

Modernization of its technological infrastructure is a key component of the Internal Revenue Service's (IRS's) long-term strategy for improving the agency's efficiency. To achieve that goal, the Biden administration proposed that a portion of its request for a 55 percent boost (after adjusting for inflation) to the IRS budget over the next decade be used for developing machine learning. With machine learning, the IRS would seek to leverage existing information to better identify tax returns for compliance review.

Using machine learning for tax enforcement is not a revolutionary idea. Among countries belonging to the Organisation for Economic Co-operation and Development (OECD, 2021), 38 percent used artificial intelligence or machine learning in tax administration in 2019, while another 34 percent were in the development stage. Countries used those technologies for both taxpayer services and enforcement. The IRS has already taken steps in that direction, most notably with the implementation in 2017 of the Return Review Program (RRP), which identifies questionable refunds on individual income tax returns using a combination of conventional approaches and machine learning.

If successful, machine learning would marshal the vast trove of data currently received by the IRS to achieve more targeted and productive enforcement actions. Still, the application of machine learning to tax enforcement faces hurdles, some of which are inherent to the methodology and others specific to the US tax system, including the complexity of the tax code, the effects of past budget cuts, and the uncertainty of future funding.

THE IRS'S VAST TROVE OF DATA

Each year, the IRS receives billions of W-2s and 1099s from employers, financial institutions, government agencies, casinos, and many other entities. Those information forms show payments to workers, clients, Social Security beneficiaries, gamblers, and other individuals. Third parties also send the IRS 1098s, reporting taxpayers' payments for items claimed for some deductions and credits, such as home mortgage interest payments to banks (for the itemized deduction) and tuition payments to universities (for the education tax credits). In addition, the IRS collects data from the Social Security Administration and other government agencies about certain characteristics of taxpayers useful for verifying their claims of child-related tax credits, such as the age and residency of their children and their authorization to work in the United States.

Thus, the IRS has ready access to big data. And because that access is generally required by legislation, the IRS does not pay to obtain most information from third parties. Nonetheless, much of those data are underused. For example, third-party information returns are matched to individual income tax returns in late summer or fall as part of the Automated Underreporter Program (AUP). In 2018, the IRS received 2.8 billion information returns (IRS, 2019b) and detected 22.3 million discrepancies in the AUP but selected only 2.9 million for further review (GAO, 2021).

Though the data are mostly free to the IRS, using it is not. Despite its name, the AUP is not fully automated. Discrepancies discovered in the AUP must be reviewed by IRS employees before taxpayers are notified to distinguish, for example, discrepancies caused by the taxpayer reporting an interest payment but on the wrong line of the tax return from those resulting from the taxpayer's failure (intentional or not) to report that income entirely. And although 99 percent of information returns for tax year 2018 were submitted electronically, that still left 34 million (GAO, 2021) that needed to be entered manually into the IRS master files, because the agency lacks legislative authority to require that all documents be filed electronically.[1]

Deep cuts to the IRS have likely contributed to that low follow-up rate. Since 2010, the IRS budget has been cut by more than 20 percent, after adjusting for inflation. Congress imposed an agency-wide hiring freeze from 2011 through 2018, a time when many experienced staff were approaching retirement age. Over that period, AUP staffing declined by 42 percent (Treasury Inspector General for Tax Administration, 2019). Another contributing factor is the even longer-term underinvestment in technology; some IRS operations still rely on programming languages (such as COBOL [IRS, 2019c]) developed more than half a century ago and unfamiliar to most programmers today.

Still, even with additional resources, the IRS would find it challenging to apply AUP to all information returns. First, third-party information does not always align precisely with a line on a tax return. Contrast the W-2 sent to employees with the Form 1099-K received by many independent contractors, among other taxpayers.[2] Both forms are intended to track the taxpayer's earnings, but only the amount reported on the W-2 can be transcribed exactly on to the line for wages and salaries on the employee's tax return. The 1099-K shows payments received by independent contractors but not

[1] An alternative is to place bar codes on paper returns, allowing the IRS to obtain the information by scanning the return. But even that simple fix cannot be implemented to all returns without legislative authority. Without legislation, the IRS could add the bar code to returns downloaded from the IRS website or obtained from other sources. However, the IRS lawyers have concluded that the agency lacks the authority, under current law, to mandate that tax-software developers include bar codes on returns, forms, and schedules created through their software (Shawn M. Baker [Services and Enforcement Program Management Office, IRS], memorandum to Elizabeth Girafalco Chirich [branch chief, Procedure & Administration, IRS] regarding "Authority to Mandate Tax-Software Developers to Embed Two-Dimensional Barcodes on Returns and Forms or to Incorporate Two-Dimensional Barcodes on Returns and Forms that the IRS Designs and Issues," December 17, 2021, https://www.irs.gov/pub/lanoa/pmta-2022-02.pdf)

[2] The 1099-K requirement, which was enacted by Congress in 2008, took effect in 2011 and was amended again in 2021. The provision requires merchant acquirers (i.e., financial institutions that facilitate credit-card and debit-card transactions on networks such as Visa and Mastercard) and third-party settlement organizations (e.g., PayPal and Venmo) to file information returns reporting the gross amount of payments to each payee.

their expenses (gas purchases by ride-sharing drivers, for example). Researchers have shown that some taxpayers overstate losses on their tax returns—thus understating their taxable income (Slemrod et al, 2017; Adhikari et al, 2020).

Second, the complexity of the tax code—combined with the complexity of a business's structure—may hinder matching of third-party information and tax returns. Currently, partnerships send Schedule K-1s to the IRS and each partner showing their share of the firm's taxable income. Partners are then responsible for reporting their share on their own tax returns and paying the individual income taxes owed on that amount. But partnership law is complex, as signified by the 23 lines for the partner's share of income, deductions, credits, and other items; 10 lines for information on the partner, including the partner's role in the business and share of assets and liabilities; and 21 pages of instructions. And sometimes it is difficult to identify all the partners: for example, when another partnership is one of the owners. As a consequence, the IRS has not routinely linked all Schedule K-1s to partners' tax returns.

HOW DOES THE IRS CURRENTLY SELECT RETURNS FOR COMPLIANCE REVIEW?

The AUP, described above, is not an audit, though it may lead to one if the taxpayer disputes the IRS's computations. Apparent discrepancies between the tax returns and related information returns can lead to notices from the IRS and ultimately a bill for unpaid taxes, plus interest and penalties, if the taxpayer does not respond to the earlier correspondence. The AUP thus provides a relatively straightforward—if underused—approach to cases where third-party information aligns closely to the amounts reported on tax returns.

The IRS uses various means to select returns for the more extensive audits, and the approaches largely depend on the availability of relevant information, the complexity of the tax return, and the IRS's resources. About 70 percent of audits are not conducted in the IRS's offices or in person (IRS, 2020c). Instead, those audits are conducted through correspondence: The IRS sends a letter to the taxpayer asking for documentation to support an amount reported on the tax return, and the taxpayer can respond by mailing supporting documents back to the IRS. Correspondence audits are often spurred by issues that arise in a review of third-party information that are informative but less specific to a line on a tax return than the data used in the AUP process.

For example, to pick returns with questionable claims of child-related benefits, the IRS relies on the Dependent Database, which contains some information on parental relationships and children's living arrangements from the Social Security Administration and the Department of Health and Human Services as well as prior and current-year tax records. Those data are used to construct business rules (with "yes" or "no" responses) to determine the likelihood of an invalid claim. Depending on the weight given to each response, the correspondence audit may be triggered by one or more disputed issues—say, whether a child claimed for the earned income tax credit lived with the taxpayer for more than six months, as required, or was claimed by the most eligible member of the household—rather than an analysis of the overall return.

Since 2014, the IRS's Small Business and Self-Employed Division has been responsible for all postrefund audits of individuals (other than for the earned income tax credit audits, which are handled by the Wage and Investment Division). The Small Business and Self-Employed Division generally focuses on the self-employed and businesses with less than $10 million in assets. Notably, those are taxpayers with income for which third-party information is currently scarce. Nearly half of the returns selected for audit by the Small Business and Self-Employed Division in 2018 were based on their discriminant function (DIF) score (Treasury Inspector General for Tax Administration, 2020). That scoring methodology is derived from the findings of the National Research Program (the IRS's compliance studies, which randomly select individual income tax returns for audit), and the scores' ranking is based on the overall return rather than on specific issues.

As with the AUP, the number of audits has fallen as the IRS budget has been cut and the number of departing experienced staff has risen. In total, the number of audits fell from 1.7 million in 2010 to 771,000 in 2019 (IRS, 2011; IRS,

2020c). The audit rate, defined as the ratio of the number of audits closed in a fiscal year to the number of tax returns filed in the prior tax year, fell from 0.9 percent in 2010 to 0.4 percent in 2019. But the decline was more marked for certain groups of taxpayers: Among earned income tax credit claimants, the audit rate was 2.4 percent in 2010 and had dropped 1.3 percentage points, to 1.1 percent, by 2019. For individual taxpayers with positive income of $1 million or more, the audit rate fell from 8.4 percent in 2010 to 2.4 percent in 2019.

WHAT IS MACHINE LEARNING AND WHAT IS ITS POTENTIAL USE IN TAX ENFORCEMENT?

Machine learning is a group of statistical tools, of which two subcategories are particularly relevant for tax enforcement: supervised machine learning (sometime called "predictive analytics") and unsupervised learning. Supervised learning is focused on a single known outcome, whereas unsupervised learning discovers patterns across datasets that might otherwise be missed. In either case, human involvement is still required. Often in consultation with subject matter experts, data scientists and engineers need to apply or fine-tune the machine algorithms that fit the problems they wish to solve, to examine the findings made by computers, and to decide how to act on those findings.

When supervised learning is applied to tax enforcement, one prominent use is determining the likelihood of noncompliance. A first test is identifying ("labeling") examples in the data; in this case, labeling tax returns which have been found by IRS employees to be noncompliant. The computer learns relationships between that labeled example and other variables (e.g., taxpayer characteristics and information from many tax forms). For instance, it might learn that indicators of noncompliance include reporting unusually low income, claiming high losses or excessive expenses, and failing to report income reported by third parties. Using those learned patterns, supervised machine learning can then predict the likelihood of an unlabeled tax return being filed by a noncompliant taxpayer. The resulting model would be used in subsequent tax years to select returns for audits or other compliance activities. The system would evolve, incorporating new information from completed enforcement actions. Common methods used for supervised machine learning include linear or logistic regressions, support vector machines, Bayesian classifiers, decision trees, and neural networks.

Unsupervised learning has no single outcome or variable of interest; instead, unsupervised learning searches for previously undetected patterns across a dataset. One type of unsupervised learning is clustering, where a model learns to find similar groups within the dataset. Tax administrators could use clustering to create groups of similar tax returns based on the many variables contained in the dataset, enabling better categorization than a person might have discovered on their own. Clustering methods include k-means clustering and hierarchical clustering.

Another type of unsupervised learning that could facilitate tax enforcement is anomaly detection, which aims to detect novel or unusual data. Anomaly-detection methods include one-class support vector machines and isolation forests. Those methods can be used in combination with clustering, wherein outliers, or returns that deviate greatly from the rest in the cluster, may indicate noncompliance.

Both supervised and unsupervised machine learning can be superior to the conventional approaches to selecting returns for a compliance activity, such as an audit. Under either approach, computers can absorb and analyze more data than the conventional approaches. By searching for relationships and patterns among variables, precise alignment between a specific data point and a line on a tax return is not required. Machine learning may make better use of data, finding more subtle patterns and complex interactions between variables that can indicate noncompliance.

But some of the challenges associated with the AUP and audits would also apply to machine learning. In particular, the quality of the outputs (audit selection) will depend on the quality of the inputs (data). The time required to perform an audit is another important consideration. Because of the lags in initiating, completing, and analyzing audits, the DIF scores are derived from six-year-old tax returns, which could reflect a different tax regime. The Dependent Database is

based on more recent data, but its focus on family-related tax benefits is narrow and the quality of the business rules is reviewed by IRS staff only once a year, months after the end of the filing season (GAO, 2016).

Supervised and unsupervised machine learning differ in the timeliness of the tax data used in the analysis. Supervised machine learning generally must be updated with complete historical data, meaning the model needs the outcome of audits to learn from data. This means that supervised machine learning may face similar challenges as the DIF scores in keeping up with current practices. Notably, this is not true of unsupervised machine learning, which may make it more helpful for keeping up with changes in the tax code and evolving tax-evasion strategies.

RETURN REVIEW PROGRAM

That concern about the lack of flexibility of rules-based systems is addressed in a relatively new prerefund enforcement tool, the Return Review Program (RRP [GAO, 2018]), which is used to detect identity theft and other types of noncompliance during returns processing.[3] By using this tool when tax returns are filed and processed, the IRS can freeze refunds to potentially noncompliant tax filers for further action or review.[4]

Though the RRP has an extensive rules-based component, it also relies on, according to the IRS, "leading edge machine learning technologies" to identify evasion strategies that emerge during the filing season (IRS, 2021b). The RRP uses both supervised and unsupervised machine learning methods for detection of noncompliance and identity fraud.[5] According to the Government Accountability Office (GAO, 2018), IRS managers meet once a week during the filing season to evaluate the data and determine if adjustments should be made to the RRP to address emerging noncompliant strategies.

The program's return on investment is one way to evaluate its effectiveness. Work on developing the RRP began in 2009, and it was fully implemented in October 2016, several months before the beginning of the 2017 filing season. From 2009 through 2019, the IRS invested $597 million in the RRP.[6] Permanently frozen refunds totaled nearly $11 billion from 2015 (when the program began to phase in) through 2019.[7] Thus, the average return on investment was 18:1 over the first decade of the program's development and initial implementation and will further increase if the program's savings continue to outpace capital expenditures.[8] That is a high average return on investment, but the

[3] For example, the RRP could flag a return for further review if it does not include a verification of earnings for the earned income tax credit. The taxpayer's employer could be contacted to confirm the income and withholding reported on the tax return.

[4] The RRP replaced the Electronic Fraud Detection System, which was put in place in 1994. The Electronic Fraud Detection System also included a machine learning component—a decision tree algorithm (Treasury Department, 2013)—but over time the system had not kept up with the evolution of identity-fraud strategies. Another concern with the Electronic Fraud Detection System was that it generated only one score for each return, whereas the RRP yields a set of scores that enables it to individually assess tax returns across all identity-fraud and other noncompliance categories (TIGTA, 2017).

[5] Various Treasury documents contain descriptions of the machine learning methods used in the RRP, though those descriptions differ somewhat. According to the IRS's FY 2022 Capital Investment Plan (IRS, 2021b), the RRP is an anomaly-detection system. The Treasury's 2021 Consolidated Privacy and Civil Liberties Report states that the RRP relies on both supervised and unsupervised learning methods, listing specific algorithms (using decision trees, neural networks, and logistic regressions) associated with supervised learning (Treasury Department, 2021).

[6] Data on actual investment costs are available in the IRS's annual Capital Investment Plan (IRS 2016, 2017, 2018a, 2019a, 2020b, 2021b): investment costs were $151 million for the pre–fiscal year (FY) 2015 period, $42 million in FY 2015, $100 million in FY 2016, $90 million in FY 2017, $109 million in FY 2018, and $105 million in FY 2019.

[7] Data on permanently frozen refunds are found in the IRS Congressional Budget Justification and Annual Performance Report and Plan: permanently frozen funds totaled $8.9 billion for FYs 2015 through 2018 (IRS, 2020a) and $2 billion in FY 2019 (IRS, 2021a).

[8] Sarin and Summers estimated a 50:1 return on investment from the RRP in 2017 (Sarin and Summers, 2019). Their calculations, however, are incomplete: They compare the amount of protected revenue, or the permanently frozen refunds, in FY 2017 ($4.4 billion) with the costs incurred only in that year ($90 million), omitting the costs incurred during the development period. Using a similar method, the IRS estimates that the return on investment in FY 2020 was 56:1 (IRS, 2021a).

decision to expand the RRP should be based on the marginal return on investment: the additional amount of stopped refunds relative to the additional dollar of investment.

A second metric for evaluating enforcement actions is the no-change rate: the percentage of audited returns settled in the taxpayer's favor, resulting in no changes to their tax return. High no-change rates may indicate poor targeting of audits, resulting in an increased burden on compliant taxpayers and wasted expenditures by the IRS. The combination of a high return on investment and a low no-change rate would generally better indicate the productivity of an audit than either one alone.

Data are not available on the no-change rate associated with the RRP, but a study by researchers from the MITRE Corporation (MITRE) and the IRS of correspondence audits may be relevant. They conducted a comprehensive analysis of several types of machine learning methods for the selection of audits of correspondence returns, filed for tax years 2013 through 2016, with either itemized deductions (Schedule A) or self-employment income (Schedule C) and compared their findings with the pre-RRP methods (Howard et al., 2021).[9] Testing multiple types of algorithms and comparing the results to audits selected through the existing business rules, the researchers demonstrated that it is challenging—though not impossible—to find a methodology that achieves both an increase in assessed revenues and a reduction in no-change audits relative to current approaches.

Quantitative measures such as the return on investment and no-change rate are informative, but they should be viewed in broader contexts. For example, a low return on investment and a high no-change rate could be caused by an understaffed IRS unable to counter the arguments of the taxpayer's tax advisers, especially when the challenged positions are complicated or ambiguous and the taxpayer can afford a prolonged dispute.[10] Conversely, a high return on investment and a low no-change rate could occur when compliant taxpayers cannot provide documentation in support of their claim of a refundable tax credit.[11] In the first instance, the taxpayer (often a large business) prevails and pays less tax than the IRS auditor initially recommended; in the second instance, the taxpayer (often an individual with low income) loses a tax benefit for which they were actually eligible.

CHALLENGES OF MACHINE LEARNING

Although machine learning may have certain advantages over the current methods used by the IRS for audit selection, it has risks as well. Though some of those risks emanate from the approach's methodology, they can be magnified by issues specific to the US tax system.

Insufficient data. A concern with machine learning is that it may perpetuate or create errors in the selection of audits. Because supervised learning is based on the results of past enforcement actions, the errors that were difficult to detect using old methods may still be undetected. Building on historical data, the machine learning model may be unduly influenced by past cases settled in noncompliant taxpayers' favor solely because of insufficient IRS resources and past

[9] For each category, MITRE and the IRS tried various separate econometric specifications: pairwise, which compared returns by ranking pairs against each other in terms of revenue; hurdle (two-stage regression), which combined two trained models with a binary classifier (for no-change audits) with a regression model (revenue); and penalized regression, which accounted for the probability of a no-change audit when training a second regression model. The penalized regression was the best of the three alternative methods in terms of both increasing revenue and reducing the number of no-change audits.

[10] Holtzblatt (2021) reports that the no-change rate for large businesses with $10 million or more of assets was 38 percent, on average, in 2019. In contrast, the no-change rate for individual income tax returns was 11 percent.

[11] Guyton and colleagues found a high nonresponse rate to correspondent audits of earned income tax credit claimants (Guyton et al., 2019). They speculate that the high nonresponse rate could occur partly because many compliant taxpayers are unwilling or unable to provide documentation to the IRS, perhaps because they fear the agency, do not understand the IRS's request (possibly because of a lack of fluency in English), or are unable to obtain the type of documents that could support their claims (e.g., proof that their child lived with them for more than half the year).

cases settled in the IRS's favor solely because compliant taxpayers did not respond. Effectively integrating advice from examiners and findings from tax compliance research with machine learning may help mitigate this concern.[12]

Another data challenge is a direct result of the recent cuts to the IRS budget. Supervised learning depends on examples that have shown a relationship between an observable characteristic (or set of characteristics) and a behavior (noncompliance). But with historically low audit rates, especially among the wealthier taxpayers, will the number of examples for the computers to analyze be sufficient, at least until audit rates climb again? The paucity of audits also makes it difficult to validate the test results when developing machine learning techniques; without historical data, the IRS would likely have to conduct audits of similar taxpayers using conventional methods in the trial period to compare with results of audits selected through machine learning.

Interpreting findings. The why and the how behind algorithms may not be obvious to the modelers and examination staff within the IRS. Unlike economic models, machine learning techniques do not begin with a theoretical model that is then tested in empirical research (Athey and Imbens, 2019). Nor are those techniques comparable with the issue-based Dependent Database, with its business rules based on the observed findings of the IRS's compliance studies. In Engstrom and colleagues (2020), the authors write, "Model complexity can make it difficult to isolate the contribution of any particular variable to the result." Yet, knowing the source of the problem may aid both tax administrators and policy analysts in developing solutions, either through the application of issue-related interventions or changes to tax laws. And more knowledge about specific triggers for audits may also provide insight into the reasons why compliant taxpayers are selected for audits.

Transparency. From a tax administrator's perspective, a lack of transparency can be desirable to prevent taxpayers from discerning the methods used to identify evasion. The IRS traditionally protects its audit selection priorities, even from most of its staff, though successful tax advisers may become adept at determining audit triggers.

However, opaque selection methods combined with the difficulty of interpreting the underlying algorithms may lead to undesirable outcomes. Dutch tax authorities were fined 3.7 million euros by the country's privacy regulators when tens of thousands of people were denied child-care benefits for years because of faulty risk profiles created by a machine-learning algorithm (Heikkila, 2022). Though the precise source of the problem is unclear, the profiles appeared to treat income and ethnicity as risk factors. In addition to the fine, a new system of checks and balances with more human oversight is being developed to prevent similar occurrences.

Behavioral responses. A challenge for tax administration is that taxpayers adjust their evasion strategies as they (or their informed advisers) pick up on the IRS's targets for audits and other enforcement actions. Relative to current audit selection methods, machine learning has two advantages for keeping up with changes in taxpayers' strategies. Unsupervised learning techniques like clustering observe data in real time, and, as currently applied by the IRS in the RRP, the findings are examined weekly by the IRS staff, who can adjust the algorithms.

Machine learning still shares a disadvantage with the conventional methods of audit detection: the inability to anticipate behavioral responses. However, researchers from MITRE and the Massachusetts Institute of Technology found that a simulation model could complement machine learning techniques and be used to inform both supervised and unsupervised learning methods (Hemberg et al., 2015).

[12] One challenge, however, is that compliance studies do not detect all noncompliance. The IRS regularly conducts the National Research Program, which consists of audits of a random sample of individual income taxpayers, to obtain measures of tax compliance. But National Research Program audits do not detect all instances of unreported income. To estimate the amount of undetected income, the IRS uses a detection-controlled estimation method, which can add substantial amounts to tax gap measures.

Complexity. The tax code is complicated, and that complexity affects the development of machine learning models. First, data scientists and the IRS staff must be able to communicate to ensure that the examples are labeled accurately and that findings are interpreted and applied correctly.

In addition, the tax code's complexity often results in differing interpretations of subtle provisions. There are many shades of gray between aggressive but legal tax strategies and illegal tax evasion (Hemel, Holtzblatt, and Rosenthal, 2022). Sometimes, resolving disputes between taxpayers and the IRS is left to the courts, though those cases may take years to conclude and the findings by courts in different jurisdictions may conflict. The large enforcement dataset maintained by the IRS has not followed taxpayers' disputes once they enter the judicial system.

Currently, at least two research projects are underway that are examining the application of machine learning to two of the most complicated areas of the tax code. In coordination with the IRS, a team of Stanford-affiliated researchers is leveraging machine learning approaches to predict partnership noncompliance, beginning with the construction of partnership entity network structures (Goldin et al, 2022). Using unsupervised machine learning, researchers from the IRS and MITRE are examining ways to improve audit selection of taxpayers from the global high-wealth population, one of the priority compliance areas for the IRS (Olson, et al, 2022).

CONCLUSION

Using machine learning could lead to improvement in the targeting of the IRS's enforcement actions. Still, the IRS will encounter challenges in the development and implementation of those tools.

One of the largest challenges is resources. In theory, improvements could be achieved simply by reallocating current resources away from conventional approaches to greater reliance on more efficient machine learning tools. But resources cannot be easily shifted. The development of machine learning tools would require hiring new employees with skills not yet acquired by the current staff. Modernization of the IRS's technology would facilitate both development and implementation of those tools. Long-term investments in staffing and technology would also require a multi-year commitment to funding.

With those investments, the IRS could yield a significant return on investment, if the experience of the RRP is repeated. At the same time, though, the agency should ensure that the new tools do not place undue burden on compliant taxpayers.

REFERENCES

Athey, Susan, and Guido Imbens. 2019. "Machine Learning Methods that Economists Should Know About." Annual Review of Economics. 11: 685-725

Engstrom, David Freeman, Daniel Ho, Catherine Sharkey, and Mariano-Florentino Cuellar. 2020. Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. Report Submitted to the Administrative Conference of the United States.

Forman, Fred, and Charles Rossotti. 2021. The Business Case for IRS Transformation. Government Executive Ebook.

GAO (Government Accountability Office). 2016. Wage and Investment Division Should Define Audit Objectives and Refine Internal Controls. GAO-16-102. Washington, DC: GAO.

--------. 2018. IRS Could Further Leverage the Return Review Program to Strengthen Tax Enforcement. GAO-18-544. Washington, DC: GAO.

--------. 2021. Tax Administration: Better Coordination Could Improve IRS's Use of Third-Party Information Reporting to Help Reduce the Tax Gap. GAO-21-102. Washington, DC: Government Accountability Office.

Goldin, Jacob, Ryan Hess, Daniel Ho, Rebecca Lester, and Mansheej Paul. 2022. Presented at 12th Annual Internal Revenue Service/Tax Policy Center Joint Research Conference on Tax Administration, June 16.

Guyton, John, Patrick Langetieg, Daniel Reck, Max Risch, and Gabriel Zucman. 2021. Tax Evasion at the Top of the Income Distribution: Theory and Evidence. National Bureau of Economic Research Working Paper No. 28542. Cambridge, MA: National Bureau of Economic Research.

Guyton, John, Kara Leibel, Dayanand Manoli, Ankur Patel, Mark Payne, and Brenda Schafer. 2019. The Effects of EITC Correspondence Audits on Low-Income Earners. National Bureau of Economic Research Working Paper No. 24465. Cambridge, MA: National Bureau of Economic Research.

Heikkela, Melissa. 2022. "Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms." *Politico,* March 29.

Hemberg, Erik, Jacob Rosen, Geoff Warner, Sanith Wijesinghe, and Una-May O'Reilly. 2015. "Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood." Proceedings of the 15th International Conference on Artificial Intelligence and Law. June, 2015: pp 79-88. San Diego, California.

Hemel, Daniel, Janet Holtzblatt, and Steven Rosenthal. 2022. The Tax Gap's Many Shades of Gray. Washington DC: Brookings-Urban Tax Policy Center.

Holtzblatt, Janet. 2021. "Too Many IRS Audits of Big Businesses Result in No Change in Tax Liability." *Tax Vox* (blog), April 19.

Howard, Ben, Lucia Lykke, David Pinski, and Alan Plumley. 2021. "Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation" In The IRS Research Bulletin: Proceedings of the 2020 IRS/TPC Research Conference (Publication 1500, Rev. 5-2021), compiled and edited by Alan Plumley, 147–64. Washington, DC: IRS.

IRS (Internal Revenue Service). 2011. Internal Revenue Service Data Book, 2010. Publication 55B. Washington DC: IRS.

--------. 2016. FY 2017 Capital Investment Plan. Washington DC: IRS.

--------. 2017. FY 2018 Capital Investment Plan. Washington DC: IRS.

--------. 2018a. FY 2019 Capital Investment Plan. Washington DC: IRS.

--------. 2018b. Strategic Plan FY2018-2022. Publication 3744. Washington, DC: IRS.

--------. 2019a. FY 2020 Capital Investment Plan. Washington DC: IRS.

--------. 2019b. Internal Revenue Service Data Book, 2018. Publication 55B. Washington DC:

--------. 2019c. IRS Integrated Modernization Business Plan. Publication 5336. Washington DC: IRS.

--------. 2020a. Congressional Budget Justification and Annual Performance Report and Plan, FY 2021 Washington DC: IRS.

--------. 2020b. FY 2021 Capital Investment Plan. Washington DC: IRS.

--------. 2020c. Internal Revenue Service Data Book, 2019. Publication 55B. Washington, DC:

--------. 2021a. Congressional Budget Justification and Annual Performance Report and Plan, Fiscal Year 2022. Publication 4450. Washington DC: IRS.

--------. 2021b. FY 2022 Capital Investment Plan. Washington DC: IRS.

OECD (Organisation for Economic Co-operation and Development). 2021. Tax Administration 2021: Comparative Information on OECD and Other Advanced and Emerging Economies. Paris, France: Organisation for Economic Co-operation and Development.

Olson, Matt, Ben Howard, Devika Mahoney-Nair, and Annette Portz. 2022. "Graph-Based Machine Learning Methods for Case Selection and Population Segmentation." Presented at 12th Annual Internal Revenue Service/Tax Policy Center Joint Research Conference on Tax Administration, June 16.

Sarin, Natasha, and Lawrence Summers. 2019. Shrinking the Tax Gap: Approaches and Revenue Protection. National Bureau of Economic Research Working Paper No. 26475. Cambridge, MA: National Bureau of Economic Research.

Treasury Department. 2013. 2013 Annual Privacy and Data Mining Reports. Washington, DC: Treasury Department.

--------. 2021. 2021 Consolidated Privacy and Civil Liberties Reports. Washington, DC: Treasury Department.

Treasury Inspector General for Tax Administration. 2017. The Return Review Program Increases Fraud Detection: However Full Retirement of the Electronic Fraud Detection System Will Be Delayed. 2017-20-080. Washington, DC: Treasury Inspector General for Tax Administration.

--------. 2019. The Use of Schedule K-1 Data to Address Taxpayer Noncompliance Can Be Improved. 2019-30-078. Washington, DC: Treasury Inspector General for Tax Administration.

--------. 2020. Individual Returns with Large Business Losses and No Income Pose Significant Compliance Risk. 2020-30-056. Washington, DC: Treasury Inspector General for Tax Administration.

ACKNOWLEDGMENTS

This brief was funded by Arnold Ventures We are grateful to them and to all our funders, who make it possible for the Urban-Brookings Tax Policy Center to advance its mission.

Janet Holtzblatt is a Senior Fellow at the Urban-Brookings Tax Policy Center. Alex Engler is a Fellow in Governance Studies at The Brookings Institution. The authors wish to thank Chenxi Lu for helpful comments and suggestions. All errors are our own. The views expressed are those of the authors and should not be attributed to the Urban-Brookings Tax Policy Center, the Urban Institute, the Brookings Institution, their trustees, or their funders.