

DocuSign Envelope ID: 7989860E-9C5D-4E9D-B455-6DA97C4127B8

Towards Reduced Administrative Burdens:
Performance Management, Machine Learning,
and Evidence Building in the Federal Government

A Dissertation

Submitted to
College of Public Affairs
University of Baltimore

In Partial Fulfillment of the requirements for the degree
of
Doctor of Public Administration

By
Joel D. Nantais

DocuSigned by:
Aaron Wachhaus
6885DD68277D441
Dr. Aaron Wachhaus, Chair

DocuSigned by:
Roger Hartley
CC36DD9AD2AC4AE
Dr. Roger Hartley

DocuSigned by:
Susan Sterett
565EA89E96EE41B
Dr. Susan Sterett

Abstract

Administrative burdens can inhibit how individuals access government services and benefits and reduce the effectiveness of programs. These burdens, including learning, compliance, and psychological costs, are experienced disparately and may cause greater negative impacts on individuals who would benefit most from the programs. This is especially true when they are “hidden” within policy and program design and implementation, thereby avoiding normal administrative law procedures that would allow citizens awareness and feedback opportunities. This study explores how existing standards for performance management and measurement through data analytics in the federal government can be leveraged to identify and measure administrative burdens with the aim of reducing their impact. This work focuses on the use of machine learning to solve implementation problems like administrative burdens, complete with the design, development, and implementation considerations that are specific to machine learning in the United States federal government. Finally, this work explores how to use the existing requirements for evidence-based policymaking and evaluation in the federal government to determine the impact of the machine learning solutions, as well as the impact of the reduced administrative burdens on the outcomes and goals of programs.

The approach of this research is to build on the existing academic literature, federal government requirements, and guidance to create three frameworks: identification and measurement, machine learning solutions, and evidence-building evaluations. Framework 1 provides a path to identifying, defining, and measuring administrative burdens within performance management processes. Framework 2 shows how to incorporate the nascent Federal government principles and guidance with academic and industry best practices to design, develop, and implement machine learning solutions in the public sector to reduce administrative burdens, which are identified and measured by Framework 1. Framework 3 is a guide to using existing federal government evidence-building and evaluation guidance to evaluate the implementations and impacts of the machine learning solutions and reduced administrative burdens.

This research demonstrates that administrative burdens in the federal government systems and processes can be identified and addressed without new legislation, regulation, or resources and that machine learning techniques are poised to provide solutions to public problems. This presents an opportunity for the federal government to refocus on providing performance data, administrative data, and information about existing uses of machine learning available to the

public in a way that can benefit the academic research field; this lack of availability has follow-on impacts on the public sector. Additionally, this study shows that the field of administrative burden research needs to adopt shared definitions, measurement criteria, and approaches in order to build on existing theory and case studies and to magnify the impact of this research on the public sector. This research provides standard definitions of administrative burdens classification and measurement; a guide for agencies to reduce administrative burdens with performance management; practical guidance for applied machine learning in the federal government; an extension of evidence-based policy and evaluation research to focus on applied machine learning as well as the impact of reduced administrative burdens on program outcomes. These contributions benefit researchers focused on administrative burdens and provide practical support to government administrators.

Acknowledgments

I need to thank my dissertation committee, Dr. Aaron Wachhaus, Dr. Roger Hartley, and Dr. Susan Sterett, for their expert and patient guidance as I worked through this research and the program requirements. Dr. Wachhaus specifically spent hours over the past several years working through these topics with me, helping me discover and focus on what was most important to me, and helping me map out a process to accomplish this research while enabling me to focus on the areas I am most passionate about. Additionally, I want to further thank the faculty at the University of Baltimore's School of Public and International Affairs and the UMBC Data Science faculty, who helped guide me on this journey, challenged me, and encouraged my work and focus. Their guidance has helped me learn about myself and how to create and contribute knowledge in my field and my professional life.

Secondly, I dedicate this work to the person that helped me throughout nearly every stage of my life, my mother. She was one of my greatest teachers through direct lessons and her example. She encouraged me to stretch myself and grow, especially when the world seemed most difficult; she showed me how to be confident and humble in the face of adversity. And she showed me how to appreciate learning and how to enrich my life through the pursuit of knowledge. Without her love, support, and insistence, I would never have begun my college journey, let alone accomplish this research. Sadly, she passed away while I was in the midst of this program and will not be with me as I graduate, but her life and teaching live in me.

Finally, I also thank my wife, Holly, who supported me every day from the moment I conceived of this journey through its successful end.

Table of Contents

Abstract..... 1

Acknowledgments..... 4

Table of Contents 5

List of Tables 10

List of Figures 12

Chapter 1 Overview of Research 14

Problem Statement - Towards More Effective and Efficient Government..... 14

Chapter 2 Literature Review 17

Administrative Burdens 18

 Overview 18

 Government Interactions and Administrative Burdens..... 21

 Types of Costs..... 26

 Intentional vs. Unintentional Burdens 30

 Negative Impacts of Administrative Burdens..... 33

 Measuring Administrative Burdens 38

Performance Management 40

 Overview 40

 Definitions..... 43

 Why Performance Management is Important 45

 History of Performance Management in the U.S. Federal Government..... 46

 Performance Management Techniques/Guidance 52

 Federal Document Analysis for Measuring Administrative Burdens..... 60

 Assessment of Performance Information in the United States Government 66

 Successful Use of Performance Information in Federal Government Agencies 69

 Summary 75

The Intersection Between Performance Management and Evaluation 75

Implementing Performance Management..... 77

Performance Management and Administrative Burdens 81

Measuring Administrative Burdens 81

Big Data, Artificial Intelligence, and Machine Learning in Public Administration..... 87

 Definitions..... 88

 Implementations of Machine Learning..... 98

 Benefits of Public Sector Use of Machine Learning 102

 Risks of machine learning..... 103

 Government Considerations for Machine Learning Use 109

 Government Approaches to Machine Learning Models..... 115

 ML in the Public Sector – Overview and History..... 116

 ML Frameworks and MLOps 126

 Specific Considerations for Public Sector 130

 E.O. 13960 Framework and MLOps..... 130

Summary..... 134

Evidence-based Policymaking 136

 Overview..... 136

 Definitions..... 138

 History of Evidence-based Policy in the U.S..... 139

 Current Evidence-based policy in Federal Government..... 144

 Assessment of Evidence-Based Policymaking 153

 Criticisms of Evidence-Based Policymaking..... 155

 Use of Evidence by Policymakers 160

 Sub-legislature Evidence-based Policymaking..... 160

 Summary and Next Steps..... 161

Chapter 3 Research Design and Methods 163

- The Research and Solution 163
- The Three Frameworks 165
 - Framework 1 - Measuring Administrative Burdens within Performance Management 167
 - Framework 2 - Applying machine learning to Reduce Administrative Burdens.... 169
 - Framework 3 - Evidence-based Policymaking to Evaluate Outcomes After Applying machine learning to Reduce Administrative Burdens 172
 - Tying the Frameworks Together..... 174
- Methods..... 175
- Research Questions 175
 - Research Question 1 - Framework 1: Identify and measure Administrative Burdens within the performance management process? 175
 - Research Question 2 - Framework 2: How can we implement machine learning solutions in public sector programs to reduce administrative burdens? 176
 - Research Question 3 - Framework 3: How do we evaluate the impact of machine learning solutions on administrative burdens, and on the outcomes of the policy? 177
- Research Methods..... 177
 - Triangulation..... 178
 - Framework 1 Methods 179
 - Framework 2 Methods 185
 - Framework 3 Methods 189
- Chapter 4 Framework 1 Results - Identification and Measurement of Administrative Burdens with Performance Management 192
- Overview and Goals..... 192
- Framework Methods: Performance Measurement for Administrative Burden Costs 193
 - Administrative Burden and Performance Management Literature Review and Document Analysis Methods 194
- Results: Agency Priority Goal Framework..... 196

- Learning Costs Measurement 205
- Compliance Cost Measurement 209
- Psychological Cost Measurement 215
- Detailed APG for Administrative Burdens 218
- Performance Measurement and Administrative Burden Information Mining 222
- Framework 1 Summary..... 227

- Chapter 5 Framework 2 Results - Reducing or Eliminating Administrative Burdens with Machine Learning 229
 - Overview and Goals..... 229
 - Framework Methods: Machine Learning Solutions for Administrative Burdens 230
 - Results: Machine Learning Framework for Administrative Burden Reductions 231
 - ML for Learning Cost Solutions 238
 - ML for Compliance Cost Solutions 241
 - ML for Psychological Cost Solutions 249
 - Results: MLOps Example for ML Solutions 266
 - MLOps Checklist Example..... 269
 - Results: Information Mining for Applied ML Solutions 272
 - Framework 2 Summary..... 284

- Chapter 6 Framework 3 Results - Evidence-based Policymaking: Evaluating the Outcomes of Machine Learning to Reduce or Eliminate Administrative Burdens 285
 - Overview 285
 - Results: Evaluation framework for machine learning solutions for administrative burdens..... 286
 - Evaluation Questions to Be Answered 287
 - Lead Office, Points of Contact..... 287
 - Rational for the Evaluation 287
 - Purpose of the Evaluation 288
 - Audience 288

Outcome Evaluation.....	288
Summative/Impact Evaluation.....	288
Dissemination Plan	289
Framework 3 Summary.....	290
Chapter 7 Conclusions and Discussion.....	291
Overview of the Resulting Frameworks	291
Implementation of the Frameworks	292
Potential Limitations of the Frameworks.....	293
Next Steps and Future Research Agenda.....	294
Appendix A – Raw Data.....	298
Bibliography	321

List of Tables

Table 2-1: Types of Government Interactions; adapted from (Heinrich, 2016; Katz, 1975)21

Table 2-2: Components of Administrative Burdens from (Herd & Moynihan, 2018, p. 56)26

Table 2-3: Administrative Burdens Diagnostic Questions 85

Table 2-4: Techniques to reduce Administrative Burdens in Program Implementation .. 86

Table 2-5: E.O. 13960 Responsible AI Principles 121

Table 2-6: Analysis of AI Principles in Agency frameworks..... 125

Table 2-7: Analysis of AI Principles in Agency Frameworks 125

Table 2-8: Agency Unique Principles Used..... 125

Table 2-9: Mapping E.O. 13960 Principles to MLOps..... 133

Table 2-10: OMB Guidance on Evaluation Standards (from M-20-12)..... 148

Table 2-11: OMB Evaluation Plan Requirements (from M-20-12)..... 149

Table 4-1: Counts of Measurement Example per Type of Cost (author) 199

Table 4-2: Count of Unique Cause and Measurement by Cost types 200

Table 4-3: Simplified Administrative Burden Measurements 201

Table 4-4: Counts of Simplified Measurement..... 202

Table 4-5: Refined Measurement Counts by Cost Type..... 203

Table 4-6: Count of Simplified Measurements by Cost Type 204

Table 4-7: Learning Cost Measurements from Current Research 208

Table 4-8: Compliance Cost Measurements from Current Research 214

Table 4-9: Psychological Cost Measurement from Current Research..... 217

Table 4-10: Summary of Opensource Results 227

Table 5-1: Types of ML Models (Burkov, 2019; Raschka & Mirjalili, 2019) 232

Table 5-2: Count of Simplified ML Solutions..... 235

Table 5-3: Simplified ML Solution Definitions..... 236

Table 5-4: ML Solutions for Learning Costs (author) 240

Table 5-5: ML Solutions for Compliance Costs (author) 249

Table 5-6: ML Solutions for Psychological Costs (author) 264

Table 5-7: Count of Simplified ML Solution per Cost Types 265

Table 5-8: Counts of ML Solution Types per ML Solution Categories (author) 266

Table 5-9: Responsible AI Principles Mapped to MLOps Process 268

Table 5-10:Count of Search Terms 272

Table 5-11: Search Terms Found by Year of Publication 274

Table 5-12: Search Terms by Year 2010-2022 (YTD)..... 275

Table 5-13: Results of Search Terms by Agency from all years 276

Table 5-14: Top 15 Agency Terms by Document Types 277

Table 5-15: Rule Titles for Top Agencies (2010-2022) 279

Table 5-16: ACUS Table of Top AI Use Cases by Agency (Engstrom et al., 2020) 280

Table 5-17: AI Use Cases Coded by Interaction Types..... 283

Table 5-18: ACUS AI Use Cases for Administrative Burdens..... 284

List of Figures

Figure 1-1: Theory of Change Model 15

Figure 2-1: PIC Playbook Performance Management Cycle..... 58

Figure 2-2: OIRA Burden Activities..... 63

Figure 2-3: GPRAMA Goals Overview 80

Figure 2-4: Relationship between Big Data and Artificial Intelligence 90

Figure 2-5: Animal and Computer Neurons (Raschka & Mirjalili, 2019)..... 94

Figure 2-6: Machine Learning vs. Artificial Intelligence (Wolf et al., 2020) 96

Figure 2-7: Types and examples of Machine Learning 96

Figure 2-8: GAO Report Summary Table (U.S. Government Accountability Office, 2021b)
..... 123

Figure 2-9: MLOps Cycle (Burkov, 2019))..... 127

Figure 2-10: History of Evidence-Based Policy and Evaluation in Federal Government (U.S.
Government Accountability Office, 2021a)) 142

Figure 2-11: Components of Evidence (M-19-23) 146

Figure 2-12: Learning Agenda Process (M-19-23)..... 147

Figure 2-13: Evidence Logic Model (M-21-27) 150

Figure 2-14: Performance Measurement Definition and Methodology (M-21-27)..... 150

Figure 2-15: Government Evaluation Types and Methodologies (U.S. Government
Accountability Office, 2021a) 152

Figure 3-1: The Three Frameworks (author) 166

Figure 3-2: Framework 1 (author) 167

Figure 3-3: Framework 2 (author) 169

Figure 3-4: Framework 3 (author) 172

Figure 4-1: Count per Cost Type 202

Figure 4-2: Example Detailed APG 220

Figure 4-3: Example APG Milestone/Indicator Tracker 221

Figure 5-1: Framework 2 (author) 229

Figure 5-2: Machine Learning Types and Examples (Granville, 2017) 233

Figure 5-3: MLOps Cycle..... 267

Figure 5-4: ACUS Graph of AI Use Case by Policy Areas 281

Figure 5-5: ACUS Graph of AI Use Case by Task..... 281

Figure 5-6: ACUS Graphs of AI Use Cases by Development Stages and Development Type
..... 282

Figure 5-7: ACUS Graph of AI Use Cases by ML Method 282

Figure 5-8: ACUS Graph of AI Use Cases by Data Type 283

Figure 6-1: Framework 3 285

Figure 6-2: Theory of Change for ML methods to Reduce Administrative Burdens 287

Chapter 1 Overview of Research

Problem Statement - Towards More Effective and Efficient Government

I believe that performance management and evidence-based policymaking can make government programs and policies more effective. This has shown to be true at the agency and sub-agency levels but still has promise at higher levels of government (Jennings & Hall, 2012; Moynihan & Kroll, 2016). One important area of research which affects both performance management and evidence-based policymaking is administrative burdens that induce disparate costs on individuals seeking government benefits or programs (Burden et al., 2012; Herd & Moynihan, 2018). I believe that advanced data analytics such as machine learning fueled by big data processing can lower administrative burdens. To do this, I believe that machine learning techniques can be applied to federal government programs to more effectively and efficiently determine eligibility, facilitate application processes, and administer government programs. This would improve the ability of public administrators to measure the performance of programs and increase the ability to determine if the programs are achieving their intended outcomes.

I believe machine learning can be used to reduce or eliminate administrative burdens for citizens in government programs. These techniques can help identify potentially eligible beneficiaries for programs more accurately, streamline or eliminate application processes, and more effectively and efficiently monitor compliance with program rules, thereby reducing and eliminating the administrative burdens of learning costs, compliance costs, and psychological costs (Agrawal et al., 2018; Fuentes, 2018). I believe that the focus on reducing or eliminating administrative burdens will further improve the ability of performance management and evidence-based policymaking to improve government programs.

To help research and achieve these beliefs, I will create and interweave three frameworks for researchers and administrators. The first framework will focus on identifying and measuring administrative burdens in programs. This will allow the accurate inclusion of administrative burdens in larger performance management models. The second framework will provide a guide to implement machine learning solutions in government programs to reduce administrative burdens. The framework will guide how to design, create, and implement effective machine learning solutions in public programs. This framework is needed because public sector machine learning implementation has unique challenges not found in the private sector, such as legal, ethical, administrative law, oversight, and transparency requirements (Calo, 2017; Gohwong, 2015). The third framework will include methods how to measure and evaluate the impact of the

machine learning techniques on administrative burdens in the program based on the intended policy outcomes of the program as adopted. This impact measurement will consist of summative evaluation techniques to understand the outcomes of the programs as compared with the policy goals.

These frameworks can then be used by administrators and researchers for purposes such as: more accurately and consistently identifying and quantifying administrative burdens in public programs; experimenting with machine learning solutions in programs to reduce administrative burdens (among other goals), or evaluating the effects of reduced administrative burdens on programs and policy effectiveness. I believe this will allow us to better catalog administrative burdens in programs throughout the public sector and build our understanding of their impacts on program outcomes. I also believe this will help encourage the understanding of administrative burdens and allow them to be more effectively discussed as part of program design and implementation.

***Theory of change:** If we can identify and measure administrative burdens, we can implement machine learning solutions to reduce or eliminate those burdens, and then we can improve the outcomes of programs to make them more effective.*

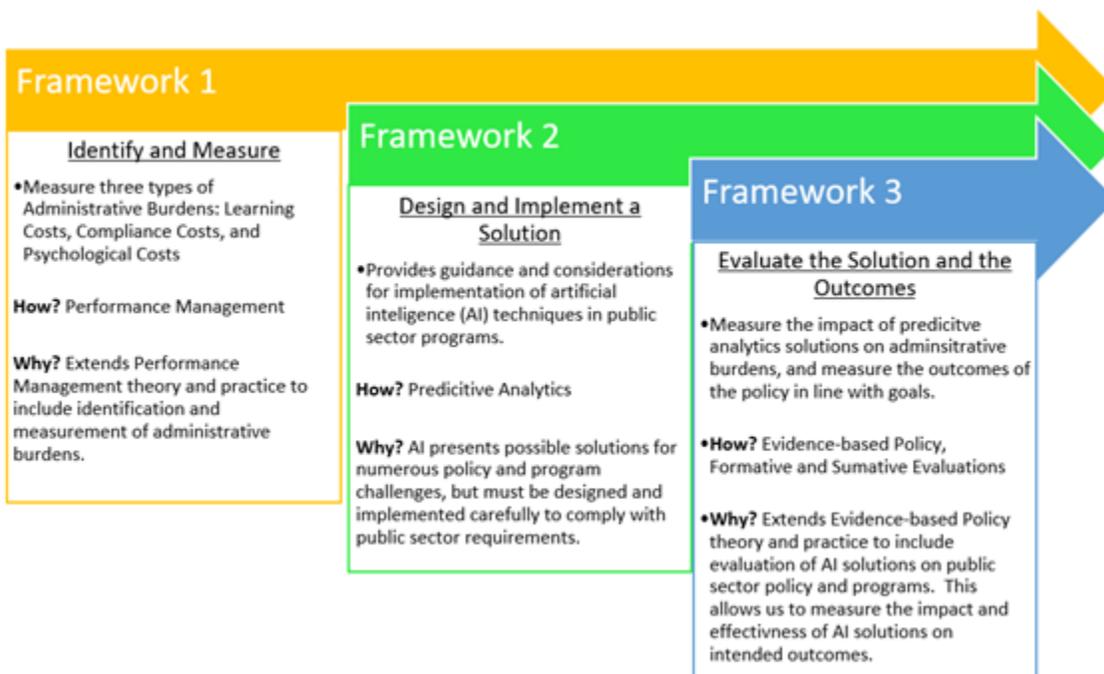


Figure 1-1: Theory of Change Model

These three frameworks together will allow me to implement and evaluate my theory of change, but all three are needed together to enable this. The frameworks can be used

independently but are more effective towards the overall understanding and evaluation of machine learning solutions and policy impacts when used together appropriately. As shown in Figure 1-1 above: Framework 1 is the measurement of administrative burdens using performance management; Framework 2 applies machine learning to reduce administrative burdens, and Framework 3 is used to evaluate the impact of Framework 2 on administrative burdens and the overall program or policy through extends evidence-based policy.

Ultimately, I believe this research will lay the groundwork for the design and use of machine learning in government programs to further the goals of performance management and evidence-based policymaking. Additionally, I believe this research will add to the field of administrative burdens research directly by providing a framework for the measurement of burdens which will both help provide an opportunity to inflict more transparency on them as well as enable evaluation of the impacts on attempts to lessen these burdens in programs. I believe both of these benefits will positively impact the design and implementation of government programs to achieve the intended policy goals. This research will also link to evidence-based policymaking by helping formulate the evaluation of the policy implementation and outcomes in line with the current focus and application of evidence-based policy (Cartwright & Hardie, 2012; Orr, 2018; Patton, 2012). In this research, I'll describe these theories and practices to show how they interconnect and how focusing on using machine learning to reduce or eliminate administrative burdens will provide theoretical and applied contributions to these fields.

In the following sections, I'll provide background and overview of performance management, machine learning, and evidence-based policymaking to define them and to show how this work will contribute to the fields, link them together, and how I believe they can work in concert to improve public administration of programs. Taken narrowly, I hope to provide clear frameworks that can be used to build, implement, and evaluate machine learning solutions to reduce administrative burdens. More generally, I believe these frameworks can also be adapted and applied to other potential administrative burden reduction solutions as well as machine learning applications for other purposes in the public sector.

Chapter 2 - Literature Review

In this chapter, I will explore the importance of administrative burdens as it has been identified in current research. I will show how these burdens implicitly or inadvertently impact the effectiveness of programs as well as the impact they have on the recipients, or potential recipients, of the program benefits. I will then discuss other important areas of research and focus that will allow me to design and build solutions to reduce or eliminate administrative burdens in the federal government to make policies and programs more effective. To do this, I will explore the history and importance of performance management (including performance information and performance measurement) in the federal government. I will look at the theories, tools, and requirements which have been shown to be effective, as well as the systems to better implement them. I do this to build onto existing performance management requirements rather than creating new requirements for performance measurement systems to identify and measure administrative burdens in the federal government.

This chapter will also detail the emerging focus on artificial intelligence, and specifically the tools and techniques of machine learning. I will highlight the unique challenges still being wrestled with when attempting to build machine learning solutions. These include how to remove and prevent codifying biases in models, how to allow for algorithmic transparency, and how to evaluate the accuracy of the model based on the impact of errors on individuals. Additionally, I will look at the emerging focus on design and implementation principles for machine learning in the public sector. I will highlight the many challenges and considerations needed to comply with the government's use of machine learning. Through this exploration, I will show how machine learning solutions in the federal government can be used to reduce administrative burdens and overcome these challenges.

Lastly, this chapter will look at evidence-based policymaking, which is focused on leveraging evaluation and scientific techniques in the public sector to more rationally and methodically test policy solutions against the expected program goals and outcomes to provide policymakers with better feedback on what works and what doesn't work, but also why things work or do not (Heinrich, 2007). This is supposed to provide better feedback to make more informed decisions about policy and expenditures, as well as to focus new initiatives on those which have either already been proven effective or provide a mechanism to monitor and test their effectiveness. Specifically, this work will look at the current evidence-based policy requirements and how they might be extended to help when designing and implementing machine learning to

reduce administrative burdens to build a system to measure the impact of these solutions on administrative burdens and the overall policy or program outcomes.

Administrative Burdens

Overview

In the United States, there are several ways that the government has enacted to provide services and benefits to individuals. Many of these are jointly administered by the federal government along with state and local governments. Some are administered solely by the federal government or by the state or local government. There are others where the governments, either federal or states, also include third parties in the implementation and administration of these services and benefits. Not simply the mechanism of administration, but the types of benefits and the goals of the programs are also quite varied. Some are direct benefits of providing cash assistance, either for housing, food, childcare, health insurance, or for a multitude of individual necessities for our modern lives. There have been politicians and researchers who have pointed to these services throughout the years as either fundamental to our lives or as the things which can add the most value and the most benefit by providing them where they are not available. However, these benefits and services do not just show up without effort. They are not automatically deposited into the bank accounts and purses of the people who need them most. Especially not for those working several low-paying jobs or with mental or physical impairments that do not allow them to work enough or at all. These government benefits are typically based on what the general public considers to be fundamentals, but because they are provided for through public expenditure, they arrive through bureaucratic processes which are impacted by the shape, structure, and policies of our government systems. Some of these processes actually make it more difficult for individuals to access them, but all processes require certain expenditures of time, energy, and other resources. The magnitude of these resources on different stages of benefits interaction can be categorized in different ways and will impact the individual's ability to access them.

For simplification, this work will focus on federally administered benefits and leave out those adopted and administered by state and local governments. Out of scope are also where the federal government provides grants to the states, and the state typically administers many or all of these program implementation steps based on guidance or requirements for the federal government. The reason behind this limitation is to focus on the interplay of federal government requirements between benefit policy and administration as well as federal requirements for

performance management, evidence-based policymaking and evaluations, and federal considerations of machine learning. In their seminal work on administrative burdens, Herd and Moynihan explore several federal and state-level programs based on the wide-ranging implementations, but also because they draw a clear line between federal policy initiatives being adopted by state and local governments (Herd & Moynihan, 2018). Additionally, research has also shown a direct correlation between federal requirements for performance management systems and the same or similar being adopted by state governments (Hatry, 2010; Heinrich, 2007; Moynihan, 2008). Therefore, by focusing this research on the federal government, I believe some lessons can be extracted for state government programs, but I will also form a basis for the expansion of this research to requirements and programs administered by states.

Based on the scope of this research, as well as building on existing administrative burden research, several core federal programs emerge as examples. These include Medicare, Social Security Disability Insurance (SSDI), Supplemental Security Income (SSI), Old-Age, Survivors, and Disability Insurance (OASDI), more commonly known simply as “Social Security,” and the Earned Income Tax Credit (EITC). Unlike its partner program, Medicaid, Medicare is administered directly by the federal government rather than mostly by the state governments. Medicare was created in 1965 as health insurance for Social Security Recipients and has undergone many changes in the decades since. General eligibility criteria are based on age (65 years old) and citizenship or lawful permanent resident status. Additionally, individuals can qualify if they are younger than 65 and also receive SSDI or Railroad Retirement Disability benefits or qualify under kidney renal disease criteria (Cohen et al., 2015). SSDI was also adopted in 1965 to provide supplemental income to individuals determined to have a long-term disability (more than one year) or which will result in death and prevents them from maintaining employment (O’Leary et al., 2015). SSDI is administered by the Social Security Administration (SSA) through central and local offices within states. However, these state offices must follow federal SSA rules and requirements, which cannot be adjusted by individual states. SSI, in comparison, is means-tested cash assistance for disabled children or adults and individuals older than 65 years (note: some individuals are eligible for SSI in addition to Social Security Benefits). Social Security benefits are more common, and eligibility is based on working years and income levels (especially related to withholding or payments based on those earning) and eligibility age. In addition certain immigration status requirements must be met, but Social Security is a universal federal program. And finally, the EITC was enacted in 1975 and updated through the decades by subsequent legislation. The EITC is administered through tax refunds

and filing by the IRS because eligibility is based on low-income levels for individuals with qualifying children and dependents. This is unique because a core eligibility requirement has been employed, and unemployed individuals have often been excluded from program eligibility based on a policy goal of supporting employment by providing assistance to low-income individuals with children (IRS, 2021). This is not an exhaustive list, but it is a sufficient list to serve as examples of existing programs to help explain the key concepts this research will cover. Despite these examples, this research will attempt to remain general rather than conforming to one or more specific programs. However, I will build on these examples when searching for existing data to help test and shape my research outcomes.

This section of the paper studies how these benefits are administered and the costs associated with finding, applying for, and receiving these benefits in our society. These programs are not free in terms that their detractors often claim. Instead, the participants and the would-be participants of these programs have several costs to bear to “benefit” from them. These costs require each individual to explicitly decide on the value and worth of these benefits in comparison to other factors in their lives. This impacts the program take-up and success of these benefit programs and, therefore, must be studied and tracked just like the costs of any other implementation and ongoing part of the program. The costs incurred by individuals may have a significant impact on the outcomes of the programs and underlying policies. This is why no measurement and monitoring of these programs is complete without including the costs and expenditures incurred by the citizens that interact with the programs.

In this chapter, I will explore these costs, which are called administrative burdens. I will explore where they arise, how they are impacted by interactions, where they are explicit, and where they are incidental. I will also explore the effect of these costs on individuals, especially individuals who are sometimes the most deserving and likely to benefit from the programs. I will also explore categorizations of the costs, breaking them down into three types of burden: learning costs, compliance costs, and physiological costs. Finally, in this chapter, I will talk about the possibilities of measuring these burdens in line with effective governing, the potential outcomes that arise through measurement, and potential ways to routinize the monitoring of these burdens through measurement.

Government Interactions and Administrative Burdens

The history of the United States federal government creating programs that provide services or benefits to individuals and thereby determining how to categorize them and how to administer them is nearly as old as the United States itself, beginning with the debate and creation of several programs to provide assistance to officers and enlisted soldiers of the revolutionary war (Jensen, 2003). Even these early programs were fraught with political debate and discourse in the early continental congress and subsequent legislature regarding what it meant to provide services or benefits and what it meant to provide these to a subset or portion of individuals based on specific criteria. For example, should the young country provide a universal pension to all those who fought in the revolutionary armed forces, or rather should it only go to a particular subset such as those injured to the point of not being able to find or

	Intra-organization	Extra-organization
Intra-organization	<p><u>1. Red-tape/Green Tape</u> (Organizational Behaviors - Internal government agency rules and norms.)</p> <p>Government encounters with other parts of the government.</p>	<p><u>2. Administrative Burdens</u> (Service, Client - when citizen seeks government service)</p> <p>Individuals or organizations encounter Bureaucratic functions.</p>
Extra-organization	<p><u>3. Government imposing on individuals or organizations</u> (Arrests, regulatory requirements, voter ID requirements - government action on entities)</p>	<p><u>4. Individuals interacting with other non-government actors</u> (about government programs or services) (positive and negative effects of informal networks, non-profits, and third-party entities)</p>

Table 2-1: Types of Government Interactions; adapted from (Heinrich, 2016; Katz, 1975)

maintain “gainful employment because of their injuries”? Additionally, early programs adopted in response to numerous individual petitions sought to provide subsistence benefits to revolutionary benefits to individuals who served a particular period of service and who were otherwise determined to be too poor to support themselves and did not otherwise have family able to provide for themselves. These criteria created divisions and adjudicatory requirements for the young administrative state in order to create processes to allow individuals to apply for benefits, determine benefit eligibility and administer to those deemed eligible. It also continued a debate in the new legislation regarding the purpose of these types of programs and how they would be used to expand the goals of the country as well as what they meant for the “national

character” and the values of the country (Jensen, 2003). The young legislature had an important choice to make, did they provide services and benefits that were somewhat universal, or did they create and administer specific criteria-based programs that were “needs-based” or were otherwise afforded to individuals based on particular value judgments or goals. For example, the needs-based program only available to veterans of the revolutionary war that were fully disabled or deemed to be destitute was limited so that it is not necessarily viewed as a benefit to veterans, but rather a country ensuring a safety-net of those veterans that are disadvantaged because of the nature or outcomes of their service. Additionally, similar land-grant programs for veterans were universal but also supported the expansionist goals of the fledging country, much like current-day mortgage tax breaks support a values-based goal of motivating homeownership over renting or unemployment insurance that requires “active job-seeking” or even drug testing as it supports values-based determinations of which individuals are “worthy” or “deserving” for these programs.

Notably, the constitution of the United States did not include any guarantees of services or benefits to citizens, but it did guarantee individuals’ right to petition their government for these things. Additionally, it empowered Congress with the ability to collect taxes and create programs to redistribute those taxes with little restrictions in response to the petitions from citizens. Researchers believe these factors, as well as the United States’ values and beliefs on individualism and other characteristics, have pushed our country towards a system of discretionary “entitlement” programs rather than broad universal programs and benefits (Jensen, 2003). When there are programs that are criteria-based, the government must create a process to adjudicate and administer based on the requirements. This, in turn, allows for those administration requirements to be imbued with similar values-based and goal-based policy and process factors. A program is seen as reinforcing, or rewarding behavior that is politically and culturally seen as positive can be made to be simple or easy to “claim,” such as a homeownership credit in taxes, whereas something more controversial, such as providing cash assistance to single, out of work individuals can be implemented with multiple requirements to ensure individuals are “actively seeking gainful employment,” are not spending tax dollars on the “wrong things such as cigarettes and junk food,” and are otherwise not defrauding the program or the “honest, hardworking tax payers who are providing for them” (Aarøe et al., 2021; Evelyn Z. Brodtkin, 1997; Herd & Moynihan, 2018). The administrative state, often in response to the complexity of program and political requirements, has a lot of leeway in how programs are constructed and administered. These choices can create difficult paths or easy paths to access the

programs. At the same time, citizens often also want our government services to be “efficient and effective” and have spent much political and resource efforts in recent decades providing requirements and tools toward these ends. The study of administrative burdens combines these two phenomena and attempts to create additional transparency about the frictions surrounding accessing public services and benefits and attempting to reconcile them transparently with our desire to also inflict democratic oversight and administrative procedures on government efforts (Burden et al., 2012; Herd & Moynihan, 2018).

Administrative burdens are the frictions experienced when individuals interact with the government because they are seeking a public benefit or program. These are most often thought of indirect, “cash assistance” government programs such as veterans’ pensions (Jensen, 2003) but can also be seen in many “hidden but direct” programs that often come in the form of social tax expenditures such as the Home Mortgage Interest Deductions or Earned Interest Tax Credit (Mettler, 2011). These are important categories because the visibility of the program can often be directly tied to whether or not the program is popular, whether it causes lower or higher levels of psychological costs, and how it is perceived by citizens and policymakers alike. Notably, the amount of social tax exemptions and other hidden direct and indirect support to citizens has steadily increased over the history of our government for a variety of reasons. Some of these include political realities of how these hidden programs are perceived by individuals – often as “earned entitlements” or even obscuring from many individuals the fact that a particular program or benefit is tied to government policy or that it exists at all (Mettler, 2011). However, not all hidden or “submerged” government programs produce administrative burdens because many of them do not interact with citizens directly but are instead government-private organization interactions which allows them to be hidden or submerged in the first place. Good examples are government subsidies and regulation of student loans or tax exemption for private health insurance, which become powerfully important for the private sector and therefore create strong lobbying and constituencies but are indirect interactions with the individuals who are able to apply for them and perhaps benefit from them or not. Indeed, some of these government-organizational interactions become so fraught that the government is forced to spend more time concerned about the desires of the private organization than the citizens that they are intended to benefit. A good example of this is the interaction between the government and private tax filing services that lobby strongly to keep their market of providing intermediary status between tax filers and the government even though many have pointed out the frictions and costs this results in on individual tax filers and the failed attempts to refine and simplify tax filing in the United

States because of these incentives (Mettler, 2011; Shybalkina, 2020; Wagner, 2013). In the following paragraphs, I want to distinguish administrative burdens from other types of government interactions and frictions. I will also examine how administrative burdens are generated, how they impact individuals, and how they impact policy and programs. And finally, I will explore the literature about them in-depth and take a normative view of their impact as it relates to public program administration, measurement, and evaluation.

Administrative burdens are not red tape (Box 1 of Table 2-1: Types of Government Interactions; adapted from (Heinrich, 2016; Katz, 1975)Table 2-1). Although it is similar – and sometimes these terms are used interchangeably – in this paper, I will be clear that they do not mean the same thing. Red tape is the phenomenon of frictions and costs associated with government-on-government interactions. These are internal frictions that add no value to the policy or program because, by definition, they are frictions that only induce inefficiencies and costs without adding anything sought-after or desired in the implementation or administration of the policy or programs (Bozeman, 1993). Red tape is experienced when a government agency seeking to adjudicate or process a benefit must comply with the rules and actions of another agency or other parts of the same agency, typically with no programming value in these rules (Bozeman, 1993; Moynihan & Herd, 2010).

Administrative burdens are also not experienced when the government is interacting with or on individuals or organizations (Box 3 of Table 2-1). This situation is easy to confuse because the same actors are part of the interaction. But instead of seeking something from the government, these are when the government imposes an interaction or a limitation on private entities (Burden et al., 2012; Heinrich, 2016). Examples of these range from legal requirements and law enforcement interactions to seeking permissions for certain privileges or activities. For example, police actions and regulatory requirements are government interactions with private entities. The individuals are not seeking a benefit or service but are instead having the public virtues and requirements of law and order imposed on their individuals or collective actions. Another example is seeking a driver's license. This is a privilege, not a benefit or service that is administered by laws and policies (and often administered by states rather than the federal government). The ability to drive is regulated by certain rules and requirements, and the act of applying for a license is to comply with those requirements before being granted the privilege. Individuals must seek the privilege and often must comply with similar learning, compliance, and physiological requirements.

The third type of interaction is when individuals and organizations are acting upon and with each other (Box 4 of Table 2-1). Because the government is not directly involved, these situations do not always raise the question of government requirements, policy outcomes, or costs associated with benefits. However, in the public administration sector, there are growing examples of government program implementation and administration by proxy through non-profits, contracting companies, and non-governmental organizations (Mettler, 2011). Even when these entities are not directly involved in benefits administration, they can have a significant impact on human services programs through their referral and associative powers (Burden et al., 2012; Heinrich, 2016; Mettler, 2011). For example, it may be either the requirement or the policy for homeless shelters to help individuals they are providing services to apply for other types of income maintenance, healthcare subsidies, housing assistance, etc. Other examples have been referred to as the growing “submerged state,” whereby most individuals no longer understand the relationship between these private entities and the government services and benefits as they often are implemented through direct incentives to the private entity or social tax expenditures to the entities in exchange for certain policy outcomes (Mettler, 2011). The process by which they inform and assist can have a significant impact on the individuals’ experiences applying for and receiving benefits or whether they ever even consider the programs. Typically, this will have a lasting impact on that individual’s perspective of the benefits program even if they never directly interact with the government entities (Jilke et al., 2018).

The fourth and final type of interaction is Administrative Burdens, and this is what I will focus on in this paper (Box 2 Table 2-1). This is where individuals and organizations directly interact with the government entity when seeking services or benefits (Burden et al., 2012; Heinrich, 2016). This is where there is a discretionary interaction, typically motivated by the seeking of assistance from the government in one form or another. Oftentimes these services and benefits are discretionary, which is why the interaction is voluntary (Herd & Moynihan, 2018). However, that is not to say that these services and benefits are not a matter of life or death for some individuals. Often, they have significant potential to impact individuals either when receiving them or not. The point is that the interaction is not mandatory or compulsory, which is why the frictions associated with the interaction become incredibly important as they shape the nature of the interaction. Additionally, these experiences shape how people view the programs, view themselves and their community, and how the policy or program is viewed, shaped, and administered.

It is difficult to understate how the voluntariness of these interactions underpins so much of the outcomes of the individuals involved as well as the programs themselves. This is important to study and understand, especially as research and government policies and programs look for measurements of efficiencies and effectiveness. Wrapping these evaluations around scientific methods, they examine the programs and make decisions on funding, continuation, and expansions of these programs. Programs are judged by measuring the impacts on the individuals targeted by the programs, both those who receive the benefits as well as those who would otherwise be eligible to receive them (Burden et al., 2012; Herd & Moynihan, 2018). It is no small statement to say that there should be no instance of program measurement of efficiency or effectiveness without looking at the impacts of these interactions and how they shape the outputs and outcomes.

<p>Learning Costs</p>	<p>Time and effort expended to learn about the program or service, ascertaining eligibility status, the nature of benefits, conditions that must be satisfied, and how to gain access</p>
<p>Compliance Costs</p>	<p>Provisions of information and documentation to demonstrate standing; financial costs to access services (such as fees, legal representation, travel costs); avoiding or responding to discretionary demands made by administrators</p>
<p>Psychological Costs</p>	<p>Stigma arising from applying for and participating in an unpopular program; loss of autonomy that comes from intrusive administrative supervision; frustration at dealing with learning and compliance costs, unjust or unnecessary procedures; stresses that arise from uncertainty about whether a citizen can negotiate processes and compliance costs</p>

Table 2-2: Components of Administrative Burdens from (Herd & Moynihan, 2018, p. 56)

Types of Costs

Learning Costs

Learning costs are the frictions experienced when working towards understanding or seeking to know about the existence of a program or policy. They are the frictions of understanding what the potential benefits are in a program, what the limitations are, and what the rules are. They come from trying to learn and understand eligibility requirements, especially as they relate to a particular individual, group, and circumstance. Learning costs are experienced by asking and trying to answer the question, “do I qualify for this program, and how could it

impact me?” Physiological costs can also impact learning costs as they can set the place of understanding and inquiry for individuals (Burden et al., 2012; Herd & Moynihan, 2018). If particular programs or program membership is viewed by society and individuals in a certain negative manner, then people may be unlikely to even consider that they may be eligible because of the negative stigma associated with it. Indeed, perceptions of programs can form set points where individuals never allow themselves to identify with “the type of person eligible” for a government benefit. In other circumstances, people may not even be aware that government programs exist, and they do not participate in social circles where these types of benefits are understood and spoken about. Or they may not even clearly be understood as government programs or benefits to begin with because of their nature and perceptions of them (Burden et al., 2012; Mettler, 2011).

There can be a range of learning costs in programs. On the high-end of learning, costs are programs in which individuals have to seek out the program's existence and eligibility requirements. Often these are small, state, or local government-administered programs. Or they are implemented through nonprofits or private organizations on behalf of the government. These can sometimes be niche programs where eligibility requirements are so selective that most people would not be eligible except for perhaps under one set of circumstances in their lives, and therefore there is not much sharing of these programs. Programs with high learning costs often devote little resources to outreach and facilitation, so either people just stumble on them or are reliant on referrals through social workers, specialty organizations, or devoted research options.

Low learning cost programs and policies are associated with programs that are high profile and well known. These typically receive significant media attention and public discourse and have high eligibility and enrollment rates. Perhaps there are also external factors that drive the marketing of these programs, such as federally back student loan programs. Additionally, low learning costs can either be achieved with simple eligibility and application requirements and understanding or automatic enrollment and application. One example is the earned income tax credit (EIC). While the overall history and rules surrounding the EIC can be complex, the existence of it and the application process is standardized and normalized for individuals who regularly file tax returns, use software or services to file, and can be associated with activities that were going to be completed for other purposes.

Compliance Costs

Compliance costs are the frictions associated with applying for and remaining compliant with program rules and requirements. Compliance costs are associated with making an

application for a program or benefits, especially important is not just the application but supporting information or verification processes associated with an application (Burden et al., 2012; Herd & Moynihan, 2018). If certain proof must be provided, and that proof must be of a certain type or process, which also is verified or obtained from a third source in a way that adds complexity, this increases the compliance costs. The mechanisms of application also control the level of compliance costs. This is more nuanced than just paper-based versus electronic applications. Depending on the target population, their resources, knowledge, and experience of an application type and process can cause increased or decreased compliance costs. Whereas one population in one situation can experience benefits and low compliance costs through online application processes, this can be experienced as high compliance costs by another population without a computer or internet access (Ali & Altaf, 2021; Chudnovsky & Peeters, 2021).

Beyond application processes, remaining eligible for benefits also impacts compliance costs. Where there are specific requirements or rules, the specific how benefits can be received and redeemed is a particular subset. An illustrative example is that of Supplemental Nutrition Assistance Program (SNAP) benefits which in some circumstances have to be redeemed at only certain stores, and even so must be used at certain checkout registers and then for only certain particular foods and products. These rules and procedures increase compliance costs for individuals receiving the benefits (Barnes, 2020). In some instances, states have adopted rules that make redemption easier by issuing SNAP benefits via debit cards which are easier for stores to redeem. In other circumstances, stores have found automated ways to sort eligible products from ineligible ones in a way that is nearly seamless for individuals and store clerks. Additionally, other states have reduced the reliance on lists of banned items, making it a simpler process and giving recipients more autonomy on how they spend their benefits. Each of these initiatives can be seen as lowering compliance costs for recipients of SNAP benefits.

In addition to redeeming benefits, remaining eligible can induce compliance costs (Moynihan et al., 2015). This can range from particular eligibility requirements such as a particular income range, that if someone begins to earn in excess, then their eligibility automatically disappears to the requirement to file subsequent proof, reapplications, or verification materials to provide evidence of their continued eligibility. Just like initial application processes, the manner in which these are constructed and implemented has direct impacts on compliance costs experienced by participants and has been shown to have a direct correlation on continued program participation and eligibility even for individuals and populations who are targeted, eligible program individuals. It is often not just the “what” is

required for these processes but also the “how” that has shown to have a significant impact on compliance costs.

Psychological Costs

Psychological costs are the frictions experienced because of the perspective, opinions, and mental models that people have formed around themselves, others, and society in terms of government programs or benefits. Simply put, a core foundation of citizenship and politics in the United States has been individualism versus government. The research is varied, but the normative beliefs, values, and opinions are even more pronounced historically and contemporary (Burden et al., 2012; Moynihan et al., 2015). People have certain perspectives and opinions surrounding government programs and benefits, especially social programs, safety-net programs, and assistance programs. These seep into their perspectives of our country, groups, and individuals. These also impact perspectives and opinions about ourselves and create friction in how they perceive and understand government programs. For example, some individuals associate such negative opinions about social programs or benefit programs that they cannot associate themselves mentally with someone who would be eligible to receive a certain benefit. This becomes a blocker for them even before they might get into understanding the program or the application and eligibility requirements. Therefore, the psychological costs are so high they don't even get into the learning costs.

In other instances, psychological costs impact how people interact with programs and frictions that are put in place between themselves and using the program benefits to their fullest extent. Compliance demands increase stress and have been shown to be associated with lower program participation (Bhargava & Manoli, 2015). Additionally, compliance demands are associated with increased stress which not only lowers participation rates but increases program dropout in particular studies. Psychological costs are increased when programs are more direct rather than indirect, as well as higher when programs are more visible than more “submerged” programs which have increased through the reliance on social tax expenditures rather than more direct payment programs in recent decades (Mettler, 2011). In some instances, this is because social tax expenditures which often are referred to as the “hidden welfare state,” such as EITC or Home Mortgage Tax Deduction, are seen as “worthy” types of assistance, rewarding individuals viewed as “positive contributors to society” because their actions align with values which are viewed positively. In other instances, this is because individuals have difficulty seeing social tax expenditures as government assistance because it is often the absence of taxation, which is difficult for individuals to equate with a direct cash assistance program. This is often seen as a

politically favorable feature by both conservatives and liberals in the United States. However, it has likely increased the psychological costs due to the stigma and shame of more visible, direct assistance programs (Mettler, 2011).

Administrative capital is defined as the knowledge and resources needed to navigate bureaucratic encounters and requirements for public programs. The increased administrative capital is associated with greater program take-up, the program continued enrollment, and the ability to navigate administrative burdens. Decreases in administrative capital are likewise associated with lower program take-up, higher levels of stress, and more program abandonment (Masood & Nisar, 2020). Some studies have shown how repeated program involvement and increased administrative burden navigation can increase administrative capital. But there is also increased administrative capital associated with higher resources, such as being able to engage assistance (either personal or professional) to navigate administrative burdens and cognitive resources (Christensen et al., 2020). The relationship between psychological costs and administrative capital is complex. In some instances, it seems that higher levels of administrative capital are correlated with lower levels of psychological costs (Masood & Nisar, 2020). But in other instances, there was no relationship between administrative capital and psychological cost. This may be because of the association of negative emotions and perspectives of being a benefits recipient and not correlated with a person's ability to navigate the benefits bureaucracy. Whereas in other instances, a lower ability to navigate the process increases stress and also increases the impact of negative associations of psychological costs, compounding each other. More research and attention are needed on the impacts of psychological costs. It may be the most important of the types of administrative burdens but may be the most difficult to accurately measure and monitor all of the follow-on impacts on variable levels of this type of burden (Christensen et al., 2020).

Intentional vs. Unintentional Burdens

How do administrative burdens arise? In some instances, they are completely unintentional from an overall policy and program design perspective (Burden et al., 2012). As policies are adopted, they are typically implemented by separate entities within the government. These entities have norms, rules, and cultures within their spheres. They also work from perspectives of administrative standards, which preach certain values such as service organizations and administrative offices, and they must balance competing priorities of service as well as efficiency and accuracy. Each of these implementation and administration goals argues for different procedures and rules.

Service is a perspective of providing benefits toward outcomes for individuals and groups in line with the proposed policy aims. In social services, these can range from the very specific issue of issuing tax credits via the IRS tax return filing and refund process to somewhat nebulous and varied goals such as integrating supporting individuals released from prison back into society. This could include specific services such as mental health counseling, job placement and training, support finding and affording housing, food, bills, commuting for some time, and follow-up legal and parole requirements. Even in this instance, success can be defined as not breaking the law again in a certain period of time or something more aspirational but more difficult to measure as “becoming a functional and contributing member of society.” In some circumstances, these policy goals can be defined as personal experiences and outcomes, and in others, they are outcomes to be experienced by communities or groups, perhaps even nations have taken all together. The point of understanding these differences is that the complexity of policy outcomes and goals can institute many different rules and requirements for a program. All well-intentioned to help administrators measure the programs against these goals. But often, these can establish costs on individuals and groups to participate in the programs.

On the other hand, many administrative burdens are established by processes and mechanisms to eliminate waste, fraud, and abuse in programs. These motivations are popular in the public sector but also not dissimilar from the private sector (Herd & Moynihan, 2018). In some ways, these motivations stem from wanting to ensure that citizens’ tax dollars are spent effectively and efficiently. These motivations also come from the same social and cultural aspects of what it means to receive government support versus individual responsibility. Some of these processes focusing on fraud and abuse in programs come from the same negative opinions of these programs, which induce psychological costs, and the psychological costs can also exacerbate these perspectives of benefits and programs. Therefore, administrative burdens can be thought of as manifestations of risk tolerance in a program.

Suppose people have little tolerance for the risk of someone who is not “eligible” or “deserving” receiving a benefit that they shouldn’t, then increasing eligibility, verification, and redemption requirements which induce greater administrative burdens, is a logical policy model. Requiring individuals to take on the burden of proving they are eligible, even to the point of setting this determination up as applicants not being eligible until they prove to the government that they are, is a very different policy adoption and implementation approach that would justify higher levels of the administrative burden as compared to one where an individual is treated as eligible until or unless information becomes available which posits otherwise. This can even be

driven by the legal standards of proof set up through administrative law and procedure standards. “More likely than not” is a lower standard that is “clear and convincing,” - and this can make an exceptional difference in the process and procedural requirements associated with a program.

Research has shown direct linkages between political ideology and administrative burdens for these exact reasons (Aarøe et al., 2021). Social programs are not adopted, implemented, and evaluated in a policy vacuum. Ideology and cultural determinations and motivations are persistent through the discussions of, the development of, and the implementation of policies and programs. These factors can give us an understanding and motivation behind the nature of intentional administrative burdens but also help us understand the context around unintentional administrative burdens too.

Therefore, I cannot say that administrative burdens are wholly negative or problematic. This is a normative judgment that is rooted in political and cultural assessments of government programs and policies, which despite some arguments for an entirely scientific approach, must always allow for a representational democracy to ascribe political and normative judgments to policies and programs. In some instances, research has shown a direct relationship between support for burdens and the perspective opinions on the “deservedness” of likely program participants. However, in each instance, support for burdens has decreased along with what was perceived to be higher levels of burdens. This means that there is a not-so-surprising belief that burdens are less justified for beneficiaries who are believed to be deserving (Aarøe et al., 2021). For example, in the United States, there is consistently high support for members of the military and veterans. This is associated with a low tolerance for administrative burdens in veterans’ programs such as the GI Bill and Veterans Administration programs such as health care and housing. For program participants of other benefits, especially those studies who were paroled from prison, levels of support for burdens were much higher. People believe that “less deserving” individuals can experience higher burdens in accessing public programs. However, there is a breaking point when the burden goes beyond what is thought of as “reasonable” regardless of the perceived worthiness of the participants. This means that administrative burdens cannot be supported if continually piled up on participants, but in some instances, little to no burdens are viewed as appropriate if the target beneficiaries are particularly well-favored or viewed as “deserving” of the public program. Interestingly in this study, education on the burden types, experiences, and impacts was associated with less support for administrative burdens for participants thought of as deserving and less deserving. This indicates that the idea of burdens may be disassociated from the actual understanding of burdens specifically, and

increased knowledge or education would be associated with decreased support for burdens across the board (Nicholson-Crotty et al., 2021).

Negative Impacts of Administrative Burdens

Impacts of administrative burdens are additive to frictions between individuals and groups and the programs and benefits themselves (Herd & Moynihan, 2018). The direct impacts are added complexity, resources required on the individuals seeking the benefits, and fewer engagements. Burdens are also distributive. They impact people differently based on their experiences, resources, and needs (Chudnovsky & Peeters, 2021). Additionally, burdens are administrative and political choices, and therefore discussions about burdens are inherently value-based and cannot be fully separated from political discourse and policymaking (Burden et al., 2012; Herd & Moynihan, 2018).

Lower Program Take-up

Impacts of higher levels of administrative burdens are fewer individuals receiving benefits than qualify for them. This means that some of the people who should be able to get benefits and programs are unable to. Whether this is by design or an unintended consequence of some other motivation, this benefit up-take aspect is important and must be explored further. Program take-up has been shown to increase as administrative burdens decrease. This correlation has been shown in both directions in research (Bhargava & Manoli, 2015; Daigneault & Macé, 2020; Herd et al., 2013).

Administrative Law Avoidance

Another impact of administrative burdens is that they are seemingly outside of the current process for transparent and representative policymaking in the United States. The United States federal government has several key components that aim to have a transparent and involved policymaking process. These include elected representatives who set the policy agendas, but also administrative procedures whereby government processes to adjudicate program eligibility and administer programs are shared with the public, receive feedback, address that feedback and make changes (Elias, 2016; Lubbers, 1997). Additionally, the federal government makes clear what the adjudication criteria are, the burdens for evidence on the government and individuals, as well as appeal criteria and procedures. These also lay the foundation for judicial review of procedures and decisions, allowing a judiciary approach to program oversights and input based on legal challenges and civil suits (Bernick, 2021). It has been shown that many administrative burdens are opaque to these processes because the frictions are not well defined, nor are

government agencies required to publish information about them (Connolly et al., 2021; Herd & Moynihan, 2018). This causes a separation between the actual costs of the program to individuals and disallows significant public knowledge of and discourse about these burdens, which have a significant impact on programs and individuals interacting with the programs.

Undermining Citizenship

When administrative burdens are intentionally applied in ways that are not transparent to citizens, it lowers the efficiency of our representative democracy because it disallows transparent understanding that is the basis for political and values-based opinions, debates, and representations. This has been a concern for submerged policies and programs, which are increasingly adopted and implemented through social tax expenditures (Mettler, 2011). But it is also a concern with the administrative process and procedural changes, which are not necessarily required to be posted for public review and comment through administrative procedures and are increasingly being used and looked at for the impacts they have on citizens as well as the programs and policy outcomes to which they are tied (Herd & Moynihan, 2018). The specific nature of the opacity of these government programs disallows citizens to fully understand the values-based decisions being made that directly impact them or other citizens. Additionally, this lack of program awareness cuts off the possibility of lobbying and feedback to elected representatives or the impact of citizens' voting decisions. Furthermore, political scientists have shown the changed understanding of citizens on these programs when social tax expenditures for homeowners, for example, are viewed as "earned rights," whereas Social Security Disability Insurance is considered a "supportive" welfare program for individuals being propped up by "more contributing members of society" (Evelyn Z. Brodtkin, 1997; Mettler, 2011).

Lower Program Efficacy

Another examination of impacts goes to evaluation policy, specifically to formative evaluation. As I have discussed, administrative burdens create more difficulty for individuals to identify programs and benefits which they may be eligible for. They make it more difficult to apply, comply and receive benefits (Herd et al., 2013; Herd & Moynihan, 2018). Therefore, administrative burden impacts can also be seen as barriers to the successful and full implementation of programs depending on the policy perspective and intent of the administrative burdens and programs (Ali & Altaf, 2021). If, for example, I am looking at the proportion of potentially eligible individuals enrolled in a program or even who apply for a program compared to applicants/enrollees, this can be used as a metric to measure the successful implementation (if

full or near-full enrolment of eligible individuals is actually a policy goal). Therefore, administrative burdens that lower the proportion of program take-up are inhibitors to successful implementation, and therefore they disallow summative evaluations of the outcomes of the programs. This means I cannot know whether the programs are successful or not at achieving their policy goals because there is not a successful enough implementation to allow these evaluations.

Disparate Impacts on Individuals

Some people have looked at this in terms of proportionate impacts too. There have been numerous studies looking at the proportionate difficulties in applying for and maintaining eligibility for government programs by those most vulnerable and in need (Chudnovsky & Peeters, 2021; Feldman et al., 2015). These studies have shown that the circumstances under which these individuals are in make it more difficult to comply with requirements for application and eligibility for government programs. If true, this means that administrative burdens disproportionately affect the most vulnerable and those many would argue would benefit the greatest extent from the programs. Therefore, I need to measure administrative burdens in proportion to individuals' circumstances as burdens will be experienced by different individuals differently.

Some persons will build up greater knowledge, experience, and assistance with government programs and administrative requirements. These are toolkits and skills which help them deal with and overcome administrative burdens. However, these require cognitive capacity, networks of support, and experience that are disproportionate in the citizenry (Alhadad, 2018; Christensen et al., 2020). A balance between skills and abilities as well as cognitive capacity in terms of navigating administrative requirements and overcoming administrative burdens is an important aspect when thinking about program design and implementation. Several examples can be drawn here. For instance, one of the largest federal government programs in the earned income tax credits redistributed nearly \$60 billion in 2020 year making it one of the largest programs by dollar amount (IRS, 2021). It impacts millions of people per year as well. Additionally, it has a relatively low administrative burden because prior knowledge of the program is not necessarily needed, as most eligibility can be determined through the normal course of filing a tax return for the year (Wojciech Kopczuk & Cristian Pop-Eleches, 2005). Additionally, there are several ways the government helps lower the learning cost by pushing information about eligibility requirements and benefits amounts based on easy-to-follow criteria in forums such as irs.gov. Non-government interactions also exist and reduce administrative

burdens because both nonprofit and for-profit organizations help educate the public as well as facilitate applications and eligibility determinations (Wagner, 2013). In the nonprofit sector, there are organizations such as Code for America helping to reduce the administrative burdens and reliance on for-profit vendors by making tax return filing as frictionless as possible and directly available to individuals through information technology solutions such as “Get Your Refund” (“GetYourRefund,” 2021). Even in the for-profit sector, a cottage industry of tax filing services has proliferated with a vested interest in increasing the amounts of tax refunds by reducing tax liabilities and increasing eligibility for programs such as the EITC. The motivations for these are often self-serving, such as the customer experience and satisfaction of a lower tax liability/increased refund check to services that make short-term loans of the refund amounts (which increase profits for larger amounts), and even for commercial establishments which offer tax filing services because it allows them to sell large ticket items to individuals based on their tax refunds (Wagner, 2013).

There is also research that shows that increased administrative burdens impact individuals at different levels because these burdens are typically relative to other factors in their circumstances and life (Chudnovsky & Peeters, 2021; Masood & Nisar, 2020). For example, someone with greater resources and education may be able to navigate higher levels of administrative burdens because they have knowledge and education resources to better understand, research, and attend to administrative requirements. They may also have access to additional resources such as hiring a professional such as a lawyer or advocate to help them navigate the application process and determine eligibility. Additionally, they may also have an added resource of time to work through requirements and comply with various compliance costs, such as traveling to appointments at an application center and working through application paperwork, and collecting supporting documentation. Someone with fewer resources because they are working more time, have multiple jobs, or perhaps added requirements of taking care of a large family (both their own children and perhaps eldercare) leave less time and resources to deal with these costs. There needs to be some measurement of these specific circumstances when considering administrative burdens to understand desperate impacts based on potential starting resources and experiences of different individuals potentially navigating the process (Chudnovsky & Peeters, 2021).

Public programs often do not adopt person-specific criteria for applications, eligibility determinations, and benefit receipt because they do not have the resources, nor do they have the mandate to tailor to individuals. Therefore, the strategy most likely to be employed to reduce

these administrative burdens without disparate impact (or at least to reduce the potential disparate impact) is to seek to reduce administrative burdens across the board. Looking at the lowest common denominator for administrative burdens which still comply with legal requirements and eligibility factors is the most likely way for the government to tackle this problem. This is a reversal from thinking about adoption and implantation based on the most risk-avoidance posture, whereby only those willing and able to meet high bar eligibility requirements can show they are worthy of the benefits or keep the benefits. These approaches are much more likely to have higher disparate impacts, which are more difficult to measure because they are more likely to result in lower take-up of programs, often for those individuals most likely to benefit from the proposed policies and programs.

Impacts on Policy and Program Outcomes

Outcomes in public administration are such an important concept. It comes from evaluation theory, and the concept is that outcomes are the policy goals. They are the changes expected to result from the tangible outputs of a program (Herd & Moynihan, 2018). For example, the output of the EITC program is the issuance or non-issuance of a license based on the inputs and activities of the applications, the tests, the adjudications, the monitoring of eligibility requirements, and the enforcement. The outcomes are the expected result. Outcomes are difficult to measure for reasons which I will get into in the next chapter. What is of interest to us in terms of administrative burdens is how they impact outcomes and our ability to measure outcomes in a program or policy.

As already discussed, administrative burdens likely impact benefit recipients and potential recipients in disparate ways (Chudnovsky & Peeters, 2021). Often these take the form of excluding or lowering the levels of take-up by individuals with less cognitive, financial, and administrative experience resources more so than other individuals who have greater resources in these areas. As explored previously in many programs, this likely means that the individuals most likely to benefit from the programs are negatively impacted more so than others (Battaglio et al., 2019; Christensen et al., 2020). These may be the very individuals for which the program or policy was designed, but for intended and unintended reasons, policy adoption choices, and implementation decisions, they are precluded or excluded from the programs at a higher rate than individuals who may not benefit from the programs as much.

As discussed previously, this may be for reasons in which a higher administrative burden barrier is enacted in order to act as a political and program filter, with the argument that only those truly needed will invest the resources necessary to navigate the administrative burdens.

The theory is that those individuals who are not truly needy will abandon the application process upon encountering the administrative burdens and performing a benefits/cost assessment of their continued pursuit (Herd et al., 2013; Herd & Moynihan, 2018). Additionally, these same thought patterns are applied to administrative burdens to exclude and root out fraud. In these instances, the political or administrative perspective is one of low-risk tolerance for anyone receiving a benefit for which they are not eligible (Aarøe et al., 2021; Abelkop, 2010). This means that in many instances, their focus, and calculus, is to institute requirements that allow for the validation of eligibility to such an extent that they become burdens on individuals. They may or may not be aware that this increases the likelihood that potentially eligible individuals have a more difficult time applying for and receiving benefits, but this is often ancillary to the motivation and goals of the administrative burdens. In many of these programs, the standard of proof of eligibility is higher, and the burden of proof is both higher and on the individuals to show they qualify rather than the burden being on the government to show someone does not qualify (Abelkop, 2010). These are important design and implementation choices which directly impact the levels of administrative burdens throughout the policy process.

The reason to revisit these issues is that when measuring outcomes of a policy program, we need to consider not just the intent of the program but the way in which the program is designed and constructed to understand the effectiveness or ability of us to accurately measure the intended outcomes (A. Kroll & Moynihan, 2017). As has long been known in the evaluation community, an unsuccessful implementation negates the ability to evaluate the outcomes. This is because the program has not actually been implemented in a way that is successful and functioning, and therefore the correlations between activities and outcomes cannot be controlled for, meaning there is no chance of causal conclusions. For example, if SSDI was not actually issuing benefits to a majority of individuals meeting the “disabled” definitions and other eligibility requirements of the policy, then it would not be accurate to attempt to evaluate the impact of SSDI on its policy goals of providing cash assistance to disabled individuals to ensure they are able to procure minimum living standards

Measuring Administrative Burdens

If administrative burdens are real phenomena, as I propose that they are, then they must be able to be measured. This will be difficult because they are subjective in the sense that they are experienced differently by differently situated individuals. And they matter for the design, adoption, implementation, and evaluation of policies and programs. They must be measured first to prove their existence in particular circumstances and measured in the magnitude of particular

situations and in relation to other variables. The first is grounded in the empirical study of administrative burdens, a foundation of the theoretical experience of researchers, administrators, and citizens. In the second instance, being able to measure their magnitude, especially in relation to other variables, allows us to better understand their correlation to other factors in the world, in individuals' lives, and in the implementation of government programs. This allows us to better understand the causes of administrative burdens and the levers to increase and decrease them. This firmly grounds them in public administration and gives policymakers, administrators, and researchers the ability to be cognizant of them and to be explicit about their uses and efforts to increase or decrease them. It also gives us the ability to measure the impact of administrative burdens on other independent variables. Most notably, this allows evaluators and evidence-based policymakers to understand the impact and interplay of administrative burdens on policy and program outcomes and goals. It also allows the testing of hypotheses of ways to increase or decrease administrative burdens.

Therefore, it is clear there are many benefits to measuring administrative burdens. Now I come to the question of how to measure administrative burdens. There are several theories on creating indexes of measurement for experienced phenomena, from the rudimentary opinion surveys to the very technical and resource-intensive mixed-methods of observation, interview, and inquiry. The path I will follow in this work, though, is to latch on to frameworks and practices already established and resources in the government in an attempt to show how existing resources, skills, and experience can be leveraged to include the understanding and control of administrative burdens. In this way, I hope to make these methods and exploration more obtainable and more realistic to conjure change in how administrative burdens are observed, understood, and used in government. Therefore, I want to explore and show how existing performance management requirements and infrastructure can be used to identify and measure administrative burdens. As I will talk about in the next section, this is important because there is growing and sustained pressure and experience using performance management techniques in the government. Therefore, instead of creating new, resourced requirements for government administrators, I want to provide a path to simply extend and expand what many are already doing pretty well to include an understanding and measurement of administrative burdens in policy and programs.

Performance Management

Overview

For the past four decades, the United States has experienced an increase in the use of performance management theories and systems in the United States federal government (Ames, 2015; A. Kroll & Moynihan, 2017; Moynihan, 2008). This has come about in multiple iterations and through external and internal mechanisms. While the focus of performance management proponents has been on the potential benefits of performance management and the use of performance information, recent work has called into question the efficacy as currently adopted in the federal government (Hatry, 2002; Moynihan, 2008). These studies often showed that performance management had not worked well on a system level for policy in the United States federal government, but there have been benefits at the agency and sub-agency levels (Moynihan, 2008). Many researchers and proponents of performance management have focused on the purposeful and effective use of performance measurement in policy decisions, especially in the legislature as congress debates the adoption of programs or in the use of budgeting for existing programs (A. Kroll, 2015a, 2015b; Moynihan, 2008). The most well-known efforts to adopt performance measurement systems may not yet show success at the government level. There are pockets of progress for performance information use at the sub-agency level, where arguably, the information being produced is more closely managed with the aims of the office, and the feedback loops are tighter, hopefully leading to more effective adoption.

Despite near-universal support for the ideas of performance management, there have been diverse perspectives on how to achieve it through the past decades in the United States. Several notable attempts have been made in the federal government, including the Government Performance and Results Act (GPRA), the Program Assessment Rating Tool (PART), and the more recent GPRA Modernization Act (GPRAMA). These formal, top-down federal programs are augmented by numerous bottom-up attempts at focusing on data-driven reviews and understanding of government programs. Notably, the city and state level “CompStat,” “CityStat,” and numerous other “Stat” operations promulgated over the past decades offer mounting examples and evidence of the approaches that can be taken for organizations and programs (Dooren et al., 2015; A. Kroll & Moynihan, 2020). They have also provided fertile ground for researchers to identify what works and what doesn’t work in performance management.

Another issue for performance management is the notable lack of data tracking and measuring the “success” of performance management. This irony is not lost on some researchers, but it does also play into the legislative and adoption story when they look at the motivations for, reviews of, and changes made to different performance management programs. But there are several bright spots in the performance management movement, which also beg for optimism that the performance management framework is able to add significant value to government decision-making for both policy and operational impacts.

There is a long history of government performance and efficiency improvement. This began early in the nineteenth century and was pushed by those adopting the tools and theories of “Taylorism” or the scientific management movement, which was aimed at finding the most efficient processes and means of production in the industrial area, but then quickly spread to knowledge economies and to the public sector (Crowley & Scott, 2017; Dooren et al., 2015). While there are many criticisms about scientific management in the private and public sector, a common theme that has lived on is the idea of monitoring processes, inputs, and outputs to identify areas of concern, waste, inefficiencies, and even effectiveness of programs. These ideals became pronounced through subsequent focuses in the United States federal government system, and the entire process sprung up to try to give policymakers and administrators new tools to be more effective in their jobs (Ames, 2015).

It is important to note here that while there have been constant iterations on the themes of performance measurement and management in the government, and each successive Congress and administration has branded a certain focus or schema on these efforts – there is little research to show that these efforts have been particularly successful and even the measure of success are somewhat vague according to legislation and administrations guidance (Dooren et al., 2015; Moynihan & Kroll, 2016). However, it is clear based on the continued rhetoric and focus that these ideas are not going to go away any time soon without a clear focus and shifting measurement and approach to government “efficiency and effectiveness,” which have been core values and goals under the government in the United States and aspirational goals of many liberal democracies in post-modern societies. Whether the efficiency supports the ideals of the sacredness of duty to be good stewards of citizens' tax monies or the effectiveness is driven by the desire to effect significant outcomes through the policies and programs, there are many political reasons to support these ideas. However, it is also important to recall the fact that these are not necessarily built on the foundations of “logical action” since there is no strong evidence that these efforts result in these goals. As pointed out by March and Olsen (1984), though, these

values have been learned and ingrained in the United States as facts derived from the “logic of appropriateness” since these ideals seem justified to politicians and citizens alike despite their specific political values and leanings (March & Olsen, 1984, 2008). The cause of whether these public administration initiatives are simply “appropriate” or “logical” is out of scope for this research since these are the current realities of administrators and applied researchers. The United States has these learned values and the requirements in the Federal government – and there are examples of instances where they can be useful and effective in achieving specific initiatives and outcomes. Therefore, this work will explore where these initiatives and requirements came from, how they are currently implemented in the United States federal government, and how they can be leveraged to approach the phenomena of administrative burdens.

The heart of performance management is to set strategic goals. Strategic goals differ from goals in the way they are supposed to be tied to outcomes or behavior changes in the policy of the program. These are not meant to be merely numerical output targets, such as “produce x number of...”. Instead, these goals should be to effect a change based on the outcomes. Therefore, the goal for x becomes, “by the end of the calendar year, increase satisfaction with USPS service delivery by 25 percent, through reducing the number of late and missing mail items by 40 percent” (Radin, 2003).

Once these strategic goals are set, the organization must collect performance information or performance data to monitor its progress towards these goals (Moynihan, 2008). In our example, this would require performance data about customer satisfaction, data on mail delivery times (including forecasted delivery targets and measurements of when they are met and when they are late), data on missing mail pieces, and obvious data on time and how the above data elements changed over time for the comparison to the baseline or starting point for the strategic goal.

Having strategic goals and performance information are the practical necessities for performance management, but as more and more researchers, legislators, and administrators are pointing out, the actual use of these things is a key component of performance management. There are many components to this use which will be reviewed later, but a basic organizational structure and leadership attention and focus on the goals and the progress towards them as organizational priorities cannot be understated as a necessary condition.

Definitions

Performance management has been defined in a number of ways over the years, just as the desired goals of performance management have taken on different. In this section, I will examine several iterations, discuss their strengths and weaknesses, and then end by describing the current operationalized definition in the federal government in the United States and also describe the parts which are supported by current research and theory.

A simple definition is that performance management is the use of performance information by decision-makers in the government (Dooren et al., 2015). The keyword in this definition is “use” since there have been many examples of the creation and tracking of performance information and performance data that were not incorporated into decision-making or shared with leaders who could have folded it into their analyses and decisions (Johansen et al., 2018). The promise of performance management is to set strategic goals for a program or process and then design it in such a way that performance information can be collected, which allows you to track progress toward those goals. Performance management is a system of generating performance information through strategic planning and performance measurement routines and then connecting this information to decisions (Moynihan, 2008).

The basic conceptualization of performance management is two-fold. First, administrators need to set strategic goals and then create and collect data to monitor their status against those goals. This is designed to pull the attention of administrators away from simply dealing with the inputs and reacting to the external drivers of their agency and mission. Secondly, the administrative leadership and managers are supposed to have increased flexibility and autonomy to organize and control their organization to best achieve their strategic goals. The idea is to diffuse control and plan to a work unit to better allow managers to change processes and procedures to more nimbly and successfully meet their goals (Gao, 2015; Hassan & Hatmaker, 2015; Moynihan, 2008; Moynihan & Lavertu, 2012). With these two components, strategic goals are clearly identified, and performance information measuring progress towards those goals is collected. And agency administrators are given the resources, autonomy, and empowerment necessary to manage their organizations towards those goals based on the feedback of the performance information.

The idea behind this is that agencies create real-time feedback on how their efforts and resources are being expended toward the goals which have been set. This sounds fairly rudimentary, but the process has not always been deployed. In many instances, especially in the public sector, performance information and performance measurement have not been part of the

program design (Mihm, 2017). Goals were set, budgets allocated, and policies enacted without a process by which to judge how effective they were. Even with an increasing focus on efficiencies, there has not always been a focus on the measurement of inputs and outputs to determine efficiencies. This has been changing over the past decades, though. More and more, the federal government has focused on creating the infrastructure and culture of performance measurement in programs (Bourdeaux, 2008; Cavalluzzo & Ittner, 2003; Dooren et al., 2015). This is often supported by policymakers regardless of their ideological position because the values of efficiency, accountability, and transparency are supported – even if for different goals (Bourdeaux, 2008; Christensen et al., 2020).

Even though more agencies have instituted performance information and measurement processes, many are quick to point out that it is not the mere presence of these systems which matters. It is the degree to which they are used by organizations to guide management and policy (Gao, 2015; A. Kroll, 2015; Moynihan & Lavertu, 2012). In recent decades there has been a growing focus on understanding what leads to the effective use of performance measurement and information, trying to understand what conditions are necessary and sufficient for the successful design, implementation, and use of these processes to help manage government programs (Gao, 2015). In several studies, it was shown that performance management requirements resulted in siloed development and tracking of performance data, but only in a compliance posture. These became elaborate and costly box-checking exercises which did not have the ability to impact leadership decisions and program outputs (Gao, 2015; A. Kroll & Moynihan, 2015).

The two necessary criteria for performance management are the autonomy of public sector administrators and leaders from a decentralized control structure and clear strategic goals and performance information measuring the status towards those goals. These two tools combined are believed to provide necessary and sufficient tools for public managers to use their leadership and expertise to achieve those goals. However, while there has been significant focus on the second condition, there has been relatively little movement toward providing public managers with the first necessary condition (Destler, 2016; A. Kroll & Moynihan, 2015; Moynihan, 2008). Without increased authority over budgets and personnel resources, public managers do not have many tools to use their own knowledge to directly impact performance. And therefore, the accountability that is proposed in the performance management doctrine breaks down, and expected benefits need to be tempered based on what is realistically within the control of administrators (Bourdeaux, 2008; Dey et al., 2015).

Why Performance Management is Important

From much of the literature, it is easy to understand the importance or believe the importance of performance management. As discussed previously, performance management is a further attempt to make government more effective, efficient, and accountable to the oversight representatives as well as transparently accountable to the public. These goals and ideals seem intuitive to most of us living in the United States and many developed democracies. However, this has not always been the stated goal of government programs, but there has been a growing belief that it should be so during the twenty-first century (Andersen et al., 2016). Some believe this grew with the growth of government programs, especially the “welfare state” since the FDR era. With the increase in government programs, authority, and expenditures, there has been a growing emphasis on “waste” and “abuse” of the government authority to collect monies and redistribute them through different programs. This has created motivations for systems such as performance management from the political-ideological spectrum (Cavalluzzo & Ittner, 2003). The conservative ideology, which emphasizes smaller government, fewer taxes, and therefore less direct support programs, wants to understand how tax monies are being spent, be aware of “waste, fraud, and abuse,” and welcomes the oversight of performance management when arguing for smaller programs or the cessation of particular programs based on poor performance or effectiveness (A. Kroll & Moynihan, 2020). Liberals, on the other hand, also want to have accurate performance information to potentially justify these programs, even to show their value to argue for increasing them. Additionally, to defend against criticism as well as to ensure that outcomes of programs are being achieved, using performance management can help identify areas for corrections, changes, or need increases in funding to meet stated goals (Moynihan, 2008).

While the above goals explain the support for performance management, it is also important to understand the theory of performance management. Simply put, the underlying belief behind performance management is understanding and articulating clear strategic goals, measuring progress against those goals, and requiring reporting and transparency to be held accountable to the goals (Andersen et al., 2016). Additionally, the ideas behind giving agency administrators autonomy and resources to allow them to manage their organization to meet those goals are intuitively logical theories that support the use of performance management (Hassan & Hatmaker, 2015). As I will talk about in later sections, it may be that these ideas and ideals are false assumptions or not actually implemented in practice which leads to a mixed assessment of the success of performance management. However, the ideas behind the implementation of

performance management are important and have become commonplace in our public institutions, just as they have in private organizations and even in our daily lives. The United States government has become a more measurement-driven and data-intensive society and our approach to the world (Hatry, 2013). This can be seen in many different manifestations, from our desire for more information about ourselves through personal fitness and health trackers to our increased reliance and desire to see more statistics and data about our pastimes, such as the prevalence of sports statistics. The United States is increasingly comfortable and desirous of quantification of public programs. This is not always possible, successful, or especially appropriate - the desire is still real, and so the ideas behind performance management in the public sector should not surprise us at all.

History of Performance Management in the U.S. Federal Government

The history of performance measurement and performance management can be traced as far back as the 1800s (Moynihan, 2008). However, as it is conceptualized today, the most common are instances of different performance movements in the 1960s and 1970s, with the most recent iteration beginning in the early 1990s (Kroll, 2015a). Performance management has its roots in the private sector and was adopted from the early scientific management movements. Scientific management believes that you can measure everything, from inputs, processes, procedures, and outputs (Moynihan, 2008). By measuring everything and applying analyses to find inefficient parts of processes and identify new procedures, companies could maximize profits by better utilizing their resources and eliminating waste. The public sector's adoption of performance management in some instances is not much different from many other instances of private sector processes and theories making their way into the public sector. Often this came from management consultants, appointed leadership who had private sector experience and preferences, and politicians looking to bring private sector "efficiency" to the public sector. For performance management, however, this may have been more overt. This is because, at its fundamental level, the movement promises efficiency, effectiveness, and transparent oversight. These have long been staples of what politicians have simultaneously demanded from the federal government while also extolling the fundamental lack of it in the federal government system (Kroll, 2015a; Moynihan, 2008).

From some political perspectives, the performance management movement is a way to justify the expense of tax dollars collected from citizens. In other instances, it has been a mechanism to identify and ensure effective delivery of programs or implementation of policies that are important to them politically or for their policy agendas. It has even been argued that

increased use of performance information in decision-making leads to lower employee attrition rates (Lee & Jimenez, 2011). Whatever the basis, performance management can be thought about in terms of measuring the input and outputs of particular processes. In the public sector, this takes on the additional definition of performance information associated with outputs and outcomes of the intended policy, measuring whether they are in line with the ultimate policy goals to provide an assessment of effectiveness in most cases but also efficiencies. In the public sector, it's important to focus on the achievements and results rather than just the financial aspects. Another important characteristic of performance information is that it is quantitative and made transparent through reports, databases, and the objectivity of standards in line with publicly stated goals. The collection of what is defined as performance information is intentional, following a designed system of information creation, collection, analysis, and reporting. This is significant in that it is different than simple collection and analysis of data, which is ad-hoc and not purposefully designed to measure a program based on a logic model or intended causal connection (A. Kroll, 2015b; A. Kroll & Moynihan, 2017; Moynihan & Kroll, 2016).

There have been high-level attempts to adopt performance measurement in the federal government in the United States. There have also been many agency-level initiatives, both from leadership down and from program management staff up, which emulate the performance measurement movement. Additionally, the United States has seen the expansion of performance measurement research, which feeds into think tanks, best practices, and professional organizations. All these forces continue to focus on federal government leaders to the proposed benefits of performance measurement. However, there also has been more recent work aimed at identifying the success of performance measurement implementation.

New Public Management

Performance management in the United States, and in many developed countries, has close ties to the theory and practice of New Public Management (NPM) (Hatry, 2002). In NPM theory, there is a focus on the idea of running government programs and organizations more like businesses which gained popularity in the United States in the 1980s. This includes thinking about citizens like “customers” and the adoption of performance management systems to increase productivity, efficiency, and management for performance. It includes some similar assumptions, such as managers having clear goals and measurement of outcomes against these goals. Managers are given autonomy in the acquisition and use of resources. The authority of operations is devolved to lower levels of public administrators. Decisions and oversight are

focused on outcomes and outputs rather than inputs and procedures. And managers are held accountable for the use of resources aligned with strategic goals and outputs, and outcomes. In some NPM systems, there is even a pay-for-performance system to reward public sector managers based on their successes toward performance goals. NPM arguments are central to performance management doctrine, increased authority and strategic focus on outputs and outcomes will allow for public administrators to manage results and be held accountable for these results against strategic goals (Moynihan, 2008). This theoretical move away from centralized control and focus on input and procedures is theoretically a necessary and sufficient formula for change in government administration.

Reinventing Government - National Performance Review (NPR)

Many believe that these measures were an outgrowth of the financial troubles of the 1970s and 1980s, which caused a renewed focus on increasing efficiencies and reducing waste. Similarly, the 1980s experienced the growth of Total Quality Management, which aspired to create performance metrics and goals for nearly every process within an organization to quantify inputs, expenditures, and profits. Like many initiatives begun in the private sector, the public sector attempted to emulate and adopt these principles as well. “Profit” control can be viewed in terms of effective expenditure of tax revenue. Additionally, even policymakers that wish to increase public programs and expenditures often find it more politically feasible to find resources that can be reallocated from existing programs rather than justifying and obtaining a new budget (Hassan & Hatmaker, 2015). Therefore, efficiency gains can create capital for new or expanded programs without the hassle of congressional assistance and work.

When President Clinton took office, a signature program was the National Performance Review (NPR) - which later became the National Partnership for Reinventing Government - which aspired to provide means to hold the executive branch accountable to the citizens by providing transparent means to measure the success of programs and find efficiencies. The reinventing government system borrowed heavily from the principles and techniques of the private sector’s use of Total Quality Management. These efforts were focused on how the government performed rather than what the government focused on. Of note, many of the goals of the NPR coincided with the use of automation, computer processing, and reduction of “red tape,” putting citizens [customers} first, and empowering government employees to “get results” (Clark, 2013; Osborne et al., 1992).

President Clinton empowered Vice President Gore to lead this signature effort, which resulted in plans for changes in size, structure, and mechanisms of performance. Often Congress

rejected several key parts of the plan, including rejecting certain budgetary and program recommendations to shrink size and costs (Moynihan, 2008). However, one of the legacies of the reinventing government initiatives was the idea that performance matters to government agencies and that continual monitoring of and striving for more efficiency and effectiveness should be a foundational goal for the executive branch. This seems somewhat parsimonious now, but it was not always believed that efficiency was the main goal of the government where programs and jobs are created through political means, much unlike the private sector (Gore, 1997). One of the main efforts under NPR was the effort to introduce performance budgeting which integrates financial information and performance information to increase data available to Congress about the performance of programs and agencies under the assumption that performance data would result in more focus on effectiveness and efficiency (Moynihan, 2008).

GPRA

As a direct off-shoot from the focus on NPR, the Government Performance and Results Act (GPRA) was passed in 1993, but not all sections were implemented until 1999. The intent of the GPRA was to create strategic goals and plans for each federal agency that allowed congress and the public to more easily measure the success of programs and agencies against these goals. The idea was to allow more oversight, which would also make it easier to modify, reward, or cancel public programs which were not successful. The idea behind this legislation was built on NPR to measure success through objective means rather than relying on political mechanisms for oversight based on values and anecdotes (Radin, 2003; Zamora & McNeil, 2012).

Through these mechanisms and oversight, the goal was to allow for feedback, improvement, and cessation or modification of programs and policies. This is a formalized approach to the idea that the public sector is supposed to be providing a certain value for the expenditures (which are mainly tax monies or fees specifically collected for services performed). There have long been calls to provide more accounting for tax expenditures and to maintain a level of rigor that ensured that value was being produced through taxes and that inefficient, ineffective, and fraudulent programs were rooted out (*Government Performance Results Act of 1993*, n.d.).

The GPRA required federal agencies to set strategic goals, measure progress against those goals, and report this progress to Congress (primarily during the Federal Fiscal Years cycle). It had three overarching mechanisms: to establish strategic plans for each federal agency (five-year plans), to establish annual performance plans for their strategic goals outlining how they will achieve their goals, and the performance measures collected to measure their success

towards those goals and to create annual reports of the success or failure against their annual goals based on their performance measures (*Government Performance Results Act of 1993*, n.d.). GPRA also required OMB to promulgate guidance for agencies on how their strategic plans must be linked to their performance information, as well as how program evaluations should support, refine, and direct agency and program goal changes. A lot of elements of the GPRA existed previously in siloes and different theories, but many see GPRA as the manifestation of long years of focus and the belief that performance information and strategic goal setting were key criteria for government focusing on results and performance (Moynihan, 2008; Radin, 2003).

PART

Assessments of GPRA were mixed, but President George W. Bush characterized it as a failure. In 2004 President Bush implemented the Performance Assessment Rating Tool (PART) through OMB direction as a mechanism to augment existing efforts. In particular, while it was argued that PART supplemented GPRA, in many instances, the executive branch prioritized the focus on PART over GPRA, even though this was criticized by those in Congress who had little input into PART since it was an executive function and not tied to the budget and oversight functions of congress (Moynihan & Lavertu, 2012; *The Bush Administration's Program Assessment Rating Tool (PART)*, n.d.).

PART assessed government programs based on purpose and design, strategic planning, program management, and program results. The output of the PART assessment was an overall Likert scale rating for a program that ranged from effective, moderately effective, adequate, or ineffective. While touted as an objective assessment regime for government programs, the program was viewed as highly partisan by lawmakers in Congress, who mostly ignored the assessments. Additionally, even government administrators reported experiences and perspectives that PART was run as a partisan instrument by the OMB, in particular, to focus on creating evidence and justifications to curtail or eliminate government programs that were not aligned to the administration's values and priorities rather than an objective feedback mechanism aimed at creating efficiency and effective government services and programs (Moynihan, 2008).

Following GPRA, and in response to several areas of concern and criticism, federal agencies were also required to comply with the Program Assessment Rating Tool (PART) as instituted in 2002 by the George W. Bush administration and overseen by the Office of Management and Budget (OMB). PART worked alongside the requirements of GPRA and

required agencies to answer assessment questions about their programs for purposes of budget justifications. The areas of the assessment covered program design, strategic planning, program management, and program results. Congress did not receive PART assessments well and often ignored them for purposes of budget assessments as they did not fit into the congressional model or they were viewed as partisan attempts to change policy (Moynihan, 2008; United States Government Accountability Office, 2005). President Obama quickly did away with PART early in his administration.

GPRAMA

Congress passed the GPRA Modernization Act (GPRMA) in 2010, and it was signed into law in 2011 by President Obama. The GPRAMA is an attempt by Congress and the executive branch to build on and correct problems with GPRA. Far from calling GPRA a failure, the government is attempting to focus on the successful parts of the performance management process and iterate improvements that are thought to impede the successful use of performance information for government decision-making. GPRAMA has several key changes which align with interesting features and research of performance management to correct missteps and failings of the GPRA as well as reestablish congressional control over performance management guidance from the executive branch as it was seen to be partisan under PART. Specifically, to increase transparency and public interaction, the GPRAMA requires agencies to publish their reports and strategic plans online in machine-readable formats. To address concerns about government leadership attention and use of performance information, it also created an emphasis on smaller amounts of higher-level goals setting, as well as identifying specific interagency priority goals. The intent behind this is to limit focus, energy, and attention on fewer goals to allow agency leadership to sustain focus and be more successful in making program improvements. This is in direct contrast to the earlier GPRA and PART emphasis on strategic plans and goals for every program, which easily overwhelmed staff and leadership focus and bandwidth (Moynihan, 2008; Moynihan & Kroll, 2016). The GPRAMA also tried to align goals and changes in strategic goals to better align with the change in presidential administrations in order to align with the cycle of the elected political leadership of the executive branch. Additionally, GPRAMA attempted to better establish measurement and performance information into goal monitoring and created annual reviews by OMB for specific follow-up, especially when progress was not being made (Kroll & Moynihan, 2015; Moynihan & Kroll, 2016).

GPRAMA realigned the schedule of required strategic goals to closely align with presidential terms to acknowledge that while objective performance data is important, the strategic goals and focus of decision-makers in our government is by design a political process, and therefore the performance management process must not only acknowledge but facilitate this reality to ensure political and agency leadership could more easily leverage performance management rather than ignoring it because it didn't fit, or impeded, their direction and control ("Implementation of GPRA Modernization Act Has Yielded Mixed Progress in Addressing Pressing Governance Challenges," 2015; U.S. Government Accountability Office, 2017).

Unlike GPRA and especially PART, there was an acknowledgment that leadership has limited ability to focus on government programs and that to be effectively employed programs, the performance information about them must be prioritized. Therefore, GPRAMA created a limited amount of Agency Priority Goals (APGs), which are aligned to agency strategic plans. These limited but high-level of APGs attempt to ensure that agency leaders can focus on and directly engage in these goals rather than being lost and overwhelmed by goals for every single program within their purview. Additionally, GPRAMA acknowledges the interrelated nature of many important policies and programs that are dependent on multiple agencies for execution. Therefore, Cross-agency Priority (CAP) goals are set in limited numbers to focus multiple agency leadership on these priority areas and create an OMB-convened group or process to review and focus on them on a regular basis. These CAP goals and their reviews require that the performance information collected to monitor progress has an outline process to focus agency and OMB leadership on progress in ways that encourage the use of the performance data for decision making to meet milestones and progress toward the successful attainment of the goals (Moynihan & Kroll, 2016; Zamora & McNeil, 2012).

Performance Management Techniques/Guidance

This section reviews current documentation, mostly that of the executive branch in the United States, to focus on the requirements and best practices of performance management in the federal government. Using these documents and their information will help me construct a performance management framework for reducing administrative burdens and measuring the progress of that focus on reduction. As discussed previously, it might be simpler to create information collection and measurement processes for administrative burdens without a focus on current performance management requirements. But this would potentially create brand new requirements and processes for federal agencies and administrators. Without mandates from the legislative branch or from a political executive, these efforts would face an uphill climb to gain

adoption, especially with a lack of resources. However, since performance management requirements and processes already exist, are resourced, and administrators have experience and expertise with them, I believe it makes more sense to integrate administrative burden measurement into these requirements and processes.

OMB Circular A-11 Part 6

As required under GPRAMA, OMB promulgated policy and guidance on how to implement performance management in the federal government. This guidance is primarily found in OMB Circular A-11 Part 6. Initially established in 2012, it has been updated in iterations, and new initiatives and guidance have been promulgated. These include sections on risk management in 2016, program and project management in 2018, evaluation and evidence building, sharing services, and category management, all in 2019. The guidance is the most up-to-date and mature version of the current implementation of GPRAMA since it has been iterated on to implement the legislation as well as weaving in other corresponding legislation, policy, and best practices that have been developed since GPRAMA was passed in 2010.

Key points from the guidance include: the focus on a “limited number of ambitious goals,” increased purposeful use of performance information through regular reviews, using evidence for continuous learning, improving performance information through transparency, and the focus on the circular’s performance management cycle. This means that leadership should be focused on limited numbers of ambitious goals. This is an important change with GRPAMA compared to prior performance management, which tended to attempt to measure and report everything within the government. Research has also shown the benefits of focusing on fewer goals and elevating these goals to leadership through routines and mechanisms (Moynihan & Kroll, 2016). The obvious hypothesis is that focusing on only a small number of goals that are important is essential to allow leaders, which are also busy responding and reacting to things that are important, to understand the importance of those goals for their overall strategy and to continue to apply focused pressure and interest needed to ensure their organizations are making progress to those goals. I believe this is also important from taking performance information from being a compliance exercise to a manageable input to leaders who can then actually use that information for decisions making, which is the criteria for purposeful performance information use one of the current measures of success of performance management (G.A.O., 2018; Gao, 2015).

Therefore, there is a direct correlation between the guidance focus on limited goals and the increased purposeful use of performance information through regular reviews (Moynihan &

Kroll, 2016). The research on prior iterations of performance management in the federal government continued to point to the creation of performance information which was then discarded when it came to leadership decision-making (White & Anderson, 2012; Wichowsky & Moynihan, 2008). Congress addressed this in the structure of GPRAMA, and research has shown that GPRAMA has resulted in higher rates of agency and leadership use of performance information and attributed this increased use to learning routines established in agencies that focused leadership on the strategic goals (with ambitious targets) through regular reporting and review requirements. Not only does the smaller number of goals help focus attention, but the regular reporting and reviews help establish learning routines that have been shown to correspond to effective performance management at the agency and sub-agency levels (Moynihan & Kroll, 2016). These routines follow the theories of learning epistemology because they focus participants on how the performance information impacts outputs and outcomes on a regularized schedule, allowing them to make adjustments and changes, which then provide them feedback in useful timeframes about the results of the program based on those changes (and other recent factors.)

The guidance in Part 6 has also begun to incorporate requirements and best practices from the Evidence Act, although there are still forthcoming sections being worked on. Already out are guidance on building and using evidence to support the learning routines from GPRAMA and requirements for program information transparency. This transparency is updated mechanisms from underlying GPRAMA beliefs and requirements that publishing performance information will allow for increased understanding, engagement, and oversight by the public as well as congress and other organizations. Research has shown little to no effects of these interactions on the successful use of performance information, but they are still firmly rooted in democratic ideals and have thus remained despite the understanding of how useful they are for improving performance management (Christensen et al., 2018; Desmidt & Meyfrootd, 2020). Most of the transparency requirements are about how and where to publish information about strategic goals, CAP goals, APGs, and the required reporting and reviews of those goals. The Evidence Act established a one-stop location for federal performance management reporting, Performance.gov, in addition to requiring agencies to make these available on their individual websites. They are also required to publish in machine-readable formats these goals, reports, and reviews as well as much performance information as can be safely made available while still taking privacy, security, and tiger limiting considerations into account.

An important section of this research is the guidance's performance management cycle. This section provides more practical guidance for agencies on how to develop and implement performance management routines both for the required CAP goals and APGs and for other strategic goals which do not necessarily rise to those levels. This is important because performance management research has also shown despite not being used by legislators, leadership at the agency and sub-agency (office or bureau) level have been more successful in performance information use (Destler, 2016; Moynihan, 2008). The performance management cycle, as defined in Part 6, is made up of the Planning, Analysis and Review, and Reporting stages. In the planning stage, the focus is on the long-term objective setting. However, it also includes not just these goals but the actions that will be taken and the resources which will be used. Additionally, agencies need to focus on the risks and challenges to their goals and how they will address them. These detailed and actionable plans are focused on three areas: Mission of the agency and sub-agency units responsible for the goals; Services of the agency, especially in interactions between citizens and the agency; and Stewardship of the agency especially related to measures of efficiency protecting against "fraud, waste, and abuse" (Office of Management and Budget, 2021).

The second part of the guidance's performance management cycle is the systematic planning, creation, analysis, and reporting of performance information which the guidance defines as "Evidence, Evaluation, Analysis, and Review." This is how agencies will monitor progress towards their strategic goals to make course corrections and provide transparency about their progress towards these goals. It is notable that Part 6 has integrated performance information and evaluation as required by the Evidence Act since these things are meant to be used in concert by the federal government (Office of Management and Budget, 2021). The "purposeful use" of these types of performance and evaluation information is established in this stage of the cycle through the learning routines established by the quarterly and annual reviews as well as reviews required for enterprise risk management.

And finally, Part 6 established the reporting requirements as the final stage of the performance management cycle to provide information about goal progress, performance, and evaluation reviews and analyses - including resulting program changes based on these reviews - to the public under the required reporting criteria of GPRAMA and the Evidence Act. Part 6 also established that these reports should provide the basis not just for interactive changes to programs but should also form the basis for the next iterations of strategic planning, whether for

the four-year or two-year cycles as required for different GPRAMA criteria (Office of Management and Budget, 2021).

Of note is the similarity between the performance management cycle as enshrined in Part 6 and many of the resources also provided to federal agencies and leadership by the PIC - which also has informed Part 6. It is not a stretch to say that while Part 6 establishes a formal OMB policy for agencies, the PIC resources go several steps further about how to develop and implement performance management in the federal government, even following the three-stage process.

The President's Management Council (PMC)

The President's Management Council has existed in similar forms as far back as the Reagan Administration and has varied focuses through administrations. It is composed of Chief Operating Officers (COOs) in the federal government, who are often Deputy Secretaries and Deputy Administrators, as well as Agency heads from OPM and GSA. This high-level of leadership ensures priority agency focus on PMC focus areas, and they have recently been heavily involved in promulgating the President's Management Agenda (PMA), which covers the Administration's top agenda priorities, which are the guiding initiatives that focus agency strategic plans, CAP goals, and AGPs (Stanley & Lutz, 2021). The PMC often is involved not just in setting these goals across the executive branch but also in monitoring progress towards these goals.

The PMC does not develop many documents or requirements on their own, but they are integral to the development and implementation of the president's management agenda, which sets many of the APGs and CAP goals. They often sit on the agency, and interagency performance review councils and teams, and they have leadership and oversight of the Performance Improvement Council (PIC) (Office of Management and Budget, 2021; Performance Improvement Council, 2019).

The Performance Improvement Council (PIC)

There are many documents and artifacts on designing and implementing performance management in the public sector. Especially with the recent decades worth of focus on the theory and practice of performance management. For this analysis, however, I am focused on the current state of performance management requirements and best practices in the federal government. Therefore, I am leaving out a significant number of historical resources which are specific to earlier implementations, state and local-level performance management requirements

(many of which mirror GPRAMA requirements), and most of the private sector implementation guidance for performance management. However, simply focusing on the current iteration within the federal government still provides a substantial number of rules, regulations, playbooks, best practices, frameworks, and checklists to incorporate into Framework 1. Significant pockets of these sources are producers by OMB, which is charged with overall policy, implementation, and oversight of GPRAMA in the federal government. The Performance Improvement Council (PIC) is a group established by GPRAMA and is made up of Performance Improvement Officers (PIOs) and staff from twenty-four federal agencies, the Deputy Director of OMB, and dedicated OMB and GSA staff. The PIC works collaboratively with OMB leadership to promulgate best practices, playbooks, and approaches to successfully implementing GPRAMA performance management processes throughout the federal government.

The PIC Playbook and Performance Management Resources

As discussed previously, the PIC was established in 2007 by E.O. 13450 and then codified in the GPRAMA. Chaired by the Deputy Director for Management at OMB and including PIOs and staff from federal agencies, they have been focused on best practices of performance management design and implementation in the federal government. PIC has promulgated and maintains many resources for federal agencies to design and implement performance management. These include specific material aimed at CAP goals, APGs, and strategic planning and reporting. They also created and maintained guidance that is agnostic to particular performance management requirements and can be generalized to any implementation. One useful resource is their goal playbook, which like A-11 Part 6, breaks performance management into a three-part cycle: set, plan, and achieve.



Figure 2-1: PIC Playbook Performance Management Cycle

The Goal Playbook provides agencies with more granular resources and guidance and develops their goals and performance management systems around those goals, as shown in Figure 1-1Figure 2-1: PIC Playbook Performance Management Cycle. It is also useful to understand that, by design, the Goals Playbook augments and builds on OMB A-11 Part 6 guidance by providing useful paths forwards for federal agencies. In “Play 1: Set the Goal”, the focus is on a small number of manageable priorities and focusing the goals on those priorities to provide consistent focus on achieving these goals. Key criteria for this stage are the focus on senior leader buy-in and visibility on the goals, which has been shown to be an important factor for the use of performance information. Additionally, the Playbook focuses on clarity up and down the chain of command on the goal purpose as well to ensure that there is clarity of understanding about the goals from the perspectives of everyone involved in activities that will be important to the goal. The goal statement needs to be refined to ensure that there is clarity on the goal in terms of being “specific, time-bound, and measurable,” which are criteria used in many performance management frameworks because these are important not just for clarity of goal understanding, but to ensure that the goals are subject to the collection and analysis of performance information and performance measurements (Performance Improvement Council,

2019). Without specific time constraints which are measurable, these strategic goals become aspirations and visions statements that could be justified and measured subjectively. Locking in these requirements is foundational to setting up a measurement framework. Lastly, in the planning stage, the goal team is focused on ensuring key individuals from senior leaders and dedicated staff are identified as are their goals in goal implementation, measurement, monitoring, and any changes needed based on the feedback.

The second stage in the playbook is the planning stage. The planning stage is about setting the plan on how you will collect and use the performance information in relation to the strategic goals. This is an important concept because it is not just about how to collect the performance information and display it to stakeholders, but it is proactively thinking about and setting rules and norms about how the information will actually be used by the organization to make meaningful decisions that align with the stated goals based on the information collected (Performance Improvement Council, 2019). In terms of what I have discussed previously for GRAMA, this step of the playbook is helping agencies understand and identify how they will implement their learning routines for decision-making - which has been shown to be a major difference between the existence of performance information and the purposeful use of performance information - which is the measure of successful performance management (Choi & Moynihan, 2019; Kroll & Moynihan, 2017; Moynihan & Kroll, 2016). The planning stage focuses on creating external awareness about the goal landscape, identification of key stakeholders in the organization, choosing effective strategies for program management, creating a performance measurement plan based on those strategies, setting a clear decision-making plan for the use of the performance information that includes both structures and cultures, and finally a communication plan to proactively message goals, plans, and decisions to all internal and external stakeholders.

Phase three of the Playbook is “Achieve,” which is about implementing the plan effectively as well as setting up an understanding of how the performance management process for a goal may need to be modified based on information gathered during the process of monitoring (Performance Improvement Council, 2019). For example, when collecting and assessing performance information, the focus is not just on communicating progress towards the goals to the important stakeholders and making program decisions to ensure progress to the goals - it is also focused on analyzing to determine if the right performance information is being collected if the correct milestones are being tracked in order to modify the performance measurement plan as needed. Similarly, there is an emphasis on making sure the right people are

being included as this can be different than what was initially planned for, or the emphasis on stakeholders can adjust based on external factors. Based on these assessments and meta-assessment, the team needs to be prepared to make changes to the goals as well as the performance measurement plan regarding the goals to continually iterate and improve the process. This is a key component of successful performance management because it is turning the performance information created and collected by setting clear goals and measurement plans into operational and policy decisions based on information collected to manage towards successfully meeting the strategic goals. This is where performance management differs from simple monitoring or reporting because it necessitates an active learning and change process. Other components of the Achieve stage are continued communication with stakeholders, both on reporting on progress, decisions, and changes but also on a continued engagement on feedback from stakeholders, which also can become data to be reviewed, analyzed, and potential for decision-making about implementation or about the performance measurement plan. The guide also focused on recognizing and celebrating success and the collection, documentation, and double-loop learning from lessons learned throughout the process to help inform future performance management (Performance Improvement Council, 2019).

Federal Document Analysis for Measuring Administrative Burdens

Since the concept of administrative burdens is relatively new from the United States perspective, there are not many documents that are directly attributable. However, I can inform Framework 1 by including requirements and information from some of these sources, which are very close but indirectly related to administrative burdens. Of note, measurement and reductions of paperwork times and burdens are closely related to learning and compliance costs (Madsen et al., 2020; Sunstein, 2020). Additionally, the recent Executive Order on Transforming Federal Customer Experience (E.O. 14058), signed by President Biden on December 13, 2021, is a significant and direct acknowledgment of administrative burdens and the impacts they have on individuals' experiences with federal government services and programs (*FACT SHEET*, 2021).

E.O. 14058

This executive order is informed by the recent and growing focus on administrative burdens, and it directs the federal government to be aware of and actively seek program implementation and design choices with the experience of individuals in mind (*Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government*, 2021). It directs agencies to consider design choices and technology with the

express aim to facilitate the interactions between individuals and the government when seeking services and to make them “simple to use, accessible, equitable, protective, transparent, and responsive...” (*Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government*, 2021).

E.O. 14058 doesn’t specifically discuss the three types of costs of administrative burdens, but it does refer to the frictions as “time tax(es),” which are experienced and paid by individuals unless the government takes measures to remove those burdens or to take them on instead of placing them on individuals. In Section 2, the E.O. specifically calls for the use of performance measurement of individuals’ experiences and the prioritization of person-centric design to reduce the frictions and amount of time that individuals must spend to access government services. It goes on to direct federal agencies to:

“continually improve their understanding of their customers, reduce administrative hurdles and paperwork burdens to minimize “time taxes,” enhance transparency, create greater efficiencies across Government, and redesign compliance-oriented processes to improve customer experience and more directly meet the needs of the people of the United States. Consistent with the purpose described in section 1 of this order, agencies’ efforts to improve customer experience should include systematically identifying and resolving the root causes of customer experience challenges, regardless of whether the source of such challenges is statutory, regulatory, budgetary, technological, or process-based.

The E.O. is unique in that it focuses prior federal government efforts on paperwork reduction and customer service to reduce frictions through a “human-centered” design and implementation approach. Rather than providing this direction in a manner that must be folded into broader performance management efforts, though, the E.O. identifies specific “High-Impact Service Providers (HISPs)” among the federal agencies and then lists a specific list of target changes to programs among those agencies. The E.O. does set up a framework for future identification of additional programs of focus, but it limits its scope to be specific about designated HISPs while requiring heads of agencies to comply with the E.O. in their strategic plans, APGs, and CAP goals (*Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government*, 2021). Since E.O. 14058 is new, it is difficult to tell what its long-term impact will be, but it is a step in the right direction toward

using existing performance management processes to focus on and reduce administrative burdens.

The Paperwork Reduction Act (PRA)

The Paperwork Reduction Act (PRA) was signed into law in 1980 and has gone through multiple iterations. The PRA is managed by the Office of Management and Budget (OMB) Office of Information, and Regulatory Affairs (OIRA) sets out several standards and requirements for the federal government when it wants to impose information collection from the public. It applies in nearly all situations with only small exceptions. It requires the government to set out specific goals for the information collection to justify why the collection is needed and how the information will be used by the government. It must also specify the plan for the collection of this information, including a detailed explanation of processes, forms, and options of providing information (electronic or analog), and potentially run a test of the proposed information collection. The PRA also requires the publication of the proposed plan, information, use, and collection techniques in the federal register for 60 or 90 days for public review and comment. After this period, the agency must respond to comments and update the proposed plan based on comments when appropriate or judged reasonable. Importantly for our purpose, the PRA also requires an estimate of the burden (usually based on time) of the proposed information collections specific to each mechanism of collection. These estimates are included in the public notice as well as reviewed by OIRA for “reasonableness” (Funk, 1987; *Paperwork Reduction Act Guide*, 2017).

The PRA’s estimates of impact and judgment of “excessive burden” have been a standard area of administrative law and compliance for many years. However, agencies typically take them as a one-off for each information collection or form that they need to require in order to administer programs and benefits and do not interweave them into other performance measurement metrics, which tend to be internal process focuses. However, the OIRA guidelines for agencies on that activities require calculation and estimation for the PRA in Figure 2-2, which show how these activities and burdens estimate track to the three types of administrative burden costs (*Estimating Burden / A Guide to the Paperwork Reduction Act*, 2022a).

Common Burden Activities	Reviewing Instructions
	Compiling materials necessary for collection
	Acquiring, installing, and utilizing technology and systems
	Adjusting existing ways to comply with previous instructions and requirements
	Searching data sources
	Completing and reviewing collected information
	Compiling and sending information

Figure 2-2: OIRA Burden Activities

Additionally, OIRA focuses agencies on what they deem to be “excessive burdens,” which OIRA is charged with screening out during the PRA adjudication process. Excessive burdens under the PRA are information that is not justified by the use of the information – so information collected but not pertinent to the decision being made based on the information as well as information collected which would risk harming someone’s privacy by identifying them with protected classes or statuses. Additionally, an excessive burden can be requiring information more frequently than the agency can justify based on the use of that information, such as monthly updated information when the enrollment for a program is only made annually based on aggregate annual information. One other important concept of excessive burdens is the approach and format of requested information (*Paperwork Reduction Act Guide*, 2017). For example, requiring forms to be completed manually, authenticated through a notary, and submitted to a central office in person during minimized business hours, only for the data to be transcribed into a digital format of structured data for processing and storage could be grounds for a violation of excessive burdens by OIRA.

Under several presidential administrations, OIRA has also published an annual report titled *Information Collection Budget of the United States Government*, which is an aggregate accounting of all PRA estimated across the federal government that purports to show the total burden placed on individuals each year by the federal government information collection. In 2009 it reported a 9.71 billion hours burden and 9.78 billion hours in 2016 (*Estimating Burden /*

A Guide to the Paperwork Reduction Act, 2022b). The Sludge Audit proposed by Cass Sunstein, former OIRA Director under President Obama, focuses his performance measurement of administrative burdens on these PRA metrics which are being produced (Sunstein, 2020). Additionally, as discussed in the next section, other researchers focus on the burdens of learning and compliance costs which can be measured and identified through the PRA burden estimates process.

The Evidence Act

The latest iteration of congressional-required performance management in the federal government is included within the Evidence Act, which will be discussed in greater detail in a subsequent section of this paper. However, it is important to include it here as it also builds on existing performance management requirements, and while it doesn't supersede the GPRAMA, it does iterate on it and include additional guidance and requirements for the executive branch. Notably, as will be described in greater detail, it re-emphasizes the importance of and requirements of performance management protocols and mechanisms. It further refines how CAP goals and APGs must be promulgated and aligned to agency strategic plans, as well as the oversight and usage mechanisms of GPRAMA (Commission on Evidence-Based Policymaking, 2017). What it adds importantly is a more explicit linking of the use of performance management by agency leadership to other existing and new requirements for evaluations. Specifically, it attempts to link together performance information which is by nature more short-term and typically output-focused, to longer-term evaluation data, which is often more focused on the outcomes of the programs (Ryan, 2019). Performance management and evaluation requirements have often been on parallel tracks in the executive branch, and many agencies have kept these activities separate for multiple reasons. But the Evidence Act builds on research and beliefs that there are more similarities than differences between performance management and evaluation and that they are similar tools that should be used for the same goals within the public sector. The Evidence Act also tries to build the infrastructure around the creation, implementation, and use of performance information and evaluation data within the government as a recognition that these are desirable capabilities for many reasons, and therefore the government must be positioned to take maximum advantage of them.

Use of Performance Management for Decision-making

As described, the performance management movement continues to play out in the United States and around the world, being implemented and adopted by fiat through legislation,

executive policy, and internal and external pressure from stakeholders and leadership alike (Kroll & Moynihan, 2015). There is a significant amount of research that shows it has thus far not worked as expected; however, there have been tangential benefits and improvements in certain areas based on the movement. The theories behind supporting performance management claim inherent benefits deriving from different drivers. These include allocative efficiency, accountability of government to the public, accountability of bureaucracy to elected officials, and technical efficiency. In allocative efficiency, it is expected that budget agents will use performance information when making budget decisions, thereby spending money on the most effective and efficient programs. Accountability to the public is often espoused because performance information will be created and provided to the public, allowing increased transparency and public inclusion in the oversight of public expenditures and programs. Accountability to elected officials is similar, except that the political leadership has greater information and performs the oversight functions of government administrators and programs, which is increased because of the setting of goals and transparent accounting of progress toward those goals. The idea behind technical efficiency is that the performance information allows for single-loop learning about programs which is fueled by the autonomy and authority devolved to managers, allowing them to directly manage programs informed by this learning which increases expertise and effectiveness (Moynihan, 2008).

The research on and oversight of early adaptations of performance management in the United States focused on the implementation of performance management systems, specifically the creation and collection of strategic goals and the performance information about those goals. In many instances in the federal government, success is tied to compliance with GPRA and PART through the analysis and review of strategic plans and performance information systems. However, it has more recently turned to the realization that the “success” of performance management cannot be simply the existence of these plans and activities but the actual use of performance information that they create for decision-making, especially in decisions which directly related to the programs themselves, as well as congressional use for budgeting and policy and programs decisions. Therefore, the success of performance management is not the existence of the components but instead the use of performance management for decision-making. The existence of the components of performance management is necessary but not sufficient. This is a key distinction because even more than avoiding the creation of compliance-focused, “box-checking” exercises, the theory of performance management is that it creates tools to be used for the specific purpose of decision-making which can increase the effectiveness of

leaders who are overseeing these programs. Therefore, there is strong importance to look not just at the existence of the performance management requirements and infrastructure but the actual use of the outputs of the performance management infrastructure for decision-making (Cavalluzzo & Ittner, n.d.; Destler, 2016; Gao, 2015; Moynihan & Kroll, 2016). This concept has become known as “purposeful performance measurement”, which is defined as designing and implementing performance measurement in a way that allows for, or ensures the use of performance information in decision-making (Johansen et al., 2018; A. Kroll, 2015b; A. Kroll & Moynihan, 2015).

Assessment of Performance Information in the United States Government

Through all these initiatives and efforts, there is near-universal adoption of performance measurement in U.S. federal agencies today. However, there is a great deal of variability in the implementation, design, structure, and where performance measurement is situated within agencies. Most importantly, there is also high variability in the utilization of the produced performance information by agency leadership. One of the most prevalent assessments and criticism of performance management over the past few decades in the United States is that, ironically, there is little data supporting the success of performance management in helping achieve its intended goals (Dooren et al., 2015; G.A.O., 2018; Gao, 2015). Many instances of implementation have been shown to be more compliance exercises than purposeful use of performance information by leadership in agencies. Additionally, many have also criticized implementation, which fails to meet criteria that would even meet the goals of purposeful performance information use because goals, performance measurement standards, and any reviews are too broad to provide meaningful conclusions about how to change or modify programs accordingly.

Several researchers have also looked at the intended effects compared to the actual outcomes of performance management policies. Researchers and proponents of performance management cite one of the fundamental cornerstones of the movement as providing better information about performance for policy decisions (Bourdeaux, 2008; Johansen et al., 2018; Kroll & Moynihan, 2017). Therefore, an important indicator of the success of any program should be the extent to which politicians use performance information in decision-making. This can come in the form of performance information being tied to strategic planning purposes, decisions about resource allocations based on performance management results, or program and policy adjustments that are made in response to analyses of performance management

information. Head (2016) explores the extent to which government organizations and policymakers are utilizing data and evidence in policy making and implementation decisions. He notes that there has been increased emphasis on the utilization of data and pilot evidence when designing and implementing programs and that democracies around the world have increasingly focused on better design, improved effectiveness, and increased efficiency of domestic and international programs. However, while there are reports of increased data creation by government agencies, there is scant evidence that this data is used by government organizations and policymakers (B. Head et al., 2014; B. W. Head, 2016). Others have shown that politicians routinely point to evidence and performance information only so far as it supports their preconceived policy positions, which are often based on ideology and constituent beliefs rather than situated in evaluations of effectiveness (House & Shull, 1988; Moynihan, 2008). Furthermore, it has been shown that the public often does not consume performance information directly, if at all. Instead, their use typically comes through filtered channels, and they are only likely to encounter it in a form where it is being used selectively to support a policy position, such as through an interest group report, public awareness campaign, or directly from a politician. This is in direct contrast to proponents of the movement who claim that increased public performance information will lead to increased citizen oversight. It's often found that individual citizens prefer to outsource their review and analysis to organizations they already ascribe to (Moynihan et al., 2015; Olsen & James, 2017).

In response to this, some have called for academics to take the needs of legislators and program managers into account when conducting policy and program administration research to better align the ideals of performance management. They argue that legislature and program staff also need to meet somewhere in the middle and be more transparent with their constituents about policy decisions, as well as the evidence they are using when claiming they are performing evidence-based policymaking or using performance information to inform their decisions. They also argue that this transparency extends to their duties to explain to constituents why they made certain policy decisions over others and when the facts necessitate a decision that is at odds with the values of their constituents. In doing so, they believe that democracies can change the nature of political values so that voters and stakeholders will incorporate facts and evidence more than currently seen (Arinder, 2016; Dooren et al., 2015; Moynihan, 2008).

In other instances, performance management implementation has been criticized as being unrealistically delinked from the actual systems of government and control. Notably, the ideals of objective performance data driving decisions have become too lofty and theoretical in some

conversations without allowing for realities surrounding political-driven control of government and agencies under the United States system. Others have argued that actual implementation, especially changes made more recently under GPRAMA, have specifically taken into account political-driven direction and control by aligning the performance management framework to the political cycles. They argue this allows it to be more objectively distinct from political ideology and be a tool of “how” government functions rather than controlling the “what” or the “why” of government policy and programs (Dooren et al., 2015; *Government Performance Results Act of 1993*, n.d.; Moynihan, 2008; United States Government Accountability Office, 2005). Research has called into question the underlying agency and structuralism assumptions and perspectives, which believe that inherent subjectivity disallows the collection of knowledge or truth in a way that is meaningful to practitioners and trying to determine causal mechanisms and policy programs that work. While the evidence-based policy movement and performance management movement has expanded, there is research to remind us that policy is ultimately a political action, and therefore evidence can only ever be one factor considered by policymakers – not the entirety. If taken too far, this sole reliance on data can become so restrictive that policy cannot be implemented until it is proven to surpass very high bars, which are incompatible with the political policy process. Additionally, the over-reliance on randomized control trials as the only source of pilot testing in a realm that requires much broader social science tools is seemingly too narrow (Sanderson, 2002).

However, despite the lack of evidence for performance information used by legislators and politicians in the United States, there has been optimism about the ability of government agencies and organizations to successfully implement performance management systems. Additionally, there is evidence that these agencies and sub-agency organizations are beginning to realize real benefits from performance information that links directly to the intent of the movement. While not on a macro-level within the government, it is promising to find pockets of the government realizing benefits from the efforts (Johansen et al., 2018; Moynihan et al., 2017).

Gaming has also been shown as a problem for performance management. Other research has shown how the adoption of performance management metrics leads to gaming of the performance information results. This can be especially true in the case of CompStat in New York City, CitiStat in Baltimore, or other “stat-like” programs which focus on a very public data-driven analysis by senior leadership. The tone and tenor of many of these instances focus on accountability and shaming of the individuals where the data show failures or lack of improvement. In order to avoid these public embarrassments or loss of resources, leadership has

been shown to game the numbers to inflate the “successes” which are reported under the system or recode reported incidents in a way that does not count negatively against them (Hassan & Hatmaker, 2015). Similarly, in areas such as education, where funding and public esteem are directly tied to performance measurement criteria, there have been numerous reported and documented instances of gaming of the information (Johansen et al., 2018). These are examples of how the adoption and implementation can lead to worse data for analysis of policies and programs. This is a real danger for the movement as it directly affects the objectivity of the performance data. False or misleading performance information precludes most intended benefits of performance management. Other research has pointed to how subjective the interpretation of performance information can be, leading researchers to question the efficacy of the process and even how to judge it. In some experiments, well-informed participants reached different and opposite conclusions based on a review of the same data and performance goals (Moynihan, 2008).

Successful Use of Performance Information in Federal Government Agencies

Despite the evidence that the ideals of performance management are often not met or that the ideals may actually be incompatible with representative democracies, there is also a good deal of evidence that the implementation of performance management and use of performance information does lead to several benefits in terms of effectiveness and efficiencies. This has been shown to be especially true at the sub-agency level when focused on specific programs and work units rather than intra-agency, high-level policy (Hassan & Hatmaker, 2015; Moynihan, 2008). Johansen, Kim, and Zhu (2018) found that certain traits of managers of non-profit organizations can encourage the use of performance information in organizational decision-making. These factors are the manager's self-efficacy, receptivity to feedback, sensitivity to the external environment, and how much they value employee input. They find that the sector can influence the use of performance information in the areas of incentives and the capacity of employees. They also find that the manager's variables have a significant and positive correlation to increased use of performance information, likely because they see performance information as an extension of feedback (Johansen et al., 2018).

It is promising that research has found evidence for the successful implementation of performance measurement in federal agencies and sub-agency organizations. Rather than faulting political agents for their lack of use of performance information in an objective manner, I believe the federal government should focus on performance management where it can be best operationalized for objective measurement and feedback. I believe this reflects the structure of

the United States government, especially in the context of professional civil service. While elected leaders are, and arguably should be, influenced by the values and beliefs of their constituents once a policy or program is implemented, I expect the professional civil service managers to utilize tools such as performance information to most effectively and efficiently implement and manage these policy programs. Therefore, it is important to continue to focus on how best to implement performance management in a way that makes it most likely that agency leaders will be able to use performance information in decision-making.

There has been a good deal of research that has been exploring exactly this question. Thus far, I have identified several important factors which are associated with these desired outcomes. These are measurement system maturity, stakeholder involvement, leadership support, agency support capacity, innovative culture, and goal clarity (Kroll, 2015a). Performance measurement system maturity relates to the existence of performance information systems within an agency that facilitates its use. In mature systems, information produced is not only valid but also readily available throughout the organization and thus more easily utilized by leadership (LeRoux & Wright, 2010; Moynihan & Pandey, 2010). Stakeholder involvement can vary greatly. In some instances, public organizations (especially many non-profits) are required by their stakeholders to adopt performance measurements. Sometimes this is in response to transparency requests, but this can also be a requirement for receiving certain funds. Some research has shown that this external focus may not greatly impact the use by decision-makers (LeRoux & Wright, 2010), but in other instances, stakeholder interest in the form of constituents does lead to successful use by decision-makers as a way to justify expenditures (Moynihan & Pandey, 2010).

Learning Routines

One of the prominent focuses on the use of performance information by decision-makers has been on the mechanisms of learning routines. This is because the theory behind the creation, understanding, and use of performance information for program management is really tied to learning theory because they create performance information to inform and allow management to learn about what is working, what is not working, and the mechanisms of their programs (Moynihan & Kroll, 2016; Moynihan & Lavertu, 2012). Some of the more recent and relevant research for this conversation is on the implementation of the GPRA Modernization Act.

The goals of the GPRAMA are to increase managerial and leadership use of performance information for decision-making, which fell short in the implementation of its predecessors PART and GPRA. A specific area of the GPRAMA was the establishment of new routines

which are meant to increase repetition of focus on the use of performance information for decision-making. The framework of examining reform efforts to implement and formalize new routines has a history in public administration research which has found that successfully establishing new routines is associated with successful reform initiatives. Specifically, there has been a focus on three new routines in GPRAMA: cross-agency priority goals, agency priority goals, and OMB-led quarterly reviews of goal progress (U.S. Government Accountability Office, 2017). Researchers believe that one of the strengths of the GPRAMA design is that it focuses agency leadership on a small number of goals rather than creating an ever ballooning number of goals pulled from an increasing number of strategic plans for every program. Additionally, the sustained focus and reviews of progress towards those goals allow agency leadership to focus efforts and resources, making it more likely that goals will be achieved. Numerous studies have shown the implemented effects of performance management systems and performance information collection have often fallen far short of their intended goals. There are many factors researchers point to for these failures of performance management policy. Primarily, although the most popular performance management systems were created through the GRPA and GPRAMA legislation, there is scant evidence that shows that congress readily uses performance information in decision-making. Even those within the congressional budget staff that are aware of the performance information available for programs often don't cite it as the main decision factor (Moynihan et al., 2017).

Notwithstanding the successful implementation and the unsuccessful implementation, there is a question that needs to be answered about how performance information is used once adopted and implemented. Thankfully, there has been some recent focus on this question. For many researchers and proponents, it is not enough that an organization has implemented performance measurement and produces performance information even when this includes outcome-level data. There is a further emphasis and question about whether that performance information is then used in organizational decision-making (Johansen et al., 2018). Ultimately, this is the end goal of the performance measurement movement and the logical question to study when analyzing the success of performance measurement.

At the organizational level, Kroll (2015) also found that the relationship between performance information use and organization performance is important. His findings show support for the idea that for performance management to be successfully used, it must be tied to a new vision, better outcomes for stakeholders, or generally significant change in the organization. Additionally, organizational strategy matters, and there is much more likely to be successful

implementation and use of performance information use when associated with stretch goals or innovations. The systems will be less effective if used to monitor existing work and goals (Kroll, 2015c).

Leadership Influence

But much research has focused on the individual leadership levels of organizations in the federal government. Kroll (2015) found the use of performance information by public managers through the physiological-cognitive model of planned behavior. This is an extension of prior research that has found that the establishment of learning forums where performance information can be discussed and made sense within a group can improve the purposeful use of performance information. He specifically utilizes Ajzen's theory of planned behavior, which posits that the performance of a behavior is contingent on an individual's attitude toward the behavior, the subjectively perceived social norm, and the behavioral control. He links this to the behavior of performance information use by citing several key research findings that show relationships between positive attitudes of management and leadership towards its use, often associated with increased use. There is also research that links social norms, often in the form of external stakeholders and peer groups, with increased use of performance information. He finds a significant positive correlation between the presence of a positive attitude toward performance information use and social norms correlates to higher intentions to use performance information. He also finds that managers' intent to use performance information is correlated to their engagement in improving data usability (Kroll, 2015b).

Others found that performance information use was more likely to be associated with altruistic feelings on the part of the manager toward their mission than self-interests - especially that of protection. Organizational culture, flexibility, and professionalism have also been found to be positively associated with performance information use. Additionally, the presence of public service motivation at the individual manager level and the organizational culture level were good predictors for performance information use (Moynihan & Pandey, 2010).

Some researchers have focused on the state-level performance management efforts adopted in response to these federal efforts (Rogge et al., 2017). State and local governments can provide interesting case examples because, although they often implement performance management in homage to federal programs or because they are nudged towards doing so through federal initiatives and grant requirements, they also have distinct power structures compared to federal agencies, thus providing varied research targets. Additionally, states have different causal factors for the effectiveness of these efforts, which can be complicated by

different government structures. For example, having a weak executive (governor) can lead to stronger state agency-level oversight which can influence increased adoption of performance management, whereas strong state executives might pay lip service to performance management but might not follow through on these promises due to short tenure or short attention spans (Bourdeaux, 2008; Moynihan, 2008).

Leadership support has been a strong area of focus. This is not surprising since the idea of leadership's control over an organization is an obvious factor to consider in any actions or decisions of that organization. Not surprisingly, research has found that leadership interest and support of performance measurement are associated with successful adoption and use in decision-making. However, there have been some interesting findings on how this causal mechanism can work. In some instances, leadership is very involved in the design and use of performance measurement systems. In others, the causal mechanisms are more subtle. In some instances, leadership is crucial in setting organizational culture, which indirectly encourages the use of performance information in decision-making by setting clear goals for the organization and encouraging a culture that prioritizes and looks favorably on data-driven decisions (Moynihan, Pandey, & Wright, 2012). In other instances, leadership creates routines such as performance information reviews and follow-up (such as those associated with CompStat, CitiStat, and other 'stat-type programs) or through the existence of leadership-driven discussions of performance information as it related to program goals (Moynihan & Lavertu, 2012). This was one of the initiatives that GPRAMA attempted to leverage by pairing down the number of strategic goals and specifying an even smaller amount which agency leadership needed to focus on annually, but also emphasizing regular oversight and review of programs towards those goals (Dull, 2008; Hassan & Hatmaker, 2015; Moynihan & Lavertu, 2012; Yang & Yi Hsieh, 2007). In order to test the effectiveness of GPRAMA in establishing new routines to encourage greater use of performance information in decision-making, Moynihan and Kroll (2016) used the 2013 GAO survey of federal managers (an iteration of a survey GAO has conducted in a comparable way several times) to look at the reported uses of performance information by federal managers. They find that the survey data show increased use of performance information after the implementation of GPRAMA in comparison to that found under PART and GPRA. They also find positive effects from the implementation of GPRAMA data review routines and positive effects on the agency priorities goals and cross-agency goals. This is supported by the theory as it indicates having limited goals and constant routine follow-up on progress helps agency leaders focus on using the performance information to make progress towards the goals (Moynihan &

Kroll, 2016). This is in direct response to research that showed that the routines that were adopted in GPRA and PART did not result in increases in performance information use (Moynihan & Lavertu, 2012). It is encouraging to find not just successful adoption and implementation of performance management but instances in federal agencies where it has been used for decision-making as well.

Organization Structure

Other research has looked at factors that have proven ineffective for impacting the adoption use of performance measurement. Interestingly, the size of an organization has not been shown to be significant, while initially, there were thoughts that larger organizations would be associated with more performance information use that did not prove justified (Kroll, 2015a). Notably, the education levels of employees and leadership also were not significant (LeRoux & Wright, 2010; Moynihan et al., 2017). Nor did the presence of financial distress for the organization or the existence of political competition, either for survival or the program or for the primacy of the agency in the policy sector (Bourdeaux, 2008; Dooren et al., 2015)

Training

Kroll and Moynihan (2015) attempt to answer questions about how training influences the successful adoption and implementation of performance measurement in the federal government. This is an important question because each successive performance management doctrine has identified training as a necessary component. In fact, the GPRAMA specifically requires OPM to identify and implement requisite training for agency staff. A large portion of training had a primary focus on educating employees about the performance management requirements and systems. Conversely, capacity training would help agency staff successfully implement and use performance information by giving them the knowledge and skills needed rather than just informing them of the requirements. Capacity training should include a curriculum that would cover how to measure performance, how to use discretion, how to learn from performance data, and how to use performance data for accountability purposes (Kroll & Moynihan, 2015).

Summary

While there is almost universal acceptance of the ideas behind performance management systems, many argue that the processing of information is fraught with problems and incongruences across federal organizations. Some constraints in the use of performance information come from politicized policy areas that are highly partisan. In highly political policy areas, values, negotiations, and persuasion often crowd out the use of expertise and data to guide decisions or problem discussions. When politics is driving policy decisions, there must be result of conflicts, trade-offs, and compromises that inherently will not align with the more objective data and evidentiary recommendations. Another factor is looking at the diversity of organizations and their leadership. Different agencies and their senior leadership will vary according to their preferences and structure. The types of organizations (such as service delivery, regulatory oversight, or policy development) and the domains (such as social policy, economic development, or environmental regulation) will also either inhibit or embrace evidence-informed policy (Head, 2016). Others have been careful to call attention to the instances where performance information is used not to guide policies or decisions but merely to justify preconceived political or policy positions. However, it is important to note that this can and should be expected from a representative democracy structure such as the United States government. In fact, several researchers point out that the objective and simple data-driven goals of the performance management movement are somewhat incongruent with political systems which demand representation by elected officials (Head, 2008; Moynihan, 2008).

The Intersection Between Performance Management and Evaluation

In future sections, this work will focus on the role of evidence-based policymaking and evaluations in the public sector. But it is important to draw the lines of similarity and distinction between performance management and evaluation. To recap, performance management focuses on the setting of strategic goals and the creation, collection, and monitoring of performance information to allow administrators to pursue those goals. Performance management is typically concerned with measuring inputs and outputs against those goals (G.A.O., 2018). Evaluation, on the other hand, is a systematic approach to a program's overall theory of change, including the inputs, outputs, and outcomes of the program and evaluating both the implementation of the program as designed as well as the success and causality of the inputs and outputs on the outcomes of the program (Orr, 2018).

In terms of an example public program, let's focus on EITC. The program takes the input of applicants filing tax returns who are employed but meet certain income levels and who have qualifying children or dependents. The outputs from this program are the adjudications of eligibility for EITC at different levels based on the inputs. A performance management program would look closely at measuring the demand for EITC applications through tax refunds as well as the adjudications of results and issuances of EITCs. There could be a much more detailed layer where the timelines, throughput, unmet demand, and the administrative burdens of the training and testing plan are accounted for so that the administrators of the EITC are able to measure and make decisions in line with strategic goals of the program.

Evaluation for the same program would look at the underlying theory for change. They would note that the program is in place not just in order to meet the eligibility requirements to impose an adjudication process but because their hypothesis is that these requirements will result in lower poverty rates for eligible individuals as well as higher rates of employment since the EITC has the goal on encouraging employment for lower-income individuals. An evaluation would approach looking at the outcomes of the EITC program to measure and test the hypothesis that the eligibility and adjudication process actually achieved the desired results, that those eligible and issued EITC have lower poverty rates of symptoms, and that they are able to maintain employment at higher rates compared to those who did not go through the process as will be discussed in greater detail in a future section the difficulties of these evaluation approaches in the public sector.

While the differences between evaluation and performance management become clear in this example, I believe that these are both useful tools for different but related projects in the public sector. Performance management helps manage implementations and towards implementation goals. Similarly, a formative evaluation can help understand the success of a program's implementation as compared to the program conception. Once successfully implemented, a summative evaluation is important to understand if the program as implemented is actually achieving the proposed outcomes and objectives as it is understood in the theory of change (Patton, 2012). Evaluations tend to be a longer-term and more resource-intensive process, so they are not always practical for the day-to-day management of programs like performance management techniques can be. Similarly, the performance measurement techniques and performance information created and collected can be very valuable for parts of the evaluations (Dey et al., 2015; Orr et al., 2019).

As I will show in future sections of this paper, the performance management, evaluation, and evidence-based policy movements have seemingly been progressing on different tracks in the federal government, but the recent adoption of the Evidence Act strives to bring these threads together to leverage their unique benefits for the shared goals of improved effectiveness and efficiency of public sector programs.

Implementing Performance Management

This research takes performance management requirements as they currently exist in the federal government and uses them to build performance measurement and management processes to identify and set goals to reduce or eliminate administrative burdens. I am not attempting to modify or change the existing performance management requirements for the federal government. Substantial changes to performance management in the federal government are outside of the scope of this research because I am focused on understanding and providing tools that can be applied in government now, under existing conditions and requirements. However, I do want to focus on the emphases and protocols which have shown to be most effective for the purposeful use of performance information by decision-makers. Therefore, this section will combine research, existing regulatory and procedural requirements, best practices, and playbooks for performance management systems and will be focused on identifying and measuring administrative burdens.

Importantly, I am focusing this framework on identifying, measuring, and reducing or eliminating administrative burdens in programs. This is important because performance management requires strategic goal setting, which must be more specific than simply “measuring administrative burdens.” In performance management, administrators must set intentionality and direction about what they want to measure and change regarding administrative burdens (Moynihan & Kroll, 2016; Wichowsky & Moynihan, 2008). As discussed previously, administrative burdens can be either intentional or unintentional. Additionally, they are open to values-based subjectivity about whether they are beneficial or harmful depending on policy goals and beliefs about programs, recipients, risk tolerances, and expenditures for programs (Burden et al., 2012; Herd & Moynihan, 2018). I am acknowledging these possibilities but am nonetheless characterizing administrative burdens as negative phenomena which are worthy of reduction or elimination in public programs. The justification for this in this paper is that they add costs and frictions which induce inefficiencies in programs, as well as hampering program implementation, which reduces our ability to measure and evaluate the program to make objective decisions about its success and resources (Herd & Moynihan, 2018).

From both a performance management and an evidence-based policymaking perspective, administrative burdens, especially those not identified and measured, are negative and undesirable. Adopting performance management techniques to identify and measure them can bring them out of the shadows into a transparent oversight and monitoring process, which might negate these negatives and allow policymakers and citizens to have a more informed and objective conversation about their worth and whether they should be eliminated, modified, or even increased. These types of citizen interactions and policy discussions are out of scope for this research, but it is important to note that Framework 1 should be viewed as a beneficial pursuit even in instances where someone believes that administrative burdens are beneficial and potentially should be increased. Several researchers have characterized administrative burdens as default negative phenomena until or unless they can be made more objectively transparent to enter the administrative law and values-based policy discussion atmosphere (Ali & Altaf, 2021; Chudnovsky & Peeters, 2021; Heinrich et al., 2021; Herd & Moynihan, 2018).

Foundational Performance Management Requirements

As discussed, the GPRAMA is currently the performance management requirement for the federal government. The GPRAMA has many facets to it, but the main structure is threefold: Strategic goals and objectives, Agency Priority Goals (APGs), and Cross-Agency Priority (CAP) goals (“Implementation of GPRAMA Modernization Act Has Yielded Mixed Progress in Addressing Pressing Governance Challenges,” 2015). Each of these three main components has strategic criteria and timelines for reviews. Additionally, the three goal-setting requirements are tied to reporting requirements within the Agency Strategic Plan and the Annual Performance Plan and Report. Strategic goals and objectives are longer-term goals that are specific to major program areas within an agency. These strategic goals are specific, measurable, and describe the agency’s role in achieving the stated outcomes through the program. Strategic goal progress is reviewed by agency heads and OMB leadership annually based on the specific objectives in the strategic plan using qualitative and quantitative data available. CAP goals are revised every four years and focused on priority areas where policies and programs are implemented across multiple agencies, which require the collaboration of efforts to effect program-wide changes to outputs and outcomes. CAP goals are reported publicly via Performance.gov, which is the federal government’s web presence for performance management. CAP goal progress is reported quarterly against CAP goal indicators, targets, and milestones. OMB and the Performance Improvement Council are responsible for reviewing these reports. APGs are focused on goals achievable within a two-year period and are aligned to the agency’s strategic objectives. APGs

may have performance indicators, targets, and timelines for results, but these are not required. APGs are reported publicly on Performance.gov and assessed quarterly by Agency Chief Operating Officers and program office staff based on data-driven reviews to identify programs and needed changes to meet the goals (P.L. 111-352, Jan. 2011; OMB A-11, Part 6).

Small Number of Priority Goals

An important theme in the research of performance management, especially for the successful use of performance information in decision-making, has been the need to obtain leadership buy-in, participation, and understanding as key stakeholders in the process (Moynihan et al., 2012; Moynihan & Kroll, 2016; U.S. Government Accountability Office, 2017). The research points to the leadership at the agency or sub-agency level since most research has shown that performance information is rarely used by the legislature in policy or resource decisions other than as a means to justify a policy decision or political argument (Bourdeaux, 2008; Gao, 2015). However, the research has also shown that in order to effectively inculcate and enhance agency leadership's ability to stay engaged in meaningful ways with performance management routines for decision-making, the number of goals, reviews, and performance data that need their attention must be limited to a small, manageable set of priority areas (Hassan & Hatmaker, 2015; Lee, 2018). This was an intentional design choice of GPRAMA based on research and feedback from prior iterations of performance management at the federal and state - levels ("Implementation of GPRA Modernization Act Has Yielded Mixed Progress in Addressing Pressing Governance Challenges," 2015; Moynihan & Kroll, 2016). This has been shown to be increasingly important as agencies have identified the need to include many of their senior leadership in "performance management councils" since programs require multiple aspects of agency resources, direction, and focus on meeting their increasingly complex strategic goals (Ayers et al., 2014; George et al., 2019). Since this is already integrated into GPRAMA, I will ensure that it is a key component of Framework 1 as well. Notably, agency leadership will have small numbers of priority goals, as shown in Figure 2-3 (Performance Improvement Council, 2019) for their performance management processes, and doubtfully will, the majority of their goals be focused directly on administrative burdens - so I will incorporate this reality into the framework to show how they can still include administrative burdens in a way that doesn't unnecessarily proliferate the number of goals which need to be set, measured, and monitored.



Figure 2-3: GPRAMA Goals Overview

Learning Routines

As discussed previously, in existing implementations of GPRAMA, the successful use of performance information for decision-making has been correlated with learning routines set in place by the requirements. Learning routines are formalized processes where organizations regularly are required to review their progress towards their strategic goals based on their performance information and account for and adjust implementation based on the feedback provided by their performance information. This allows them to learn from the data and course-correct their processes and potentially even their strategy (Moynihan & Kroll, 2016; U.S. Government Accountability Office, 2017; Vought, 2019). Learning routines and positive results from them are improved when leadership is engaged in the process. This does not always need to be based on significant engagement. Rather the research has shown the signal importance of learning routines, and the use of performance data for improvement is key for leadership actions to improve the organization's performance management processes (Lee, 2018; Moynihan et al.,

2012). Importantly, the “how” of this matters, too, since there have been negative results measured when performance management routines are used as oversight and basis for punitive review, as a means for public embarrassment, or have direct impacts on resources and performance standards. These are often associated with the “stat” type implementations (e.g., CompStat, CitiStat), which often require challenging and confrontational public performance reviews where leaders are brought to task for their performance data which is negative. Research has shown the culture created by these interactions inhibits organizational learning, likely because it results in efforts to hide “failures” by gaming the performance measurement systems to avoid negative repercussions (Destler, 2016; Moynihan et al., 2012). Similar results were observed in the PART process during the Bush Administration because many viewed the results of the program and performance scoring as a means for political program and policy control rather than an opportunity to learn from and improve the program (Herd & Moynihan, 2018; United States Government Accountability Office, 2005).

Performance Management and Administrative Burdens

While I will look at many aspects of performance management theory, practice, and research in this dissertation, I will be doing so to inform a practical tool to allow researchers and administrators to apply performance management to measure administrative burdens. I will not be attempting to modify or change existing performance management theory or requirements in the federal government beyond this inclusion of administrative burdens. This is an important point for my research since there is an incredible amount of performance management research showing how the federal government can and should modify its practices, how legislators should use performance management, and how the public could interact with performance management in different ways that would directly impact policies and programs. However, what is more important for this research is how existing requirements and best practices can best be used to include the measurement of the costs of administrative burdens. Using existing requirements, even if they are imperfect, enables researchers and administrators to have a tool that can be implemented now and with existing resources rather than a theoretical tool that will require changed legislation or government regulations before it can be applied.

Measuring Administrative Burdens

As discussed already in this paper, administrative burdens are emerging in importance within public administration theory and administration. Focusing on these frictions can help the government and the public better understand many facets of public programs (Herd, 2015).

These include aspects that are important to the performance management and evidence-based policymaking moments, such as the efficiency of program administration; effectiveness of program implementation and outcomes; transparency of government programs and budgets; open government; and evidence about program outcomes. Therefore, I believe it makes sense to ground my framework in the existing guidance for performance management systems in the federal government as well as the existing theory and research about measuring performance management to identify ways to identify and measure administrative burdens. Performance management requires strategic goals which are specific, time-bound, and measurable. I am going to focus on identifying and measuring administrative burdens with the express goals of reducing them or eliminating them in government programs.

As stated earlier, there are no agreed-upon methods or frameworks, so we're going to navigate what is in existence now and build a generalized framework that can be applied to many types of programs and situations in the federal government. To accomplish this, I will look at each type of administrative burden cost and highlight the operative information from existing research to identify ways to collect and develop performance information for them.

Administrative burden costs are lived experiences of the individuals interacting with the government (Burden et al., 2012; Chudnovsky & Peeters, 2021; Herd & Moynihan, 2018). Therefore, remain cognizant of our measurement criteria when designing performance measurement systems. Additionally, administrative burden costs can manifest differently for different programs and different individuals. Therefore, it will not be possible to compile one performance measurement system for programs but rather a design framework for different programs to apply in their unique circumstances. This is also important because as programs administrators focus on reducing and eliminating administrative burdens, performance measurement systems must be flexible to account for these changes both in the performance information collected and analyzed but also in the experiences that changes can move significant performance management challenges to other areas of the program or have unintended consequences on a different process. This necessitates a multi-pronged approach for measuring different aspects of each program based on each cost, which is also adaptive as performance management reviews identify needs to add additional performance information to other program aspects.

As has also been discussed in this paper, administrative burdens can be caused by myriad aspects of a program's design and implementation (Herd & Moynihan, 2018). Furthermore, administrative burdens are felt differently by differently placed individuals - meaning that they

are unequally distributed (Ali & Altaf, 2021; Herd & Moynihan, 2018). Certain individuals or situated individuals are either more or less able to overcome administrative burdens. Said another way is that these burdens cause more friction for some individuals than others. Research has already shown that often in social programs, those who are most eligible for the program or those who would benefit most from the program are impacted to a greater extent and therefore more likely to not be able to access the program or to access it to the full extent of the program. Additionally, those who have particular mental, physical, or social-economic factors in their lives are often less able to understand, comply, and proceed through many types of administrative burdens (Carolyn J. Heinrich et al., 2021; Chudnovsky & Peeters, 2021; Herd, 2015). These are important points not just for understanding the impacts and distribution of administrative burdens but for thinking about how to identify and measure them within the confines of a program. Measurements of individuals' experiences are difficult without constant feedback mechanisms, so looking for an individualized way to understand and measure administrative burdens based on disaggregation of individuals throughout the program, especially focusing on those who may be unequally impacted, is going to be important for a full understanding and creation of performance information.

“Is the citizen burdened? And what is the benefit [to the burdens]?” These are questions that Judge Richard Posner asks in his econometrics-based legal judgments when determining if the government has gone too far with particular administrative requirements and determining whether they are deemed to be legally consistent with the legislative requirements of programs and policies (Peeters, 2020). In these questions, I find a basis for reviewing administrative burdens quantitatively and qualitatively in a manner that can be applied to performance management programs. For example, having particular outcomes identified based on the legislative requirements and then opening up performance measurement to provide a quantitative measure of the costs imposed in the “how” these goals are obtained for a qualitative understanding of the burdens as compared to the desired outcomes. For instance, a legal requirement to adjudicate a benefit for individuals based on a means-test requires the collection of information to make that means test. However, there are many ways to collect that information. On one end of the spectrum, requiring individuals to provide certified paper documents from their employers can impose significantly more compliance costs than using administrative data collected by the IRS to automatically provide validated income data. Now that I have two measurements, a qualitative assessment can be made about the benefits of one method as compared to the other. When the employer-certification process can be viewed as

“reducing fraud,” the same can be said about taking already validated IRS data. This linking of administrative data has no anti-fraud value, but the compliance costs are nearly non-existent for the citizens, so the choice becomes obvious when the goal is to reduce administrative burdens. Not all examples will be straightforward, but the theory of these comparative choices based on the performance information is quite simple and becomes easy to explain how and why certain decisions are made.

Based on the above understanding of administrative burdens, our framework must provide different performance measurement techniques which can be adapted and applied to different costs. Additionally, there are various ways to measure different administrative costs - allowing a menu of options for administrators to choose from depending on their particular program needs. In creating this menu, I was helped by looking at the different aspects and questions to be asked when identifying potential administrative burdens and costs. To this, I will apply performance measurement techniques to further link these cost identifications with the appropriate performance measurement protocol.

Herd and Moynihan (2018) began outlining tools and concepts to approach the identification, or administrative burdens especially focused on the collection of feedback about processes and impacts on individuals, as outlined in Table 2-3 (Herd & Moynihan, 2018). Other administrative burden researchers have mainly focused on case-study analyses of particular programs or implemented instances to qualitatively discuss the presence, magnitude, and impacts of administrative burdens. However, these research methods are not easily adopted by program administrators, legislators, and researchers wishing to look across the fields of administrative burdens rather than specific programs.

<p>Take-up</p> <ul style="list-style-type: none"> •What is take-up rate for eligible beneficiaries?
<p>Inequality</p> <ul style="list-style-type: none"> •Does take-up rate vary accross populations?
<p>Learning Costs</p> <ul style="list-style-type: none"> •Is it easy for potewntial participants to: <ul style="list-style-type: none"> •find out about the program? •establish they are eligibile? •understand what benefits are provided? •learn about the application process?
<p>Compliance Costs</p> <ul style="list-style-type: none"> •How many questions and forms are there to complete? •How much documentation is needed? •Does the participant have to input the same information multiple times? •Is the information sought already captured via administrative data? •Is it possible to serve the person in a less intrusive way, such as phone rather than in-person interviews? •Do applicants have easily accessible help? •How freaquent is reenrollment? •How much time must people commit to the process? •What are the financial costs?
<p>Psychological Costs</p> <ul style="list-style-type: none"> •Are interactions stressful? •Do people receive respectful treatment? •Do people enjoy some autonomy in the interaction?

Table 2-3: Administrative Burdens Diagnostic Questions

As derived from the above identification concepts, it becomes clear that performance information collected about administrative burden costs can be done through surveys to program participants, applicants, or those who are simply trying to understand programs. However, in some instances, it is going to be easier, more accurate, and less resource-intensive to identify program administrative data, which can measure administrative burden costs rather than relying on surveys or creating new information collection processes. As discussed previously, one important tool which has been shown successfully in many public programs to increase take-up and participation (because of reduced administrative burdens) is automatic enrollment (Grimmelikhuijsen et al., 2017; Herd & Moynihan, 2018; McIntyre et al., 2021). This is

typically done through the use of administrative data when available. Other concepts of potential ways to reduce administrative burdens can be illuminating to help our understanding of how I would measure administrative burdens through performance measurement processes. For example, approaching programs through the lens of automatic enrollment can help frame the understanding, measurement, and discussion of any policies or procedures which impede automatic enrollment or that require manual or intervened processes. Based on the implementation of automatic enrolments, the program administrators can use the performance information about application processing times, rates of requests for additional information or verification, times of applicants spent interacting with the program application processes, and the rates of administrative denials (because of missing application requirements) as a means to measure administrative burden costs as compared to the same measurements for an automatic enrollment process.

<p>Learning Costs</p> <ul style="list-style-type: none"> • Make information and application processes accessible, online, and easily searchable • Use simple language, understandable to the target audience, and multiple language options • Provide calculators to estimate benefits • Communicate choices in simple terms, such as categorize options in gold, silver, bronze • Provide reminders (text, mail, phone) • Use outreach campaigns to shape the public perceptions of a program and provide information
<p>Compliance Costs</p> <ul style="list-style-type: none"> • Integrate multiple forms with similar questions into one • Allow multiple options for documentation to be accessible • Allow standard deductions (such as medical expenses) rather than require extensive documentation • Use administrative data to verify status and pre-fill forms • Allow online or phone interviews/submission in addition to in-person options • Solicite responses and documentation only to degree necessary to perform the task, or legally mandated • Provide enrollment help: auto-completion, phone, or in-person help, either via public actors or third parties • Make administrative centers geographically accessible • Provide help outside of the traditional nine-to-five bracket • Allow third parties to enroll at point of contact (hospital, churches, community groups)
<p>Psychological Costs</p> <ul style="list-style-type: none"> • Send messages of welcome and inclusion to potential participants • Build a cultural ethic based on respectful interaction and an ethic of help • Give individuals the opportunity to articulate their story and provide feedback • Offer participants an Ombudsman or other clear mechanisms to express dissatisfaction

Table 2-4: Techniques to reduce Administrative Burdens in Program Implementation

Since we do not have a rigorous tool or method for measuring administrative burdens, I begin to see that by focusing on identifications questions for administrative burdens and potential changes to processes and policies to reduce administrative burdens, I can begin to identify what

performance information will be useful to create a performance measurement process of administrative burdens in programs. That being said, since administrative burdens are the experienced frictions of individuals, I also know that there will need to be thought about how to collect performance information that measures on an individual or group basis rather than aggregate across the entire population. Additionally, from the literature, it is clear that administrative burdens are subjective across different programs based on several variables (e.g., public perception of program participants, perception of the programs, the focused policy objective of political leadership), and therefore performance measurement systems may need to be specific to each program and even different among different implementations of the same program (Ali & Altaf, 2021; Heinrich et al., 2021; Herd & Moynihan, 2018). Once the performance measurement systems are built for programs and implementations, though, I am confident that I will gain the ability to compare administrative burden costs across programs and time by generalizing specific program measurements to higher-level concepts in performance measurement processes. For example, even though the specific causes of compliance costs in a program may change over time, I will be able to generalize to compliance costs in the application process and compare those costs across time, including changes in processes and policies as well across program implementation. This ability will be key when wanting to test the impacts and outcomes of programs based on the changed levels of administrative burdens to study and evaluate our efforts.

Big Data, Artificial Intelligence, and Machine Learning in Public

Administration

In this section, we're going to look at an overview of big data, artificial intelligence, and machine learning which have been revolutionizing many areas of our lives and the world in recent decades. We're also going to look at how these technologies and techniques have been supported and integrated into public administration and the challenges and prospects of this intersection. This is distinct from implementation in the private sector because of the nature of the challenges, the required features of public policy and administration, but also because of the potential outcomes and how they might affect our lives, our society, and our system of government. There are both risks and potential rewards with significant impacts. Rather than pile up a list of risks to make excuses as to why administrators cannot or should not use them in public administration, we're going to take a balanced approach seeking to build a framework to show how to accurately and thoughtfully design and implement solutions in order to benefit

public administration in the form of reduced administrative burdens. But also how to do so in ways that mitigate potential risk and take into account the unique challenges and risks in the public sector.

In this section, I will cover the basic definitions and attributes of artificial intelligence, big data, and machine learning. I will also explore the current techniques, frameworks, implementation policy, and best practices. An important and emerging area will also be covered that is looking into the unique challenges and considerations for ethical, legal, and accurate artificial intelligence based on the emerging focus on and concern about biases and accountability. Additionally, I will focus on some of the emerging requirements and considerations in the legal and regulatory sphere for artificial intelligence in the United States federal government. Let's begin with definitions and examples of these techniques, especially focused on public sector problems and implementations.

Definitions

Big Data

The concept of big data has been around for decades, but it has found more fame as it began fueling revolutionary AI breakthroughs in recent decades. The growing computational power of computer processors and the plummeting costs of data storage have been combined with novel techniques such as machine learning, data mining, and new data visualization techniques to drive the big data revolution in many areas of our world (Desouza & Jacob, 2017; Giest, 2017). There is some disagreement about exactly what the definition of big data is but there is convergence around several key factors: the data is too large, too raw, and unstructured to be kept and used in traditional relational database structures (Kim et al., 2014). Others have used a varying number of "V's" to define big data: volume, variety, veracity, variability, and velocity (Giest, 2017). But there is also the official U.S. government definition of big data as enumerated by the National Science Foundation: "Big data sets are large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, e-mail, video, click streams, and/or all other digital sources available today and in the future" (White House, 2014).

The federal government in the United States is actually one of the earliest collectors and users of big data as it is known today, even before there was technology that enabled the current uses of big data. In fact, much of the current landscape of big data has been shaped by government collection, dissemination, and utilization of data. For example, the Census Bureau

has long been collecting and analyzing data sets that meet several of the “V’s” of big data. Other examples include the National Oceanic and Atmospheric Administration made large weather datasets available to the public, and the Department of Defense built and then made publicly available the Global Positioning System (GPS), which was combined with geospatial data from several other agencies in the 1990s (Desouza & Jacob, 2017; White House, 2014). Additionally, the Department of Labor’s Bureau of Labor Statistics historically collected and analyzed data which is rightly classified as big data even though it has been analyzed using traditional statistical methods. This happened even at times when most in the academic community and the private sector focused on sampling rather than full population data analysis. It is also worth noting that these examples of government data collection and analysis, as well as making their data available to the private sectors, shaped much of the landscape of big data uses currently. For example, in the financial sector, the mortgage loan industry and the insurance industry benefitted from government data collections for much of their statistical modeling of customer bases, risk scoring, and actuarial analysis. It’s likely that without government investment and sharing of data, these industries would not have been able to make similar investments in their nascent data stages (Giest, 2017).

The U.S. government invested heavily in big data, paving the way for the private sector. There is no argument that now the private sector has developed and deployed methods to use big data and collect big data in ways that have revolutionized many industries and are changing our culture. As seen in other areas, now the government is playing catch up and trying to implement similar techniques and technologies of big data analysis and application in the public sector (Desouza & Jacob, 2017; White House, 2014). However, it is important to look at the similarities and differences between the public and private sectors to fully understand and account for the constraints, limitations, as well as risks, and mitigations that are important in understanding the government’s use of big data.

The potential benefits of big data in the government are many, especially in light of big data’s ability to provide fertile ground for machine learning. Machine learning has several use cases, including forecasting, categorization, and regression analysis. Specifically, by harnessing big data and machine learning, private sector companies are able to target their advertisements, predict likely customers for certain products, better predict customer churn (and attempt to keep them from leaving), and identify likely future surges or lulls in business (Agrawal et al., 2018; Tetlock & Gardner, 2015). The government has other potential use cases for big data and machine learning. The potential benefits of machine learning include more effective decision-

making, more efficient decision processes, and less biased decision-making (Maciejewski, 2017). There have already been using cases in national security, the military, and anti-fraud measures at the IRS (White House, 2014). In other areas, the government can leverage big data to more effectively and efficiently administer government benefits, decreasing misuse while automating much of the application processes (Moynihan et al., 2015). These implementations are what I will focus on in this work, with the goal of identifying machine learning processes that target the reduction of administrative burdens.

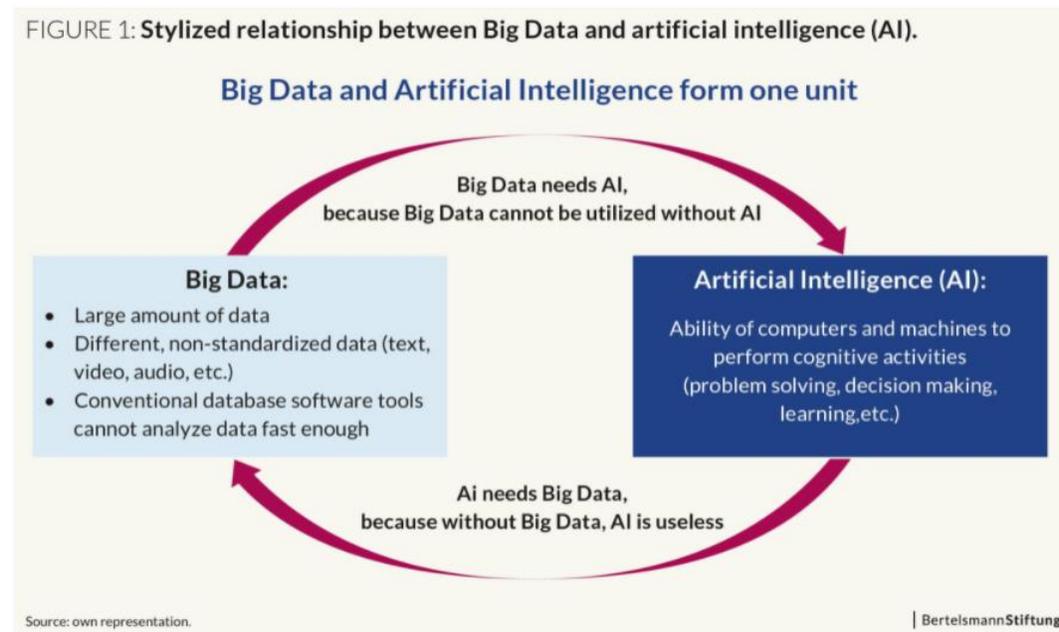


Figure 2-4: Relationship between Big Data and Artificial Intelligence

As depicted in the figure above, the relationship between big data and artificial intelligence is one of fuel and furnace. Big data is a critical component of artificial intelligence techniques, including machine learning. Although big government data has been processed in traditional ways in the past, there exists promising new implementations and use of this big data to fuel new artificial intelligence implementations in a variety of areas. As discussed, the core of many big data definitions is the need to process and store it in novel and automated ways. One of the ways to process big data for useful purposes is artificial intelligence. In fact, in many ways, artificial intelligence and machine learning as they are known now would never be possible without the existence of big data that drives them and make them possible. Therefore, it is somewhat symbiotic because big data fuels artificial intelligence techniques, but in many instances, artificial techniques are needed to process and use big data for most purposes (Agrawal et al., 2018; Desouza & Jacob, 2017; Kim et al., 2014).

Artificial Intelligence

Artificial intelligence (AI), like big data, has suffered from broad definitions and overuses that have clouded many of what the term actually represents. Many definitions of AI focus on the criteria of machines mimicking natural animal and human intelligence, especially through the ability to perceive and learn from the environment and taking actions that raise its chances of achieving its goals (Calo, 2017). In 1950 Alen Turing published a proposition to set aside the idea of machine intelligence as a goal or measurable benchmark. Instead, he argued, what mattered was the “manifestation” of intelligence of the machine since this was observable and measurable. He proposed an “imitation game” that would demark the threshold for an “intelligent” machine. Put simply, what became known as the “Turing Test” is when a machine operates so expertly that observers cannot distinguish its behavior from that of a human (Kissinger et al., 2021). This has subsequently become the goal and test of more forms of artificial general intelligence (AGI) but does not represent a clear definition for all forms of AI.

In the Federal government, there have been several recent definitions that help clarify the term. The National AI Initiative Act of 2020 and the John McCain National Defense Authorization Act of FY 2019 defined it as including:

“(A) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.

(B) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.

(C) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.

(D) A set of techniques, including machine learning, that is designed to approximate a cognitive task.

(E) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting.”

Artificial intelligence, unfortunately, often brings up the idea of benevolent or malevolent human-like robots and machines which Hollywood has helped people think about turning into either destructive forces in the world or some of our strongest companions. Both ideas within science fiction are not necessarily wrong, but we are a far way from those potential realities. In fact, most current iterations of AI are either so narrow that they can only perform or assist with very specific tasks under specific conditions or, more generally, perform tasks more poorly than many human children. No one is quite sure what the development timeline of these more advanced AGI applications is, but many are convinced, based on historical progress, that they are inevitably coming (Agrawal et al., 2018; Calo, 2017; Kissinger et al., 2021).

Artificial intelligence is hardly a new field, as it began in the early 1900s in several scientific fields, including mathematics, philosophy, and information theory. As digital computers became a reality in the 1950s, the idea of artificial intelligence became a goal and desired outcome for many researchers and engineers. What was generally pursued then and what became the popular understanding of artificial intelligence is computers that could understand, reason, and respond much like humans. This has become known as general AI as opposed to the field of narrow AI, which is focused on smaller, more discrete tasks (Agrawal et al., 2018; Jobin et al., 2019).

The government began to invest in AI research and development heavily in the 1960s and 1970s, which further grew available schools and facilities focused on understanding, developing, and implementing AI techniques (Kissinger et al., 2021). Many people believe that it was a matter of a short period of time until AI abilities matched those of humans and fulfilled the promised goal of general AI. However, even though significant progress and displays of AI implementations were shown in the early decades, several began to worry about the ability of their methods to reach the goals of general AI. A robot that was able to successfully play and win games of checkers and chess through significant programming and computation of specific moves was hitting the reach of computer memory and computational power. Additionally, these required a significant amount of accurate programming to achieve in ways that were not scalable or sustainable for general AI goals. Because of these factors, ‘AI winter’ set in during the 1980s and 1990s as much funding moved elsewhere (Agrawal et al., 2018; Calo, 2017). However, several academic and applied researchers continued to work on the theory and methods that would allow different techniques to make significant breakthroughs decades later. Like many scientific endeavors, these continued theoretical efforts began to combine with technological improvements in data storage and process techniques as well as computer computational

improvements, which have combined to more than thaw these AI winters and begin to make rapid progress in AI development and implementation (Calo, 2017; Kissinger et al., 2021).

Machine Learning

One of the main reasons the AI winters ended and AI has made such progress in the past decades is because researchers made a paradigm shift in their efforts. Instead of trying to program computers with intelligence, relying on the expertise of computer scientists and domain experts to provide the needed values and understanding for machines, computer AI researchers developed methods to allow the machines to train themselves, known as machine learning (Kissinger et al., 2021; Raschka & Mirjalili, 2019). Machine learning is a subset of AI, which is more narrowly focused on a method to teach computers to learn typically narrow tasks and does not focus on many of the other areas of AI typically thought of as AGI. Scientists reached this conclusion after they realized the limitations of the programmer/expert approach were always going to be narrowed by the number of lines of code and experience that could be fed into a computer. Instead, they proposed an approach that would give the computer parameters to understand the inputs and the desired outputs but then train itself, known as machine learning. As discussed, AI tends to refer to any processes where computers “act like” a human, including functions and thinking processes. These complex processes and behaviors have become known as artificial general intelligence (AGI). This is in comparison to narrow artificial intelligence, which is programmed to perform a single task. Many implementations of narrow AI have been achieved through machine learning. Machine learning is typically thought of as a narrow AI because the methods and tasks associated with it tend to be niche areas that few would consider similar to AGI. Instead of a humanoid AI robot, there are instead powerful computer programs that can teach themselves to become masters of a certain game, translation ability, image recognition, or other tasks which are being applied to many different areas of our lives (Agrawal et al., 2018; Burrell, 2016; Kissinger et al., 2021). Much of what we interact with on a near-daily basis is the result of narrow AI machine learning processes. So, to keep the explanation linked, big data fuels much of the use of artificial intelligence, and machine learning is a subset of artificial intelligence. The key features of machine learning are, in fact, training computers to learn themselves, typically based on the processing of big data sets through algorithms programs to learn narrow rules and tasks from those big data sets.

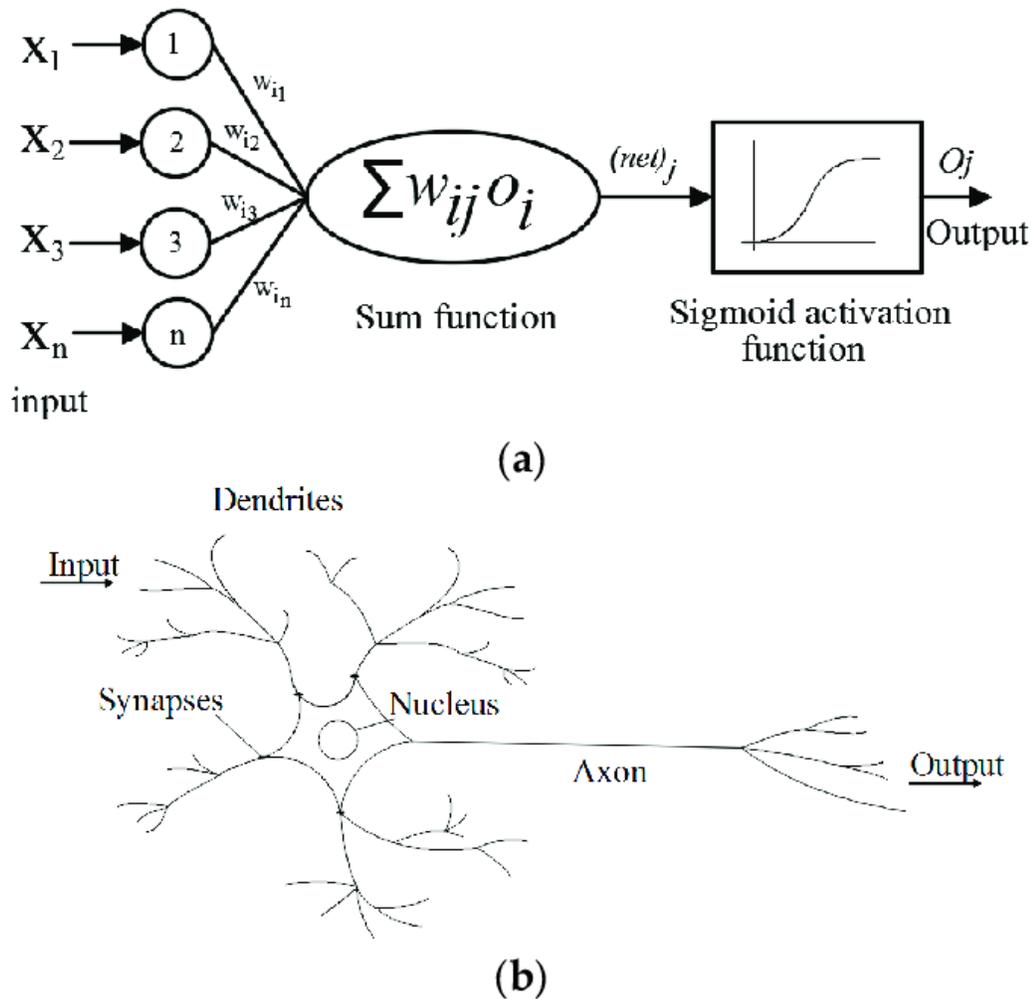


Figure 2-5: Animal and Computer Neurons (Raschka & Mirjalili, 2019)

One additional key feature of this breakthrough was the development of computer algorithms that are modeled on the neural networks found in human and other animals' brains. The comparison is shown in Figure 2-5, where a rough approximation of human brain synapses to ML algorithms is depicted (Raschka & Mirjalili, 2019). The theory and insight worked well and have led to a plethora of narrow AI breakthroughs which have become applicable to many areas of our lives and economy - sufficiently thawing the AI winter and increasing research in the public and private sectors around the world. Programing computer algorithms that could mimic nodes and synapses of brains, as shown in the above figure, to build computer neural networks that would allow them to learn and train themselves for tasks specified by the programmers. Instead of programming specific “if-then” type statements that form rules into a computer program, the computers were programmed with learning mechanisms that focus on inputs and target outputs with carefully formulated feedback mechanisms to identify when their

learning is correct or not. For example, prior to machine learning, if programmers wanted a computer to be able to recognize images of cats, they attempted to program concepts of a “cat” into logical lines of code. They would attempt to logically program the concept of whiskers, tails, noses, and ears in ways distinct to cats and not other animals. You begin to see the difficulty in trying to account for all of the positive and negative examples, especially as you realize that an image of a cat can take many forms and positions depending on how the cat is sitting, lying, curled up, jumping, or otherwise active. And you try to develop concepts of “cat-ness,” which do not apply to other animals similarly, to be able to help a logical operation of computer programming make those distinctions. This is an exceedingly difficult thing to do from an inductive perspective. I will approach this problem through the deductive method through machine learning in a bit to compare the approaches.

Machine learning takes advantage of recent breakthroughs in both increased computer storage of data (which has several benefits for machine learning) and compute power developments to allow computers to train on many more data than previously possible under the old paradigm, much faster and more broadly than possible previously. These computer neurons were first developed in 1956 by John McCarthy, but they were not able to be tested with much rigor until the data, storage, and computing power developed sufficiently in recent decades. These computer neurons have been fueling the rapid growth of narrow AI-based machine learning implementations. Even recent advances in artificial neural nets and deep learning models have built on this foundational design to discover these advances.

It’s important to understand the conceptual mechanisms behind machine learning. Unlike in previous programming, each logic was programmed and understandable. With machine learning, we gain the significant advantages of the computers’ ability to learn tasks, make novel connections and predictions, and the automation of many aspects of both programming and operations which were not possible previously (Raschka & Mirjalili, 2019). But we also lose the ability in many instances to completely understand the connections the computer has discovered between the input data and the prediction outcomes. Even when they are very accurate, the computer finds those connections in ways that are confusing or incomprehensible from an inductive logical standpoint. This has been beneficial in many aspects, including health-related applications where machine learning has discovered important new chemical combinations to fight infections or concerns, but humans and computers do not understand the underlying causal properties of “why” they work. For some, this is incredibly problematic, as it raises concerns about unintended consequences, agency, oversight, and transparency (O’Neil, 2016). However,

this can also be seen as a new shift towards thinking and discovery, which is aided through the mechanisms and use of machine learning and human partnerships to do research that was previously impossible through traditional methods.

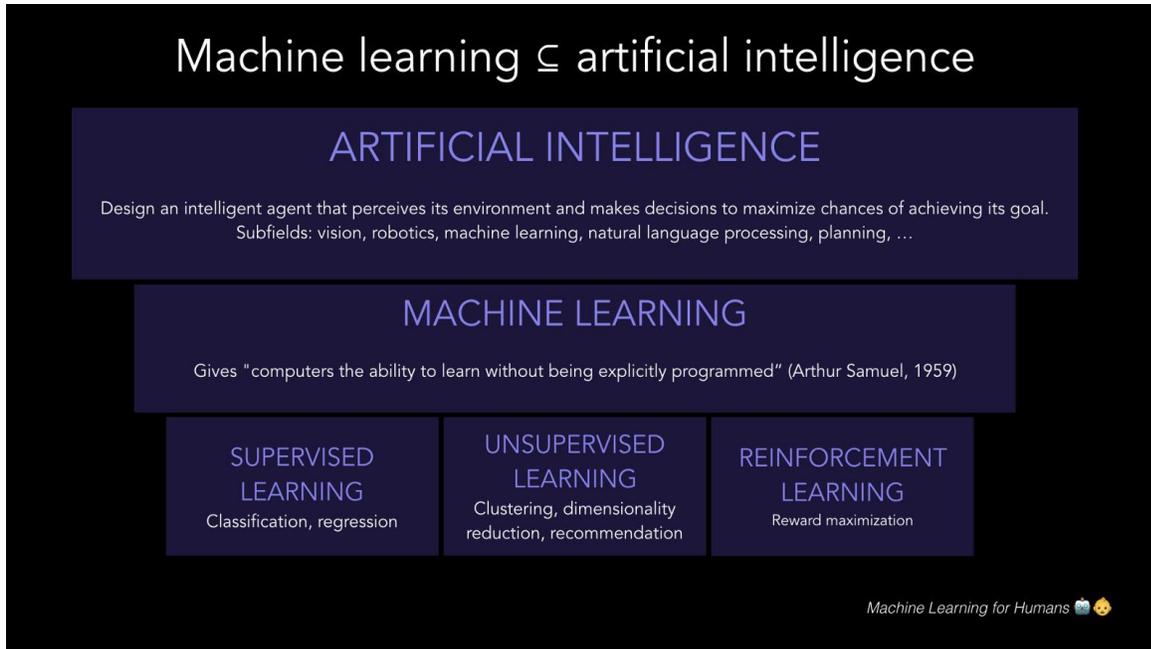


Figure 2-6: Machine Learning vs. Artificial Intelligence (Wolf et al., 2020)

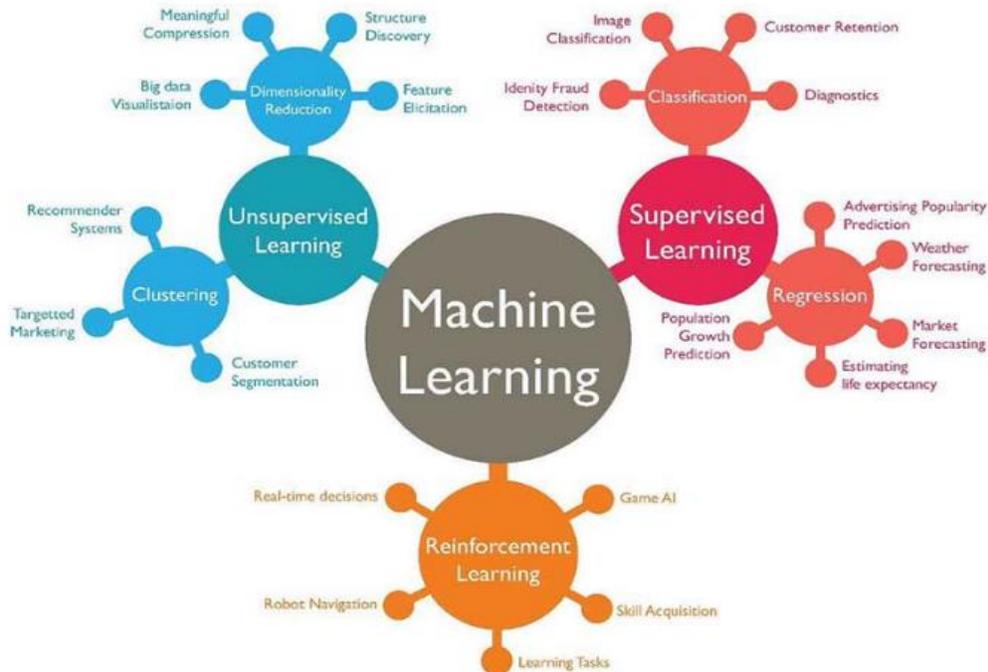


Figure 2-7: Types and examples of Machine Learning

There are three general types of machine learning currently, as depicted in Figure 2-6 and Figure 2-7: supervised learning, unsupervised learning, and reinforcement learning. Each of these is beneficial for particular tasks, which have a range of different potential applications, as

shown in the figure above. Supervised learning is where the computer receives data which is split into testing and training data. The training data that the machine learns from is labeled to provide the initial inputs for the learning process (Burrell, 2016). For example, a dataset of images of cats and other animals which are labeled as “cat” or “not a cat” is used to train a model to identify images potentially of cats. Without these labels and “supervisions,” the machine couldn’t independently formulate the understanding of “cat,” nor does the machine understand more than how to predict if an image is more or less likely to contain a cat. Through the learning of many labeled images, the model formulates associations of data points that are correlated to the presence of a cat in an image, and then this trained model can be applied to a new dataset of unlabeled images to predict which contain cats.

Unsupervised machine learning is where the data set does not have the labels of the target variables the computer is learning how to predict. Instead, the computer processes the data to determine patterns, rules, or correlations on its own which may or may not provide useful implementations. Importantly, in some implementations of unsupervised learning, you do not always know what the target variable or output of the computer model should be (Raschka & Mirjalili, 2019). This is true in particular applications of cluster analysis, where the unsupervised learning model is finding potential divisions of distinct groups within the data set based on attributes that may or may not be intuitive when reviewed later. This method can be useful in a number of applications, but as compared to the cat prediction example, an unsupervised model could be given thousands of pictures of animals to train on and then be used to create different categories of animal photos. If done correctly and accurately, one of the categories is likely to be “cats” - but derived from learning and separating the data rather than learning through tagged images. Obviously, the computer would have no concept of a cat, but learning the similarities and differences in the data could distinguish images that contain cats from those which contain dogs or birds. Unsupervised learning has become adept at providing incredibly useful and novel patterns, correlations, and predictions but without always a discernible causal explanation of these correlations, which can be difficult in certain implementations. However, they have also provided distinctly unique and useful outcomes that humans may never have been able to develop without their aid (Baylor et al., 2017).

The third general type of machine learning is reinforcement learning or partially-supervised learning. It is similar to unsupervised, where it does not require labeled data for training but instead interacts with the world during its training and/or deployment and then uses feedback through that interaction to train towards a target cumulative reward parameter (Raschka

& Mirjalili, 2019). Reinforcement learning is a useful adaptation where the computer model needs to interact with the world and be able to learn additional information, patterns, and rules based on these interactions. The unsupervised learning model is seeking the optimal or near-optimal solution based on the environment that it knows, where this optimal outcome or policy is not understood or modeled. This has been especially useful in scenarios that favor long-term versus short-term reward payoff, such as gameplay (Go, checkers, backgammon), operational challenges (elevator optimization, telecommunications), and locomotive controls (autonomous driving, flying, robotic control) (Gerrits, 2020).

Implementations of Machine Learning

Private Sector

As usual, while the government supported and funded the early development and implantation of big data, artificial intelligence, and machine learning, it really gained steam and became a part of our lives with the advent of many private sector implementations and use cases. From how people connect with individuals and groups on the internet, search engine results, the advertising people see, the costs of many services, including health care, car, and life insurance, discounts for travel and hotels, translation software and apps, spam blocking for our email and phone calls, weather forecasts, and many other areas of our lives that are direct or indirect applications of machine learning are ubiquitous (Agrawal et al., 2018; O’Neil, 2016).

The private sector has countless implementations of machine learning from which to draw examples. Early success was found in ways to increase the effectiveness of internet searches. Early internet search engines were successful because of page rank algorithms which were elaborate employments of coded logical statements and feed by time and computing-intensive reading of the internet and linking of behavior between pages. As the available information and web pages grew, these early methods became less sustainable (Agrawal et al., 2018; Dwivedi et al., 2021). Additionally, as the improvements in machine learning spiked, it became much more reasonable to program search engines to teach themselves how to best search and display webpage results. This benefited from the growing number of users and web searches, which amassed massive data sets for the ML models to train on and be provided feedback (in the case of clicks and page visits) about which predictions were more accurate.

Additional public sector use cases were developed and implemented with surprising success. Amazon is another example of an entire industry being “disrupted” through the use of an ML-focused business model, advertising, and supply-chain management. Amazon developed

ML models to help recommend and predict purchases, inventory, and supply chain improvements. They actually lost money for many years because their goal was to collect massive amounts of data to feed these ML models rather than simply focusing on profit margins. They have leveraged this data and growing amounts of internet interactions to develop new lines of commerce, products, and supply chain models, which has led to massive growth in profits and market share (Agrawal et al., 2018).

Other examples include historically industries such as insurance which relied on traditional statistics to predict outcomes sufficiently to find profitability between product fees and payouts. The insurance industry has further benefited from ML models that can learn from massive historical data as well as datasets that were not able to be effectively integrated through standard statistical practices (such as information about road safety, health outcomes, and web search datasets) to drive more accurate predictions which allow them to offer their plans in ways that remain profitable (Agrawal et al., 2018). Additionally, the automobile insurance industry adopted even more specific instances of data collection from the Internet of Things (IoT), enabling devices to predict and present incredibly personalized predictions and products. You can now get an IoT device for your car to help feed ML models about your driving, which directly impacts your auto insurance plans and rates, or you can link your wearable fitness trackers to your health issuance information to drive plan changes based on your activity and lifestyle (Kissinger et al., 2021; O'Neil, 2016).

More examples surround predicting what customers will buy and who will “churn” by unsubscribing or leaving in order to target them for retention techniques. These have improved the success of advertising, sales, and other efforts. It has created an entire industry around personalized advertising which uses ML techniques to target ads and sales of products to specific individuals based on their data instead of larger, general audiences thereby increasing the effective return on these efforts (Agrawal et al., 2018; O'Neil, 2016). Additionally, many other areas of commerce have been impacted in ways that allow us to fundamentally change how the retail sector approaches business.

Similar new products and markets spring up through ML-fueled advances in areas beyond retail. Internet security developed both predatory ML-based algorithms to target individuals online as well as the defensive measures which help identify, predict, and protect us from cybersecurity issues, fraud-based spam email, internet transaction fraud, and many other areas much more seamlessly than without ML techniques (Berendt & Preibusch, 2014; Dhar, 2020). Research has also benefited from the ML revolution in ways that are not always

transparent. Astronomy is one example where there has been a fundamental shift in techniques. Instead of collecting visual, radiographic, light spectrum, and other inputs from a variety of satellites manually, the fields have moved to broad-spectrum data collection and storage of celestial phenomena, which can then be mined and analyzed through ML techniques to run experiments, research understanding, and build complex theories and models of the universe (Raschka & Mirjalili, 2019). Many other areas of research have similarly moved towards this type of approach.

Medical research and healthcare is an especially interesting example because ML techniques have not just changed the way that most medical research is being carried out, from chemical analysis of new drugs and drug techniques to understanding complex relationships in human anatomy, physiology, and health. But also it has changed the practice of medicine as ML has been fueling assistive technologies, which are creating much more accurate and accessible processes to identify and diagnose health issues than manual methods of the past. This has been surprisingly true in radiography, where ML models have become better at spotting cancers and other anomalies than trained healthcare professionals can be without. Similar advances are happening in other image-based ML work, such as health problems with eyes and eyesight based on cell phone camera images and a range of other techniques (Agrawal et al., 2018).

ML processes that have fueled the advent of accurate and cheap Natural Language Processing (NLP) have also brought products, services, and processes that have revolutionized the success of computer language translations, voice commanded software and hardware, voice transcription, and a growing area of computer assistants which can help us organize our lives through Iota devices such as home systems, cell phones, and our cars. NLP is fueling many other areas of research, including how people interact with the world and each other and how data can be leveraged in formats that were once much more resource-intensive to analyze, such as books, road signs, text in pictures, and even hand-written notes. Not only does NLP help us interact with these data more easily, but it has also helped us use ML to analyze, interrogate, and even produce data from these different data sources. Some examples of advanced ML-NLP uses have shown how computers can generate logical and accurate text, conversations, and analyses after processing text (Wolf et al., 2020). It is in these ML techniques from the private sector that the public sector may make many effective improvements to policy implementation, which lower administrative burdens.

Public Sector

The public sector has long lagged behind the private sector in adopting, implementing, and using new information technology and computer science techniques and infrastructure. In some ways, this has to do with limited budgets, long and methodical requirements for procurement and contracting, as well as hiring, retention, and remuneration for public sector employees (Anna-Katharina Dhungel et al., 2021; Calo, 2017). These specific issues in depth are out of scope for this research, but I do need to understand and acknowledge them as reality when we look at the emergence and use of ML in the public sector. Often, top technologies are out of reach from public organizations for several reasons. One, public sector agencies require more secure and reliable infrastructure than the public sector. This typically means using technologies a few versions behind the more up-to-date version because there has been time to make certain it is secure, but also it is more stable than the new implementation, which still has bugs being worked out through early adopter use (Gerrits, 2020).

Additionally, the demographics of the public sector IT professionals are different of the private sector. Many have explored the reasons for these, but several include the balance between remuneration and mission. Some are motivated more by public service motivation than larger paychecks and the work culture of the private sector. There are also differences in benefits, job stability, lifestyle, and retirement opportunities that impact these differences (Dwivedi et al., 2021; Katzenbach, 2021). The public sector also typically lags behind the private sector in technology software and infrastructure advances. Therefore, individuals who are most focused on cutting-edge technologies might favor a private sector or research sector career instead. Another aspect is the thoughtful and bureaucratic processes of government hiring. The federal civil service system is exceptional as compared to the private sector, but focusing on this as a comparative example is useful. Data scientist as a job has been growing rapidly for the past two decades in the private sector. Data scientists often have a mix of education, training, and experience in statistical methods, computer programming, and domain knowledge which allow them to design and implement advanced analytics, including machine learning, in specific areas they are working on. However, the Office of Personnel Management (OPM) just released data scientist position descriptions and a job category in 2021 and 2022, respectively. While the public sector has had a growing appetite for this skill and knowledge in the federal workforce, it takes time for the hiring processes to catch up and make this clearly possible. Still, time will tell how newly hired data scientists will fare as measured by recruitment and retention in the federal government as compared to the private sector.

Benefits of Public Sector Use of Machine Learning

In many instances, our current statutes and regulations require the government to make predictive decisions that directly affect citizens, albeit to varying degrees. Some of these decisions pertain to the eligibility for benefits, but many others are within the bureaucratic encounters interaction type, where the government is imposing limitations. These are instances when the government is deciding whether individuals will receive punishments, whether individuals' civil liberties will be limited, and many other areas (Berman, 2018). The repercussions of these decisions can vary greatly, from minor inconveniences such as having your bags checked at the airport to more impactful decisions which deprive a citizen of their civil liberties by placing them in prison. Criminal Justice has multiple examples of applied ML, including decisions about where and when to send police patrols based on predictions about where crime is likely to occur; or instances when determining whether to allow for the posting of bail by an individual awaiting trial by predicting how likely it is that the individual will appear for their trial and how likely is it that they will commit another crime, especially a serious crime involving violence (Agrawal et al., 2018; Berman, 2018).

Assuming that government deciders are doing the very best they can in these situations, these deciders still have to deal with the fact that they cannot have perfect information about the situations and individuals they are required to make decisions about. Additionally, behavioral science research has shown the depth of individuals' biases, whether they are known biases or unknown biases in making decisions (Ariely, 2009). Further, individuals are not very good at making predictive decisions about future events based on statistics of past events, absent a tool to assist them (Silver, 2012). Recent research has also suggested that the success rates of even experienced individuals making predictions are often not better than random chance (Tetlock & Gardner, 2015). Combining these limitations and risks with the requirements of our system to make these important adjudicatory decisions gives rise to the need to establish better ways to inform these decisions. Proponents argue that better data is needed to be able to make better decisions (Heinrich, 2007). Others are now arguing that government decision-makers also need tools to use better data to make better decisions and that big data and predictive analytics can achieve these goals (Desouza & Jacob, 2017; Gamage, 2016; Höchtl et al., 2016; Lane, 2018; Maciejewski, 2017; Mergel et al., 2016; Rogge et al., 2017).

There are some emerging proposals for use cases of ML in programs where citizens are seeking services or benefits from the government. Many of these seem to be focused on finding fraud or misuse in programs, but a few are looking at ways to make the application for and

administration of benefits and programs more efficient and effective, especially from the perspective of the citizens being served (Giest, 2017; Kim et al., 2014; Lavertu, 2016). Beyond the ever-present search for fraud and abuse in programs, government administrators need a different approach if they are ever going to leverage predictive analytics to make government benefit programs more efficient and effective.

As discussed previously, ML has numerous implementations which can be used for a growing variety of purposes. A consistent theme of ML design and implementation is to use ML to create more accurate, effective, and efficient processing of data, especially when able to replace manual methods involving human beings. This overall theme is consistent with much of the public sector administration goals and builds on the existing processes and theories to develop IT-based public administration solutions. Many researchers and administrators were focusing their work on specific areas of public administration and applied ML. In this work, I want to acknowledge these potential benefits but will not go into depth about these areas other than to enumerate use cases as they might be applied to the reduction of administrative burdens.

Risks of machine learning

In addition to the above benefits to public sector use of ML, there are also risks and challenges that must be adequately considered. Risks exist in implementing any ML application, but in the government sphere, the risks increase as the outcomes potentially have greater effects on an individual. The risks of false positives and false negatives are heightened. Additionally, these risks are all but certain. For example, no predictive algorithm is perfect, and each has the potential for non-random and random error. Even if a machine-learning algorithm was accurate 99 percent of the time, applied to 300 million citizens, you would still have 3 million false positives or false negatives. In recommending movies on Netflix, that may be an acceptable error rate, but when recommending whether to launch a surveillance program or arrest a citizen, people might think it is not acceptable. In other instances, a predictive algorithm can mistakenly be programmed with inherent discriminatory biases on many grounds, such as race, sex, or sexual preference, which do not represent causal factors for an outcome (Berman, 2018; MacCarthy, 2018).

Because government agencies are, in many instances providing a service that is outside of the private sector “market economy,” our society cannot rely on market forces to correct for these risks and errors. Policies enacted by legislatures are either carried out by publicly funded institutions, public-private partnerships, or by private companies contracted and directed by government staff. This can lead to several types of government inefficiencies. One instance is

highlighted by principal-agent theory, which states that managers and employees have different goals (i.e., the employee attempts to do as little as possible and the manager uses different means to achieve their goals). In the public sector, this leads to the creation of hierarchy and rules aimed at achieving the outcome with the minimum costs. However, due to the public lack of ability to monitor everything, there is an information imbalance in favor of the agency and at the cost of the citizens (Weimer & Vining, 2017). Therefore, there are likely to be information gaps between the public and how the government is using machine learning. This can be compounded by the nature of government classifications and the sensitivity of law enforcement and national security information. In these instances, what would drive a course correction and norming in the private sector will likely go unnoticed by government agencies outside of the market, further entrenching the bureaucracy. One way that the government attempts to account for this is through forced transparency, such as under the Freedom of Information Act and well as the rules prescribed in the Administrative Procedures Act and the Paperwork Reduction Act. These require public posting of draft rules, allowing public comment and rectification prior to finalizing the rule. Additionally, they prescribe that government agencies must report data being requested and how it will be used (*Administrative Procedure Act (5 U.S.C. Subchapter II)*, 2016; *Paperwork Reduction Act Guide*, 2017).

Other areas for caution are the risks in the machine learning models themselves. This can occur when they are programmed using biased data or when the data does not accurately account for the breadth of inputs, outputs, and outcomes making up the actual situation (Bellamy et al., 2018; O’Neil, 2016). For example, in predictive policing models, a risk would be using only data provided by police resources as this would only include crimes, arrests, or other actions by police officers. This could result in an echo chamber of decisions based on past biased decisions such as higher policing and arrests of certain minority groups (Calders & Verwer, 2010). At the very least, the model would suffer from incomplete pictures of past “crimes” if it is limited to those crimes that were reported to or discovered by the police. As researchers know, not all crimes are reported. Instead, the machine learning model needs to include third-party data sources of crimes (as many are under-reported by as much as 50 percent) and outcomes that go beyond how police organizations count (Robinson, 2017). In the private sector, competition will force these types of changes and best practices as companies compete to win the business of police agencies. Profits will be maximized based on not only more accurate algorithms but also public perception and government purchaser understanding of the underlying data used and the training of the algorithms. However, limited competition can also lead to government failures

since most government agencies have a monopoly on the service or product they provide to the public, or they do not have sufficient knowledge or understanding of vendor products (Weimer & Vining, 2017, pp. 172–175).

Some of these problems can stem from bad data, but they also can be found when the programmers do not understand the data they are working with and how it was gathered. This can be common in the government as few agencies have the internal skillsets for big data, machine learning, and machine learning and typically rely on outside contractors or off-the-shelf product solutions (Berman, 2018; Lane, 2018). The protections of a professional public service (which are required for a functioning complex bureaucracy that can provide uninterrupted services despite political power changes) will also lead to government policy failures. The same protections can lead to the retention of underperforming employees, the difficulty of quickly adapting or changing program staff, and the loss of more skilled and productive employees to the private sector, where they may be better compensated (Weimer & Vining, 2017, p. 172). Researchers note this problem with many government institutions where big data solutions and machine learning are often built up over time, adapting data sets and collection methods that were designed for very different purposes (Lane, 2018). Additionally, the cloistering of specialized data staff (often acquired from outside the operations section of the agency) and the chain of data operations (collections, compiling, cleaning, transforming) that is required for machine learning has the potential to add biases or inaccuracy for a model (Janssen et al., 2017). Because the decisions made by the government have the potential to affect citizens significantly, care and oversight are needed. Additionally, if the public does not have reason to trust government predictive analytic decision-making, then the government does not gain trust and efficiency as possible outcomes of the adoption.

Algorithmic Risks

ML is susceptible to bad data, wrong algorithms applied to the data, and results on ML being applied to the wrong decisions or uninformed about the nature of the ML results in ways which are more harmful than helpful, especially in the context of the public sector. Many of these risks fall into categories of bias or incorrect interpretations and inferences of the predictions. These are two different but important aspects.

Biases in ML are very real and becoming more widely known and understood, but still, there is a significant risk to the unknown in ML biases because of the nature of their training and predictions. In many instances, the ML algorithm acts as a “black box” in that it predicts correlation, but people cannot identify causation to understand exactly what predicts the

outcomes specifically (Bellamy et al., 2018; Kissinger et al., 2021; O’Neil, 2016). Additionally, this is compounded by the fact that ML predictions and associations are typically novel compared to a human understanding of relationships, so they lack a familiarity with associations that people are often used to (O’Neil, 2016). This happens because of the large number (and growing numbers) of hidden layers in ML neural nets, which process associations between data points at a scale that is nearly impossible to do in a manner that is backward understandable about which variables were processed and used specifically to make the predictions (Raschka & Mirjalili, 2019).

Probably more documented are biases that are built into ML programs because of the biases inherent in the training data used to train the models. As discussed previously, ML models require big data to train on. In some cases, the bigger the data, the better. Often researchers are training ML models to make predictions about future cases and using training data from prior human decisions or outcomes. This can be particularly troubling because there very well may be known and unknown biases in these prior human decisions that, if not accounted for, will be training into our ML model, perpetuating these biases under the guises of an accurate ML model (Bellamy et al., 2018; Fejes & Futó, 2021; Katzenbach, 2021; O’Neil, 2016). For example, the practice of “redlining” is well documented private and public process of actively creating a racial disparity in home loans, home sales, and home ownership decisions in the United States. If these decisions are included in a data set used to train an ML model that will be used to help determine who will receive future home loans, then the ML algorithm is being trained on these biased, racist decisions – even potentially labeling these decisions as “correct” for a supervised model. This process, if not accounted for, will perpetuate these biased decisions against non-white applicants going forward. Even more damaging is that these decisions will no longer be viewed by many as being “subjective” because they are being made by ML-based on big data representing the result as an arithmetic prediction which many people have been shown to defer to as though it were an objective fact (Ariely, 2009).

Ethical Risks

Just as there are growing understanding and concerns about potential biases in ML models, there are also growing understanding and concerns about the ethic of ML models making decisions that impact our lives. This is happening more as people understand how pervasive and impactful these ML decisions can become in their lives. There are ethical conversations around the potential to create echo chambers in news, interactions, and social media based on ML models targeting us through our prior interactions online, which may cause

us to miss out on opportunities to understand or learn about opinions, news stories, or individuals which do not ascribe to our prior behaviors. This is ethically concerning because, without an acknowledgment or understanding of this, it has the potential to limit what people understand and view about the world in ways that will potentially make our society more polarized and fractional (Kissinger et al., 2021; O’Neil, 2016). But it also has the potential to allow marketers and companies to control what people believe, set prices based on our likelihood to purchase something, or micro-target us in ways that interact with our impulses beyond how people understand them (Agrawal et al., 2018; Dwivedi et al., 2021).

The AI ethics debates also center around what decisions and what levels of decisions society is going to hand over to ML models versus what needs to be kept by human beings. Notably, this debate is robust in the realm of ML-enabled weapons and warfare with significant, real-world implications as the growing use of ML-enabled technology becomes available to militaries. One of the impactful questions is what levels of the decision will be made by ML versus humans. For example, ML models are becoming more accurate at targeting weapons systems, responding to data about predicting attacks or even the attacks themselves. But at what point when a decision has the impact of potentially taking lives are citizens willing to let the ML model or a human make that decision? It is an interesting understanding because as we know, humans will and have made many mistakes in this area. However, even an ML model which can be proven to be constantly more accurate than a human decider will likely still raise an ethical dilemma about the decision to take lives being outsourced to a machine. In many instances, including the United States Department of Defense, the current compromise is to “keep a human in the loop” (Busuioc, 2020; Kissinger et al., 2021; J. A. Kroll et al., 2017). This looks different in different iterations, but the basic premise is that while ML will be used to monitor data, make predictions, recommend a course of actions, and then carry out those actions, a human will be required to make the ultimate decisions. Many questions remain about the value of these humans in the loop, how much autonomy they will truly have, and how much they will understand about the ML model and processes which led to their decisions which potentially makes a significant determination about the level of trust they should put into those decisions. Ultimately, keeping a human in the loop who merely confirms what the ML model recommends every time without adding a layer of thought, judgment, or human-specific context is more like “this theater” rather than a significant counter-weight to the ML models in the first place. Time will tell as this issue is debated further and more examples, both bad and good, are made available to use to research.

Potentially less life-threatening but more impactful are going to be non-military decisions that impact many more people on a daily basis. These will be decisions about credit worthiness, employment decisions, policing and the prison programs, advertising targeting, decisions about education institutions, and even prices of insurance. More and more, these decisions are being informed by or made directly by ML models. These decisions are being made without the understanding from the impacted individuals about what data is being used by the ML models to make them, nor even how that data is being used by the ML models. This takes away a specific understanding of agency and autonomy from individuals in a number of ways. First, if you don't understand what data of yours is being used to inform the decision, then you don't know how accurate this information is or how to correct inaccurate information. This can and has led to decisions that would have been different if more accurate information had been supplied or inaccuracies corrected (Busuioc, 2020; J. A. Kroll et al., 2017). However, often even the companies employing the ML models do not always understand what all data is used or how that data is used by the model (O'Neil, 2016). Secondly, this opacity in the ML model and data disallows an understanding of what a person can do to change the desired decisions or outcome in the future. By using the ML models, we have often cut off our ability to explain the decisions and, therefore, cannot explain what done differently would lead to different decisions (Busuioc, 2020; Liu et al., 2019). In this scenario, you don't necessarily know what you can do to improve your future chances of getting that job offer, that loan, or the lower insurance rates. And lastly, there are concerns that data and information that was never thought imaginable is being used to make these decisions by ML models and that people should have some agency over the use of our data being provided free to companies and governments. This comes down to an issue of consent, and especially future consent and being able to decide that people do or do not want our data to be used by ML programs. In Europe, citizens have seen this playing out in a number of debates that have crossed into the legal realm, but the illustrative phase comes down to "the right to be forgotten," that is, to have your data taken away from companies and the public to no longer be used in any manner (Chassang, 2017; Keller, 2018). This has an interesting ethical component because it can mean that associations of your choices in movies and music can not be used to inform decisions about your creditworthiness, but it also means that you could request information about your past criminal history to be erased. It opens up interesting conversations about what is correct to be a known public ally and what can and should be used to inform certain decisions about you and the community (Aridor et al., 2020; Liu et al., 2019).

Legal Risks

In the legal debate, the world is experiencing the growth of legal questions related to the above ethical and biased questions. These legal questions extend the issues of accountability, autonomy, accuracy, and oversight of the government on these ML models and the results of their predictions. Additionally, another side of the legal debate, where these ML models are used by the judicial systems itself to inform questions about policing policy and operations, questions about availability or setting of bail for prisoners, sentencing questions for convictions, and even questions about parole of individuals who have met the terms of their sentencing (Busuioc, 2020; Coglianesi, 2019; J. A. Kroll et al., 2017; Liu et al., 2019; O’Neil, 2016). Question of predictive policing, for example, has raised legal questions about the rights of police to target certain neighborhoods or individuals based on ML predictions – many of which lead back to underlying questions about data biases, accuracy, and ethical questions about privacy and oversight of the ML prediction (Citron & Pasquale, 2011; Coglianesi, 2019; Matsumi, 2017). More research and work have been done on the growing number of private company ML models currently used to help set bail amounts or inform parole decisions. Significantly, much of this was launched by a ProPublica investigation and reporting about the lack of accuracy and amount of biases contained in one specific company’s algorithms which unfairly targets individuals who are black for higher bail amounts and predications against parole (Hälterlein, 2021; O’Neil, 2016; Robinson, 2017). Another important legal line of inquiry is to what extent the government can use your own data against you to target, convict, or recommend against parole in line with legal, due process clauses and prohibitions against unlawful search and seizure, which were developed to give individuals rights against the state in terms of what society deemed to be “over-reach” which may have been easier to place legal and societal boundaries when it was limited to physical interactions rather than interactions of data and ML models (Coglianesi & Lehr, 2017, 2019; Liu et al., 2019).

Government Considerations for Machine Learning Use

As discussed previously, the government is investing in and focusing on the implementation of ML models in the government, trying to catch up and capture the gains and efficiency, and effectiveness that the private sector has found in many areas. Additionally, the government is investing in the idea of ML in a number of different implementations to provide the public sector with capabilities and focuses that the private sector doesn’t need to worry about because these are specifically the domain of the government in the United States. These include

decisions that have significant impacts on individuals' lives, from decisions about program and benefit eligibility to questions about justice that impact civil liberties, compliance, and individuals' interactions with the community around them (Engstrom et al., 2020). These challenges are daunting for new technology, which society is still wrestling with from an intellectual level as well as an ethical and liberties perspective. In this section, I will explore some of the considerations of ML design and implementation that are both unique to the public sector as well as more important because of the range of services provided as compared to the private sector, as well as the power and potential impact of public sector decisions being made or informed by ML. These are not trivial, and I believe that the United States and many democracies have only just begun to understand and struggle with the appropriate balances from technology, legal, and ethical perspectives. As ML implementation in the public sector progresses, our society undoubtedly will begin to grapple more and more with these issues in a legal and political context. I believe much will be informed in the political sphere as our understanding and opinions of ML used by the government are pushed into our politics and therefore led by elected decision-makers who are informed by researchers and public opinion. However, it is important to notice and pay attention to the important legislative (or lack of) and legal discussions and interpretations since these too will inform and be informed by the political discussions and decisions.

Several notable government-specific ML designs and implementation considerations include existing administrative law requirements, ML-specific legal and regulatory requirements, issues surrounding the due process and protected rights as applied to ML models and decisions, oversight requirements (e.g., congressional, judicial, public), political and public opinion realities, and the issue of government staff skills and knowledge to design, implement, and maintain ML programs. I believe it is also important to call out that another challenge is to balance all of these considerations with the practical aspect that there are many areas in which ML is going to benefit individuals and our government significantly. I fear that often many people focus on the potential harm and negative outcomes, which are important considerations, without also extolling the potential benefits – especially as compared to the current existing situations.

Administrative Law and Oversight Requirements

Administrative law requirements such as the Administrative Procedures Act (APA) and the Paperwork Reduction Act are going to have an interesting interplay with federal government ML applications. For example, the APA and PRA require information collections to be

explained, justified, and open for public comment in many cases. How will this work when information collections are fed into ML models for predictions, decisions, or other specific applications? Will it be sufficient that the public notices in the federal register note the use of the data in ML models? Or will agencies have to detail which information elements specifically, the design and accuracy calculations of the model, the implementation and monitoring of the model, or other more granular information which may not be possible in all ML applications. For example, current APA and PRA requirements necessitate the publication of implementing regulations that detail how the government agency will collect information and use it to make adjudications of services or benefits, as well as the oversight and administrative review criteria and processes. These levels of details, depending on how they are interpreted, may not be available for all ML implementations, specifically if the government is using an ML algorithm that is not directly explainable or understandable because of the many hidden layers of the neural net.

Some researchers believe these are going to be some of the thorniest issues but have several proposals to help ameliorate them. One idea is to limit the government use of ML to those types which are more explainable and understandable, such as decisions-trees or random forest algorithms, which can be analyzed and produce scores of variables that are weighted in the ML decision-making outputs more so than other types of models (Coglianese & Lehr, 2017; Raschka & Mirjalili, 2019). However, this would limit the government in the types of algorithms that can be used. This could be especially problematic if those less understandable models are also more effective, accurate, and applicable to the specific government use-cases. For these reasons, researchers are working on more explainable ML models which rival the accuracy currently experienced with the often “black box” performance of many deep learning, unsupervised models being employed in the world (Baylor et al., 2017; Gerrits, 2020). There are other options that the government can use to comply with administrative law requirements with existing ML models. One would be to publish in the federal register information about the data used for training, implementation, and details about the ML model itself that opens up the model to the same types of public review and comment as other adjudication and processing regulations. This could cause harm to individuals and government programs however, because specific details about the algorithm itself can open the ML model up to exploitation, gaming, and susceptibility to exploiting underlying data, including individuals' personal identification information depending on how the model is deployed (Coglianese & Lehr, 2019; Engstrom et al., 2020). Another option would be to keep the details of the ML model from the public to guard

against these risks but to leverage the position and authority of the OMB or a similar governing body to have access to and review the design, implementation, and review of agency ML models for the concerns about biases and to verify that the data collected and used in the way that is proportional in the federal register. This would still require the public to trust the oversight and governance body and limit the amount of understanding and interpretation the public can make (Coglianese & Lehr, 2019; Hildebrandt, 2018). However, not all citizens would have the necessary expertise to even understand the intricacies of a ML model and implementation, so a trusted governance body might be beneficial for multiple reasons. Another options is to keep secret the specifics about the ML model for the reasons enumerated, but to create an interactive “what if” process for individuals to interact with the model to see how different inputs of their data might result in different ML outputs and agency decision making. This would be similar to many credit monitoring services which offer “what if” scenarios to see how changes to your personal situation and information would impact your credit score (Wachter et al., 2017). This would not only protect the model design and implementation, but potentially provide individuals with a more helpful way to understand the proposed ML models impact on them and the information that the government agency is proposing to use, allowing them to provide more cogent feedback on the proposal, and allowing for greater public comments and understanding about the government’s processes (Busuioc, 2020; J. A. Kroll et al., 2017).

Legal and Regulatory

Currently, in the United States, there are few laws or regulations which directly impact the design and implementation of ML in the federal government. The United States government, both the legislative and the executive branches, has so far tried to strike a balance of signaling the importance of purposeful, ethical design and use while also trying to bolster the government’s ability to implement and use ML for all of the reasons discussed already. This is strikingly different than the trends being seen in the European Unions and many European countries, which are passing specific laws which are aimed to regulate and specify individuals' rights as they pertain to the public and private collection and use of personal information gathered online and how ML is designed and implemented, specifically by government entities (Aridor et al., 2020; Goodman & Flaxman, 2017; Keller, 2018). Not specifically law, but with the force of law for federal government agencies, the United States has several executive orders that scope out the current requirements and policy. The first is Executive Order 13859 - *Maintaining American Leadership in Artificial Intelligence* which outlined the government’s approach to regulation and oversight of ML, mostly in the public sector. E.O.

13859 mainly emphasized that the government's policy was to help support the growth and expansion of American AI uses and research but tasked agencies with coordinating any oversight or regulation of AI under the leadership and coordination of OMB's OIRA. Additionally, any attempt at oversight and regulation should be guided by ten principles: public trust, public participation, scientific integrity, risk assessment and management, benefits and costs analyses, flexibility, fairness and non-discrimination, disclosure and transparency, safety and security, and interagency coordination (E.O. 13859, 2019). E.O. 13859 also created the General Service Administration's (GSA) AI Center of Excellence and AI Community of Practice for government AI uses, which was later codified in law by the AI in Government Act of 2020 (*40 USC 11301: Responsibility of Director*, n.d.).

A subsequent Executive Order 13960 – Promoting the Use of Trustworthy AI in the Federal Government was signed in December 2020. Unlike the prior E.O., 13960 was specific in laying out requirements for federal agencies to design and use AI in the government, requirements for agencies to catalog and, when practical, publish their AI uses cases, and a common federal government AI policy surrounding the principals, and specific requirements for GSA and OPM to help promote and attract AI skills to the federal workforce and to help agencies acquire AI technology and skilled personnel. E.O. 13960 is significant because of the nine principles laid out that government agencies must comply with when designing, purchasing, and implementing ML solutions. These principles are that the ML models will be: lawful and respectful of our nation's values; purposeful and performance-driven; accurate, reliable, and effective; safe, secure, and resilient; understandable; responsible and traceable; regularly monitored; transparent; and accountable (*Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government – The White House*, 2020). These are fairly far-reaching principles and requirements for federal agencies. However, E.O. 13960 is notably lacking clear definitions of these principles and a framework for how they will be defined, implemented, and held accountable within the federal government. Of note, the E.O. was signed near the end of the Trump administration, and the Biden administration has done little visibly to uphold or further these principles but has also not rescinded the E.O. in any way. Many agencies are developing their own ML frameworks and principles which will give specificity to their own approaches to ML, including adherence to E.O. 13960, but there are no deadlines, oversight, or compliance requirements. Additionally, there is no official coordination between agencies, so it is likely there will be significant incongruence in how these principles will be interpreted and applied across the federal government.

Political and Public Opinion

Another consideration that our federal government is going to need to make is political, and public opinion of ML use in the federal government. Political calculations are an understated reality for government administrators, which is a fundamental design of our representative democracy and the reason our agencies are headed by political appointees and elected officials – to ensure that our government agencies are operating in line with the “will of the people” as defined by their elected representatives. This means that as the political opinions of ML applications ed and flow into people's minds, government agencies will also be worry about implementing ML applications for their programs, absent an understanding of consent from the leadership and popular opinions will be taken into account. However, as already discussed, there is a full-throated embrace of aspiring toward ML applications through development and acquisition. This tension plays out in ways that make the bureau of federal agencies slow to adopt ML until or unless they can be shown to be more useful than potentially harmful, and agencies will likely shy away from any ML applications which are potentially risky or use them in ways which are potentially risky because no agency wants to have the use-case example that leads to judicial or legislative oversight because of negative push-back.

In practice, this likely will mean two approaches. First is one that has borne out in state and local government examples where the public sector acquires ML applications through a vendor. These applications are typically bought “as is” or customized by the company, but importantly the ML applications are considered proprietary to the company, and the details are not disclosed to the public nor even the agency at times. In exchange for this opacity, the company takes over the liability to ensure the ML application complies with best practices, is scientifically sound, and meets any legal or regulatory requirements. In many instances, this has also meant that the proprietary ML models are not subject to judicial or public oversight or scrutiny because they are legally considered the property of the company and not part of the agency which is acquiring it for use in their program (Coglianese & Lehr, 2019; Gerrits, 2020). In the second approach, the government agency designs and develops its own algorithms for an ML model and implements it rather than purchasing it or the services from a vendor. This approach requires much more sophistication in the agency workforce, their infrastructure, and their program design. For these reasons, this is not the most popular approach. The key to this approach as well is that the agency must prove to itself that the mL design and implementation will not put the agency at risk in the court of public and political opinion. This is a difficult requirement to meet because it means proving a negative. Additionally, because of the absence

of coordinated rules and tools within the federal government that have already been discussed, there is no checklist the agency staff can show their leadership that has been complied with, leaving leadership wary to support any ML applications that could be slightly risky.

Government Workforce

In order to really reap the benefits of ML applications at scale and to the extent possible while ensuring the utmost care in design and implementation, the federal agency workforce is going to need to grow and acquire talents, experiences, and skillset necessary for ML applications. This means not just skilled data scientists but also data engineers, data policy experts, machine learning engineers, leadership that understands and can lead these teams through the difficult work and transitions needed, and additional information technology staff to make sure these teams have the necessary technical tools, environments, and maintenance needed for mature ML programs. These things are all critical components as the private sector has discovered and codified, but the public sector, as usual, is still trying to play catch-up. As a significant marker, OPM only just released the first general service job series designation for a data scientist in December 2021, after allowing for a parenthetical working title to be applied to existing job series in late 2019 (e.g., program analyst, information technology specialist, actuarial, etc.) (OPM, 2019). Merely creating a job series is only the first step. The government still needs to grapple with attracting and retaining the necessary talent. Like many government IT jobs over the past decades, it is likely to be a continued challenge for the government when trying to compete with the private sector, which enjoys significant advantages in compensation, hiring agility, and clarity of roles and responsibilities. When looking at government IT roles as a marker, the government is still trying to find the right balance between civil service protections and employment patterns and competitiveness, with congress and the executive branch aggressively making opportunities to try to make a government position easier to apply for, shorten the period between job acceptance and entry on duty dates (especially when security clearance are needed), and retention benefits (which include both remuneration benefits to compete with private sector pay and professional development and work flexibilities such as seen in the private sector) (Gamage, 2016; Gore, 1997).

Government Approaches to Machine Learning Models

As described by Hildebrant (2018), there are two implementation schemes of ML emerging in the government sector: code-driven regulation and data-driven regulation. In code-driven regulation, agencies use ML techniques to automate decisions and adjudications based on

code which is informed through the legal understanding of benefits requirements. For example, an agency that works with legal and policy expect to design a computer program that automates the application, adjudication decision, and benefits granting processes (Hildebrandt, 2018). This program code is informed by the laws and regulations, as well as the administrative burdens of the program. Just as the laws and regulations were publicly debated, these decisions can and should be presented to all stakeholders in a way that makes them transparent and ideally allows for feedback. Data-driven regulations on the code are derived from prior adjudication or interpretations of the law and regulations (think here of unsupervised machine learning programs). This has the potential to mask not only prior biases of decisions but also mask the input and effects of sub-regulatory administrative burdens. It is important to recognize these two distinct uses because they require different considerations and have different audiences (Hildebrandt, 2018).

ML in the Public Sector – Overview and History

As discussed previously, the public sector has identified the development and use of ML as a top priority in recent years, attempting to benefit in similar ways that the private sector has developed in recent decades (Anna-Katharina Dhungel et al., 2021; Kissinger et al., 2021).

ML in Government Literature

Like many areas, the federal government is playing catch-up in AI and ML applications from the growth seen in the private sector in recent decades. This is for a number of reasons, and it is important to note that this is by design as much as it is a failure of the federal government to move quickly and nimbly in data science, just as it does in information technology. It is important to understand that there are clear benefits for the government to be using technology that is considered several years behind “cutting edge” compared to industry and research institutions. This is because the fault tolerance and risk tolerance in government are much lower, and in many instances, the potential outcomes of failure are much higher in the context of risk and results. For example, if your online purchase transactions don’t go through, this would be irritating and would cost the company you are buying from money. But if your tax return was not calculated properly or paid on time, this could cause many additional follow-on outcomes and inconveniences. In the ML aspect, these potential negative outcomes are also multiplied by the magnitude of the outcomes being predicted or processed by the ML algorithms. These factors lead government agencies to feel more comfortable with versions of software several iterations behind the current release because this ensures many, if not all, of the bugs, are already

worked out and patched and ensures that from a cyber security and infrastructure security aspect, there are fewer potential vulnerabilities.

Whiles these benefits are valid and real in federal agencies, this doesn't negate the fact that some of the adaptations are behind because, in many instances, the government is as large, bureaucratic, and slow as many complaints make it out to be. This obviously leads to many instances of IT infrastructure, software, and staff skillsets being well beyond where leadership would like them to be in terms of current. The reasons for these issues and the potential solutions are out of scope for this paper, but the overall point that the government has real considerations and challenges to address, which are just as applicable to ML models as they are IT issues, is important to keep in focus while administrators address the other aspects.

Also important is to realize that even though the government is well behind the public sector in terms of ML use in government programs, the government is also responsible for much of the current state of ML and AI knowledge, research, and usage as it exists today and as it is changing the world and how people interact with the world. This is because much of the early pieces of the ML revolution (just as many things) were funded, supported, and made possible because of the federal government's assistance. From funding early research in the 1950s through the 1970s, as well as creating and making many big data sets available for different iterations, and even now through the permissive and supporting regulatory environment, the AI revolution has happened in the United States and much of the world because of, not in spite of, the federal governments' efforts (Agrawal et al., 2018; Donoho, 2017; Kissinger et al., 2021).

National AI Initiative Act of 2020

One substantial law is the National AI Act of 2020 was enacted on January 1, 2021, and requires the president to create organizations to coordinate and direct the federal government's focus on AI in the public and private sectors. Specifically, the National AI Initiative act required the creation of the National Artificial Intelligence Initiative Office (NAIIO) within the White House Office of Science and Technology Policy (OSTP) (United States Congress, 2021). The Director of OSTP appoints the Director of NAIIO who leads NAIIO's team to oversee interagency coordination of NAI, serve as the central point of coordination and information collection and dissemination for interagency, private sector work, and academic research and coordination on NAI, provide technical and administrative support to the Select Committee on AI and promote access to "technologies, innovations, best practices, and expertise" among federal agencies. NAIIO collates and disseminates much of its work and guidance through its website, www.AI.gov (OMB, 2022).

AI in Government Act of 2020

The AI in Government Act of 2020 was signed by then-President Trump and codified GSA's AI Center of Excellence and AI Community of Practice in the federal government. Both of these initiatives were developed to help accelerate and coordinate AI and ML projects within the federal government. The AI Center of Excellence was one additional Center of Excellence in GSA's programs which are aimed at increasing the ease of adoption, development, and use of government agencies in emerging technology solutions. For example, the GSA Center of Excellence for Cloud Computing is focused on creating government-designed and approved cloud computing platforms that adhere to all of the security, privacy, and regulatory requirements for government agencies, thereby alleviating the need for agencies to seek development and approval individually. Instead, leveraging the GSA's work, they can adopt cloud computing and storage solutions without as much technical work, maintenance, and approval processes. This is another development looking to benefit from economies of scale in GSA Centers for Excellence, alleviating individual agencies in having to go through these same steps each time a new agency or sub-agency unit wishes to make use of new IT resources. Similarly, the AI Center of Excellence is attempting to lighten agencies' loads in needing to develop ML-specific technologies, techniques, and approaches for similar problems (*Artificial Intelligence / GSA - IT Modernization Centers of Excellence*, n.d.). The AI Community of Practice was designed to help bring together the correct entities within agencies to focus their expertise and help each other across the interagency to share best practices, harmonize approaches, and create professional associations across federal agencies.

E.O. 13859

As discussed previously, there is little in the way of formal legislative or regulatory requirements for ML in the federal government. However, there are two important Executive Orders which have the force of law for the executive branch. The first is E.O. 13859 – Maintaining American Leadership in Artificial Intelligence, which did not impact the federal government's use of ML significantly. Its focus was mainly on the federal government's approach to the regulation of AI in the United States. Overall, it directs a liberal environment that supports the private sector's growth of AI applications and its commercial usage of it. It went so far as to direct that any potential regulatory efforts affecting AI had to be coordinated with and reviewed by OMB's OIRA, which would be able to ensure compliance with the E.O. both written requirements and the spirit of the political leadership of the executive branch.

OMB released Memorandum M-21-06 with the subject: “Guidance for the Regulation of Artificial Intelligence Applications” on November 17, 2020, as directed by E.O. 13859. This Memorandum solidified the federal government’s approaches to the regulation of AI and ML in the public sector, reiterating that the overall goal was to nurture the development of AI in the United States, and therefore, the first question each agency needs to consider is whether any regulation is even necessary in the first place. Specifically, out of scope is any ML adoption by federal agencies themselves. However, there is a direction in the memorandum to consider agency actions that can support the growth of AI development and implementation in the private sector. These include providing access to federal government data, metadata, and ML models, which would be useful for private-sector AI use cases. This is in line with the requirements under the Open Data Act but also harkens back to instances of government gifting of global position system data, weather data, or census data which has long benefited the private sector use and development of ML models. Beyond this, the Memorandum also requires agencies to engage with the public and with private organizations to understand their uses of AI, help them develop voluntary AI standards and practices which may be tailored to industry-specific needs, and to coordinate any potential uses of regulation of AI through the interagency processes headed by OMB’s OIRA (OMB, 2021).

E.O. 13960

E.O. 13960 was signed in December of 2020 and was more directed at the federal government’s use of AI. It also changed federal agencies by focusing on the development and implementation of ML for the purposes of improving the efficiency and effectiveness of government programs but also detailed nine principles for ML that agencies must comply with. Notably, these principles were not well defined, nor were the compliance requirements formally instituted, nor were agencies given resources to ensure compliance. However, several agencies have worked through their own efforts to develop agency roadmaps, AI strategies, frameworks, and best practices – many of which comply or aspire to comply with the nine principles enumerated in E.O. 13690. These principles are as follows:

E.O. Principle	E.O. Explanation
Lawful and Respectful of our Nation’s Values	Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation’s values and is consistent with the Constitution and all

	other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.
Purposeful and performance-driven	Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed.
Accurate, reliable, and effective.	Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained and such use is accurate, reliable, and effective.
Safe, secure, and resilient.	Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.
Understandable	Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate.
Responsible and Traceable	Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these Principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable.
Regularly Monitored	Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate

	performance or outcomes that are inconsistent with their intended use or this order.
Transparent	Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including the Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information.
Accountable	Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI and shall monitor, audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI.
	Source: (<i>Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government – The White House, 2020</i>)

Table 2-5: E.O. 13960 Responsible AI Principles

E.O. also requires that the Director of OMB publish a roadmap for guidance that OMB will develop to ensure agencies meet the requirements of these principles within 180 days of the E.O., which would have been by the end of June 2021. This roadmap is supposed to include a plan to push proposed development and information to the public as well as seek public comments, suggestions, and feedback on the roadmap as well as the AI guidance for the federal government (*Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government – The White House, 2020*). However, as happens in many presidential administrations transition, this section has not yet been complied with, nor has the Biden administration rescinded or overruled E.O. 13960, so time will tell at this point whether OMB will take more coordination and direct role in federal agency design and use of ML models for government programs. Section 6 of the E.O. also directed a compilation of voluntary interagency organizing bodies to help agencies implement these principles to be published by the

federal government CIO Council (*Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government – The White House, 2020*).

Section 5 of E.O. 13960 directed agencies to develop an inventory of their AI and ML implementations for their agencies, and to the extent practicable (that it does not introduce national security risks), they should make these inventories available to the public through their websites and through the federal government’s Chief Information Officer (CIO) Council to seek public input on the criteria for the inventories and provide guidance to the agencies’ CIOs on how to complete and publish their inventories (*Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government – The White House, 2020*). These inventories seemingly also have not been completed by agencies, at least not in ways that are available to the public as described by the E.O. – likely also due to the administration changes and lack of follow-up by OMB or the White House thus far.

Sections 7 and 8 of the E.O. focused on increasing the capacity within the federal government to design and implement ML solutions by directing GSA and OPM to create a pathway through the Presidential Initiative Fellowship program to hire and retain AI and ML talent in the federal government workforce. These should be through available fellowships and rotation programs and shall also create guidance on using these programs to increase the number of agency staff with “AI expertise.” Section 8 also directed each agency to designate an official responsible for coordination within the agency of the principles of AI design enumerated in the E.O. working in coordination with the agency’s data governance body (as described and required under the Evidence Act of 2019).

Federal Data Strategy

While not specific to AI or ML in the federal government, it is important to note the Federal Data Strategy and action plan published in 2021. Important because of the relationship between data and ML, as well as the fact that these functions are typically intertwined in the federal government, many agency data strategies either specifically address AI and ML use or serve as a foundational step for ML use in the federal government because the agencies need to have access to the data in formats and of high quality to enable the development of any ML models (Coglianese & Lehr, n.d.; Fejes & Futó, 2021). Action 7 of the Federal Data Strategy specifically enumerates the requirements for federal agencies as outlined in E.O. 13859 and E.O. 13960 as actions that need to be monitored and completed, but also that these actions will help build the foundations for increased use of ML in the government (*Federal Data Strategy 2021 Action Plan, 2021*).

GSA Centers of Excellence for Artificial Intelligence

As discussed previously, the AI Act required GSA to create a Center for Excellence for Artificial Intelligence. This Center has not published much since 2020, but notably, they did create a “CoE Guide to AI Ethics,” which maps federal government considerations for AI ethics to the three stages, Design, Develop, and Deploy of an MLOps framework (General Services Administration (GSA), 2020). This guide is more of a checklist without providing the appropriate tools or answers for agencies. The only other notable document this Center has published is an article for agencies to determine what their readiness level is for AI and ML development and deployment focused on factors surrounding data, workforce, and technical infrastructure, as I have discussed in other sections (*Artificial Intelligence | GSA - IT Modernization Centers of Excellence*, n.d.).

GAO AI Accountability Framework

In June 2021, GAO published an “Accountability Framework” (Figure 2-8 below) for AI in the Federal Government. This framework has four components with two subcomponents, each mapping to different considerations for agency AI design, development, and use. The four components are Governance, Data, Performance, and Monitoring. The GAO Framework defined governance as a process to “promote accountability and responsible use of AI” and distinguishes between governance at the organizational level and at the systems level. At the organizational level, agencies would specify the goals in their ML projects, as well as organizational roles and responsibilities for ML in the agency. At the systems level, this would typically entail the design and development specifications of an ML program.

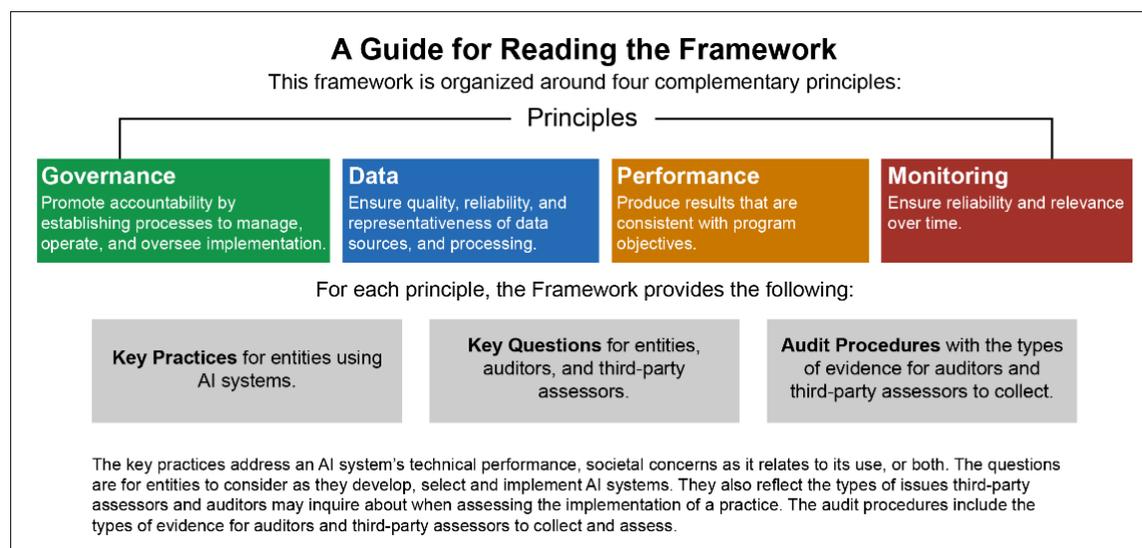


Figure 2-8: GAO Report Summary Table (U.S. Government Accountability Office, 2021b)

The GAO Framework tackles similar data considerations as found in many private sector ML frameworks, focusing on documenting data sources, the reliability of the data, the design of the ML model through data variable selection and transformation, and the review for data biases and reliability as well as security and privacy considerations. In the performance section, GAO focuses on the system-level performance of the model but also the “component level” performance, which is the performance of the ML model in whatever system it is developed for. For example, not just the typical performance metrics and monitoring of the ML model, but how that model’s performance fits into the program, it is designed and developed for. Do the outputs of the ML model, for example, accurately predict or categorize eligible applicants in an adjudication system for public service or benefit? And finally, the GAO framework guides agencies on monitoring the model in production to include monitoring the performance of the ML model to ensure accuracy, identify and account for model drift, and to ensure the outputs of the model are traceable to the use of the outputs. The monitoring for federal agencies should also include an assessment of relevance for the system or process it was created for and identify where the model might be “scaled” to be used for similar programs or use cases (U.S. Government Accountability Office, 2021b). There is not much which is novel for federal agencies in the GAO framework, but it is notable that these are specific recommendations for federal agency use of ML, and it codifies many best practices and ML lifecycle phases seen in MLOps. Of note, GAO publications like this are likely to become the basis for future GAO investigations or reports on specific federal agency programs or processes. Additionally, the GAO teams and authors may also serve in advisory roles to congress when developing legislation or practicing other oversight functions, so it would be prudent for agencies to be aware of the GAO guidance.

Federal Agency AI Frameworks and Strategies

There are several different frameworks and best practices being adopted throughout the federal government. Mostly by agencies since OMB is not pursuing an overarching federal government framework or guidelines currently. Despite that lack of coordination, many of the principles and focuses are similar in nature, but there are some important differences. In this section, I will explore the documents produced and made publicly available by federal agencies and the executive branch. Below is a review of the different frameworks and best practices to see what they have in common and what differences exist. Importantly, this is a developing area. Just as is required in E.O. 13960, agencies, like corporations, may feel compelled to spell out in a public manner for themselves and the public what their approach to ML in the public sector will

development, and monitoring as espoused by MLOps (U.S. Department of Health and Human Services, 2021).

ML Frameworks and MLOps

An emerging field has been developing along with the increased use of ML for all types of use cases and operations to help guide practitioners and teams when designing, developing, and implementing ML models called “MLOps,” which is shorthand for Machine Learning Operations and an extension of a software development operations approach designed as an agile toolkit for teams to follow a structured development path. There are many similarities between software development and ML development, but also notable differences that justify the related but separate approach of MLOps (Greco, 2021). MLOps focuses a team on the different steps that need to be followed and considered at the different parts of the ML model development, deployment, and monitoring stages. It is also important to understand that this is known as an iterative process because most ML models are not a one-time run model but instead implemented in production and need updating to keep them valid and accurate and to improve them based on feedback, changes to data or the target of the model, and to ensure that they are still producing the desired outputs. We’re going to cover MLOps at a high level, as well as the specific consideration for an ML model because this is important to understand as a baseline for ML in public sector use for reducing administrative burdens, but also because it is important to show how these academic and industry standards fit into the existing requirements and considerations for government use, and where they need to be improved upon for government use too.

As can be seen in Figure 2-9: MLOps Cycle (Burkov, 2019)) below, MLOps represents a continuous practice of design, development, monitoring, feedback, iterative changes, and so forth. These steps for an ML project are important from a technical perspective, but we’re also going to use them to overlay the requirements for the public sector design and use of ML to see how it needs to be extended to address the requirements and concerns of the public sector. First, though, I will explore the MLOps process and step in a bit more detail to understand how they fit together

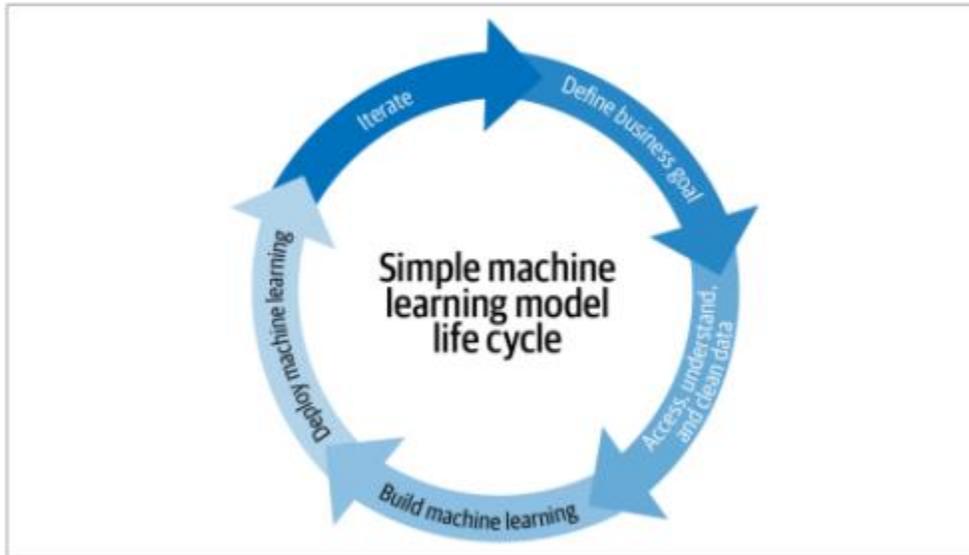


Figure 2-9: MLOps Cycle (Burkov, 2019))

Model Development

In the ML model development stage, the organization determines the problem and whether ML can help solve that business problem. If there appears to be an ML solution, then the underlying necessary condition for a successful ML model is investigated, especially the required data at a high enough quality for the ML program. In addition to the quality and availability of the correct data, privacy rules, regulatory requirements, and use permissions for the data will be investigated to ensure that the data can administratively and ethically be used for ML.

Once the above questions are resolved, the development of the ML model can begin. This typically involves technical steps, including feature engineering and selection to determine which variables with the data set to use given the target variable (those that we wish to predict or the outcome variables). Then the team will select an appropriate ML model type for the data and the problem, as well as tune the hyperparameters for the model based on the data and target outcomes. A growing popular use case at this stage is the deployment of “auto ML” tools which run many iterations of different ML models and hyperparameter tunings to recommend or select the appropriate ML approach based on desired outcomes (Breuel, 2020; Greco, 2021). These still need to be reviewed by appropriate parties, but the advent of auto-ML has drastically increased the design phase of ML solutions. Once the appropriate model is chosen, training and evaluation of the model can proceed in-depth. This is the process of actually training the ML model and reviewing the different iterations to optimize the models and outputs. This stage is

incredibly important because this is where the accuracy, validity, and bias of the models are reviewed and corrected if possible. Two additional steps in MLOps are reproducibility and a responsible AI review. Reproducibility in MLOps is similar to reproducibility in academic research, whereby the ML team must ensure thorough documentation of all of the key factors of the ML model, data, and environment so that each experiment and model can be reproduced at a later time. This has not always been the case, and for the same reasons as in research, but also because of the growing importance of oversight and regulation, this has become a critical component. A responsible AI review is what it sounds like. In addition to the ML engineers' review of the accuracy and validity of the ML model outputs, there needs to be a developed methodology to review each model for accuracy (different algorithms' accuracy can be measured in different ways), additional tests and reviews for biases in the data and the models are checked, and the model inputs and outputs are evaluated against the business use case of the model to determine that it seems to fit against the intended solutions (Alla & Adari, 2021; Raschka & Mirjalili, 2019; Treveil et al., 2020). Additionally, this is where tests regarding the model outputs as compared to non-ML solutions might be tested and reviewed. For example, if an ML model is being developed to make predictions or adjudication decisions for a program, the developer needs to compare the outcome of the model against the decisions that were made previously by humans to determine if the model is more or less accurate. There may also be consideration of a threshold of increased accuracy, which determines if the ML model is accurate enough since a negligible improvement in accuracy may not be enough incentive to take on the difficulty of deploying the ML model into production.

Model Deployment

Model deployment is a distinct and important stage that many don't consider until they must address it. Since model development is the realm of research and experiment, it tends to be the focus of many concerns and considerations of MLOps. However, once a model is produced, the organization has to deploy the model in a production environment. This is similar to the deployment stage of software development but has some unique characteristics. In the public sector, these challenges are often visible to both the public and the organization as it often requires availability downtime of programs, forms, and applications, as well as changed interaction types with programs and applications. On the organization side, not only does the ML model code need to be rewritten or produced into a production environment and system, but this stage needs to be monitored and reviewed to ensure no unintended changes to the model or how it performs happen. Additionally, depending on the purpose of the ML model or how it will

work within the intended program, there often must be other software and process changes for the organization. There are two types of ML model deployment, dynamic deployment and static deployment (Burkov, 2019). For example, will the ML model run as a service or provide live scoring? This would be an example of the ML model running based on real-time input or updates from interactions with webpages or the completion of information collection forms that then automatically use the ML outputs to drive action. An example of this would be an ML model chatbot which is an interaction with an individual who is exploring information about eligibility requirements, benefits information, or application requirements in real-time. In other instances, ML models would be run as a batch process as part of a new or changed business process (Burkov, 2019; Treveil et al., 2020). For example, using ML models to help adjudicate program eligibility would be run after certain application information is submitted but would still require changes to the current adjudication software and business processes for the organization.

Model Monitoring (Operations)

Once the model is deployed, the organization needs to have a plan and a process for monitoring the model. This is especially true for a model which runs as a service but is still necessary for all ML models. The monitoring of the model is an important process because the model can experience “drift,” which is changes in accuracy or validity of the model, typically through changing input data overtime (even potentially changing because of the implementation of the model) or changes based on other factors such as eligibility requirements or the population of participants. Additionally, for accuracy, changes to model biases also need to be monitored and corrected. Without a plan developed and adhered to, the ML model can become worthless or particularly detrimental to a program or its intended outputs (Alla & Adari, 2021; Mäkinen et al., 2021).

The model monitoring stage is also where evaluations should be designed and performed to evaluate the impact of the ML models on the intended outcomes of a program. For example, in adopting ML programs to reduce administrative burdens, the government needs to monitor through our performance measurement systems the impacts of measured administrative burdens, as I determined in Framework 1. The government would also want to perform formative and summative evaluations to more systematically review whether the ML program is developed and deployed properly within the public program, as well as an in-depth evaluation of the changes in outcomes based on the ML program aimed at reducing administrative burdens (A. Kroll & Moynihan, 2017; Orr et al., 2019; Patton, 2012). Based on evaluations, as well as monitoring performance measurement criteria, the monitoring stage is likely to result in iterative changes to

the ML model and program, which would lead to new requirements that will move into the design stage again. This is why MLOps is most often depicted as a cycle since what is learned should be used to make corrections and changes to the existing program (Treveil et al., 2020).

Specific Considerations for Public Sector

As discussed above, MLOps has become the industry standard in moving through the lifecycle of ML model creation and deployment. Specific considerations for the public sector come in within each stage and should be part of the MLOps model for public sector organizations. In the development stage, specific considerations to the data availability need to have a specific public sector lens. This is because often, the government must contend with data collection and use standards the private sector doesn't need to consider. This is where the administrative law requirements of the APA and PRA would be reviewed to ensure that there is specific permission or no prohibition from using collected data in this manner (*Administrative Procedure Act (5 U.S.C. Subchapter II)*, 2016; *Paperwork Reduction Act Guide*, 2017). Since the APA and PRA are not specific to the uses of ML at this time, the organization would need to ensure that the intended outputs and outcomes of the ML model are in line with the approved uses of the data. For example, application data that is used to determine eligibility for a program is approved whether that adjudication is being done by ML or a manual process. The replicability of the ML model must also be robust enough to comply with any federal government oversight requirements, especially those from the judicial branch or congressional inquiry. While there have not been many examples of judicial oversight of AL or ML in the public sector yet, there is an acknowledgment that this is an area of growing interest and concern (Busuioc, 2020; J. A. Kroll et al., 2017). And finally, there will have to be the inclusion of reviews in each stage to ensure the ML model complies with legal and regulatory requirements. Although there are not many yet outside of E.O. 13960, it is likely that administrators will see increased requirements for the public sector use of ML that will need to be accounted for in the MLOps process.

E.O. 13960 Framework and MLOps

Now that I have covered the fundamentals of MLOps, I am going to revisit the current required principles for ML development and implementation in the federal government as outlined in the nine principles in E.O. 13690, and I am going to map them to the different stages and steps within the MLOps model to show how agencies can meet these and where MLOps needs to be extended to incorporate the federal government considerations. This will not be a

detailed and exhaustive explanation of how to apply the required steps but a robust outline and mapping that will help researchers and administrators understand the basics of ML implementation in the federal government, to then help show how to apply ML for reduction of administrative burdens.

E.O. PRINCIPLE	E.O. EXPLANATION	MLOPS MAPPING
LAWFUL AND RESPECTFUL OF OUR NATION’S VALUES	Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation’s values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.	Model Development: Initial concept and model research and design
PURPOSEFUL AND PERFORMANCE-DRIVEN	Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed.	Model Development and Model Monitoring:
ACCURATE, RELIABLE, AND EFFECTIVE.	Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained and such use is accurate, reliable, and effective.	Model Development: Model training and assessment; Model Monitoring: model accuracy and drift assessments, as well as model evaluations.
SAFE, SECURE, AND RESILIENT.	Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.	Model Monitoring: Assessment of deployment model, including model logging and cyber security monitoring.

<p>UNDERSTANDABLE</p>	<p>Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate.</p>	<p>Model Development: When developing and testing the model, can the results of the model be explained and transparent.</p>
<p>RESPONSIBLE AND TRACEABLE</p>	<p>Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these Principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable.</p>	<p>Model Development and Model Deployment: Documentation for oversight and reproducibility. Also, the deployment processes and standard operating procedures within the programs they are being deployed.</p>
<p>REGULARLY MONITORED</p>	<p>Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order.</p>	<p>Model Monitoring: As part of the ML monitoring plan, assessing the model against E.O. 13960 requirements (along with any additional legal or regulatory requirements).</p>

TRANSPARENT	<p>Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including the Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information.</p>	<p>Not specifically addressed in MLOps, but this would be the process of including documentation, compliance checks, and other pertinent information in public and oversight workstreams to ensure there are no “hidden ML projects” within the federal government.</p>
ACCOUNTABLE	<p>Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI and shall monitor, audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI.</p>	<p>MLOps Governance: This could be the agencies detailing their MLOps processes and having an outside entity monitor or audit them, or this could be judicial or congressional oversight of agency MLOps processes.</p>

Table 2-9: Mapping E.O. 13960 Principles to MLOps

As seen in Table 2-9, the requirements of E.O. 13960 can be accounted for in federal agencies’ MLOps processes and plans since many of the compliance exercises are either already part of the MLOps framework or they have natural timing in the existing MLOps process. For example, the E.O. principle of accuracy, reliability, and efficiency overlaps well with the testing and monitoring of the accuracy and validity of an ML model once it is designed and trained. Additionally, there is a requirement for continued monitoring of the ML model’s accuracy and validity in the Monitoring stage which would also comply with the E.O. It is clear that the Trump administration didn’t develop E.O. 13960 out of thin air. It was informed by researchers

and practitioners of ML solutions and adapted to the goals and policy of the federal government, so it should not come as a surprise that there is a great deal of overlap between E.O. 13960 requirements and the best practices from the academic and private sectors. However, the federal government has yet to link the E.O. 13960 principles to more holistic guidance about how they should be implemented in each agency, as was required by section 4.

If the current administration intends to make good on this requirement, OMB can benefit by tying the implementation of the principles into existing ML operational guidance and best practices such as MLOps. This would benefit federal agencies because many are likely following these types of practices since they are industry standards, but it would also benefit them because it would provide a clear mechanism to implement and check compliance with the implementation of the principles. Detractors may be concerned about integrated forms of compliance with the principles in the same design, development, and deployment processes and teams because it may not provide sufficient “independent” oversight and auditing that some have called for government AI (Coglianese, 2019; Mergel et al., 2016; U.S. Government Accountability Office, 2021b). I believe it is imperative to implement these principles as close to the teams doing the ML development as possible because these principles should be interwoven into all parts of the process, including problem definition and solution ideation, in order to have the greatest impact. However, I do not believe this prohibits an additional independent entity from auditing the principles as well if that is desired or required.

Summary

As discussed in this section, the government is further implementing ML solutions for federal government use cases and problems, just as the private sector has done. Even the concerns from detractors are getting heard, and the federal agencies are drawing from best practices, lessons learned in the private sector, and emerging legal and policy guidance to ensure that ML solutions designed and developed for the federal government will be done so in compliance with these considerations and aspects. There is no reason to think that these efforts will not continue to grow, and the federal government will continue to experience improvements to policies and processes just as the private sector has enjoyed in the past decade. More and more, the timing is ripe for the federal government to also look at using these ML solutions to solve the challenges posed by administrative burdens, thereby improving the effectiveness and efficiency of government programs and services. In the next section, I will examine the next step, ensuring that administrators and researchers can evaluate how administrative burdens (and especially the

reduction of them) can impact government services and programs, as well as measure and evaluate the impact of ML solutions in this area.

Evidence-based Policymaking

Overview

The evidence-based policy movement is focused on adopting and implementing policies and programs in the government which effectively produce desired outcomes as proven by evaluation and experiments (Cartwright & Hardie, 2012; Haskins, 2018). This focus is similar to the performance management movement because it looks to increase the use of performance information and measurement, empirical evidence, and data. However, it is less focused on the process of how things get done and more focused on whether the program, successfully implemented, actually achieves the intended policy outcomes or goals that it was designed to do (Heinrich, 2007). Evidence has recently been defined by Congress to include “foundational fact-finding, policy analysis, program evaluations, and performance measurement” (Vought, 2019). However, this is broader than many other definitions of the movement, which tend to focus on scientific methods applied to gaining causal inference about program outcomes based on their inputs and activities (Cartwright & Stegenga, 2011). Many point to the growth of the evidence-based policy agenda from the application of evaluation methods in the public policy cycle. However, there have also been growing calls to more fully integrate performance management into evidence-based policymaking (A. Kroll & Moynihan, 2017).

The evidence-based policy movement seeks to empower or require policymakers to adopt programs that have been proven effective in other areas and therefore may have a higher likelihood of being effective elsewhere. Another approach is to adopt programs that can accurately be evaluated for their effectiveness once implemented. This allows the programs to be monitored and changed or defunded if determined not effective. These processes attempt to get away from the adoption of policies or programs and leave it up to the executive branch to administer those programs without a built-in feedback process to be able to determine if and how effective these programs are in achieving their intended goals (Orr, 2018; Sanderson, 2002).

Like performance management, the political support of evidence-based policy derives from the desire to show effective and efficient use of tax dollars in the public sector. There are increasing calls from legislators, experts, and the public for greater efficiency and effectiveness in government. For many, this means doing more (or the same) with fewer resources. This seems logical since much of our political energy surrounding government oversight is focused on eliminating waste and achieving goals as effectively and efficiently as possible (Cartwright and Hardie, 2012). The evidence-based policy movement is heavy on quantitative evaluations,

especially prizing Randomized Control Trials (RCTs) as a means to be able to evaluate causal outcomes, but there is also a growing research literature on the use of qualitative methods and mixed methods approach (Khagram & Thomas, 2010). RCTs are scientific approaches that use random selection and assignment of participants for experimental and control groups to eliminate several types of potential biases which undercut the ability to make causal inferences from the experiments (Cartwright & Hardie, 2012; Khagram & Thomas, 2010).

The United States government has pushed agencies closer to this standard. The Office of Management and Budget (OMB) has released several directives and best practice standards citing the need to include RCT experiments, statistical data evaluation, and data creation and analysis (Pawson, 2002; Sanderson, 2002; Stack, 2018). The 2010 Coalition for Evidence-Based Policy also concluded that while scientific methods were a necessity for program evaluation, nothing could be as good as randomized controlled trials (Commission on Evidence-Based Policymaking, 2017). Beyond the federal government in the United States, the British government, as well as a multitude of U.S. state and local governments, have also required exact percentages of funding to be linked to policies and programs that are backed up by statistically based evidence (Jennings and Hall, 2011). The United States federal government took this a step further by passing the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act) in 2019. The Evidence Act requires agencies to focus on creating data and evaluating programs and policies to justify which implementations should be funded. It also requires the creation and open sharing of government administrative data, the creation and use of statistical evidence, and cross-agency and private-public data sharing to further these goals. Importantly, the Evidence Act also begins to tie together the use of performance management and evidence-based policymaking and evaluation, whereas these practices were often separate and stove-piped if they were pursued at all in agencies (Commission on Evidence-Based Policymaking, 2017; Vought, 2019).

The Evidence Act contains several thrusts, all aimed at creating the foundation for and the motivation to create and use evidence in programs (Keiser & Miller, 2020). The Evidence Act attempts to pull together the disparate requirements of performance management, evaluation, and administrative and procedural requirements into a process that can more easily be used by leadership to make better-informed policy and program decisions as well as to more effectively and efficiently manage programs. While this seems monolithic, it may actually decrease resources needed by agencies that are already required to comply with these areas and may focus

attention on these tools for their intended purposes rather than treating them as compliance exercises (Androutsopoulou & Charalabidis, 2018; J. A. Kroll et al., 2017; Stack, 2018).

Definitions

Like performance management, evidence-based policymaking strives to improve the effectiveness of government through the use of information and scientific evaluation techniques to help make decisions on program design, funding, performance measures, implementation, and management (Heinrich, 2007). The term “evidence” in evidence-based policymaking is often defined based on the needs of the program evaluation but can be generally defined as statistical information necessary to reach a conclusion regarding some aspect of the policy. Statistical data is seen by many as necessary but not sufficient for evidence-based policymaking. The 2017 Commission on Evidence-Based Policymaking’s final report listed several necessary factors for evidence-based policymaking. These include robust government data, organizational expertise in understanding and analyzing data, inter-organizational research and data sharing, and long-term design and planning for evidence-based approaches to programs (“The Promise of Evidence-Based Policymaking,” chapter 5).

Others have defined evidence-based policymaking in more simple terms of using scientific methods to answer specific questions for policymakers. They call for implementing methods that adequately deal with reliability issues as well as policy validity in evaluations of programs and policies. Many proponents believe that evidence-based policymaking techniques can help policymakers determine whether and how “established results [of existing programs] bear on a policy prediction” and how to evaluate those policy predictions taking into account all the evidence (Cartwright & Hardie, 2012; Cartwright & Stegenga, 2011). There are some proponents who argue that only policies or programs which have been proven effective through rigorous methods should be adopted or implemented elsewhere. There is understandable logic to this position, as new programs are often complex, expensive, and require significant time and resource investment to implement before they make measurable impacts. Therefore, there is much less risk to only adopting on a full scale those programs which have been proven effective elsewhere in order to increase the chance of success. Many critics of this approach, however, point out how this ideal is not always achievable by legislators or politicians in the political realm. The government itself has pushed agencies closer and closer to this standard. The Office of Management and Budget (OMB) has released several directives and best practice standards citing the need for the inclusion of randomized controlled trial experiments, statistical data evaluation, and data creation and analysis. The 2010 Coalition for Evidence-Based Policy also

concluded that while scientific methods were a necessity for program evaluation, nothing could be as good as RCTs (Jennings & Hall, 2012).

History of Evidence-based Policy in the U.S.

While some trace the roots of the evidence-based movement to the early 17th century enlightenment period (Heinrich, 2007). Others agree that the evidence-based policy movement developed initial followings among professionals in the 1970s, but it took on the form and favor it currently enjoys at the end of the 1990s. This is likely because the United States and the United Kingdom experienced rapid growth in government budgets in the 1960s, mostly due to social welfare programs, coupled with an increased desire to apply social science techniques to solve public problems (Heinrich, 2007). There was increased optimism that science would provide immediate impacts on government effectiveness and efficiency, perhaps building up after the world experienced vast scientific expansion during and since the end of World War II.

Several researchers point to the post-WWII era as the advent of the evidence-based policy movement, which focuses on rigorous research methods to build evidence about “what works” that is used to focus public resources on those interventions in policy design and implementation. These methods were born in the medical field, which, prior to adopting the methods of randomized controlled trials (RCTs), were mostly based on anecdote and unscientific evidence (Baron, 2018). While the 1940s and 50s experienced a dramatic increase and acceptance of RCTs in medical trials, these were codified in the United States in 1962 when the FDA changed regulations to require them for any new pharmaceutical approval (Baron, 2018).

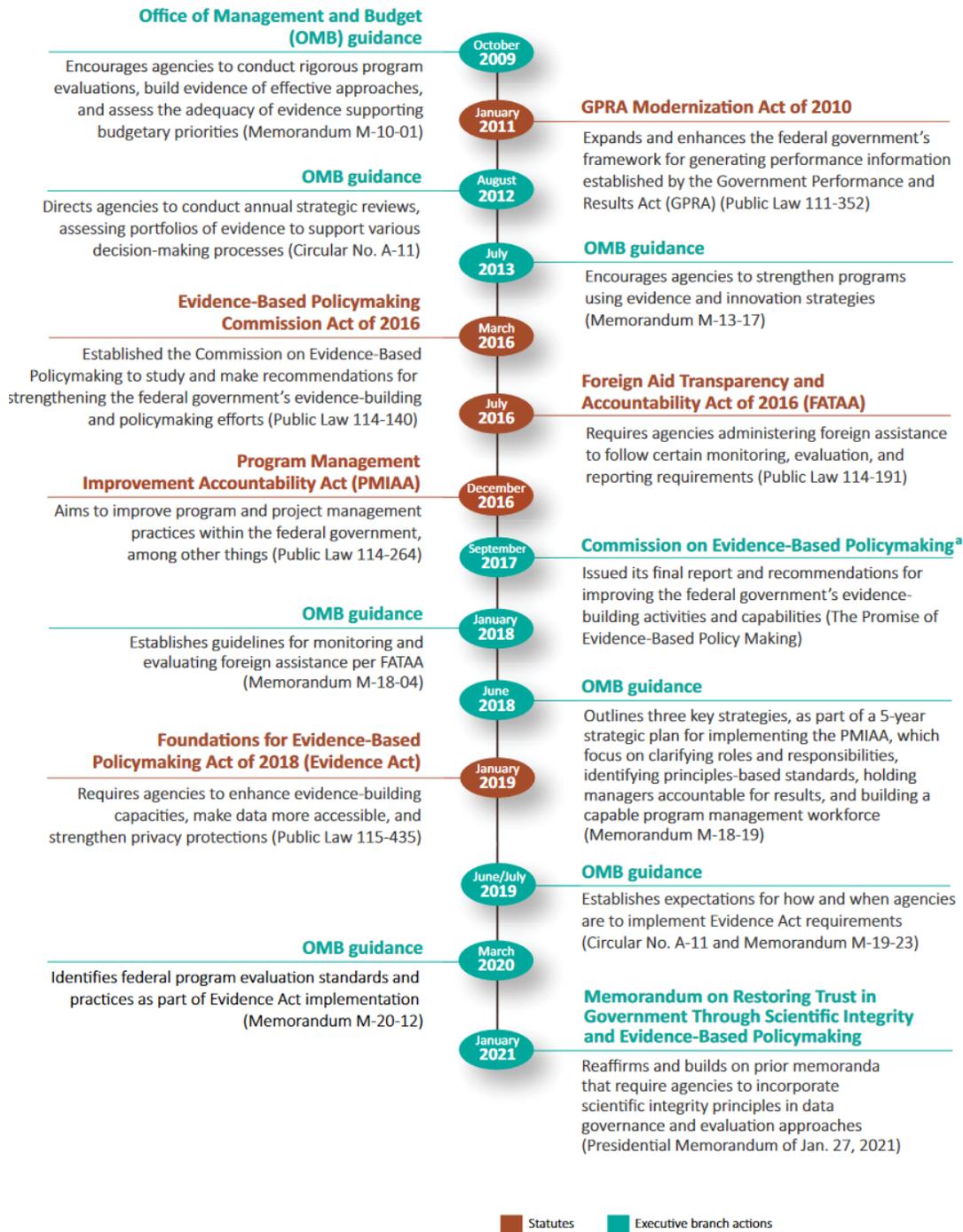
These methods quickly poured into other scientific and policy areas, including social policies, which had been dabbling with RCTs since the 1930s but began to proliferate in the 1960s and 70s in the United States (Baron, 2018). Just as in medicine, the goals were to seek rigorous analysis of policies and processes using scientific methods in order to inform policymakers of the ideal policies and programs to enact in order to be effective (Head, 2015). Some examples include the 1930s Cambridge-Somerville Youth Study that evaluated counseling and group recreation for low-income adolescent boys; the 1961 Manhattan Bail Bond project to test the effect of releasing certain defendants without bond prior to the trial; the 1962 Perry Preschool Study providing preschool to children under four from low-socioeconomic backgrounds; the 1974 National Supported work Demonstration to evaluate offering subsidized jobs to “hard to employ” people and then unsubsidized job placement; and the 1971 RAND Health Insurance Experiment to test different levels of support with health-care costs and the resulting outcomes of individuals health (Baron, 2018; Cartwright & Stegenga, 2011; Pawson,

2002). The policy impacts of these early RTCs in social policy were varied, but they set in motion the ability to combine rigorous experimentation and social policy design and implementation conversations, especially in ways that could contribute to or at least navigate sensitive political discussions and realities.

These early experiments set up additional use of evidence-based policy in the 1980s and 90s, especially surrounding welfare and employment policy and programs. For example, the Manpower Demonstration Research Corporation's (MDRC) "Welfare/Work Demonstration" in the 1980s led to HHS funding large-scale RTCs of different programs aimed at supporting the poor with the goals of employment and income supplementation (Baron, 2018). In 1988 Congress further supported these efforts by passing the Family Support Act of 1988, which required HHS to use RTCs to evaluate welfare programs (Evelyn Z. Brodtkin, 1997). These efforts and the interest in evidence-based policy transformed into a landscape where presidential administrations, Congress, and state government began allowing and, in some instances, requiring certain carve-outs in policy design and implementation rules to allow for and promote RCTs and other experimental methods to evaluate the impact of policy choices. In 2002 congress passed the Education Sciences Reform Act that established the Institute of Education Sciences (IES), which provides independent research on education policies for the Department of Education (Commission on Evidence-Based Policymaking, 2017; Khagram & Thomas, 2010). Several non-profit and private organizations also formed that supported evidence-based policy movements by entering into public-private partnerships to design and test policy and/or promoting the use of evidence-based techniques in federal, state, and local governments, such as the Jameel Poverty Action Lab (J-PAL), the Coalition for Evidence-Based Policy, and many others (Baron, 2018; Cartwright & Stegenga, 2011; Haskins, 2018).

In recent years the United States federal government further enshrined the theory and requirements of evidence-based policymaking in requirements, resources, and organization structure for executive branch agencies. This takes to form of budget requests and justifications, policy design and adoption requirements, and programmatic evaluation, auditing, and feedback mechanisms (Orr, 2018; Stack, 2018). Evidence-based policymaking theory and ideals are interwoven through performance management and evaluation criteria and have begun to bring those areas together and increase the insistence on and opportunities for more rigorous scientific methods to support not only the implementation but also the design phases of policy. Ideally, in many eyes, RTCs or similar methods will be required before new policies are adopted on a grand scale, but at the very least, they should be used to determine if a policy implemented is effective

and how that program should be changed over time or if it should be rescinded (Commission on Evidence-Based Policymaking, 2017; Khagram & Thomas, 2010). This has culminated currently in the adoption and implementation of the “Evidence Act” for the federal government, which many states follow suit as they tend (Jennings & Hall, 2012; Ryan, 2019). In Figure 2-10 below, GAO outlines their view of these requirements and landmark legislation and executive guidance as they instruct the executive branch on the best practices of the evidence-based policy movement since 2009. Some of these are specialized to a particular government sector or task and so are not dealt with in this work, but many of these are covered in detail and show how they build on each other and tie together.



Source: GAO analysis of select laws and executive branch materials.

Figure 2-10: History of Evidence-Based Policy and Evaluation in Federal Government (U.S. Government Accountability Office, 2021a))

There have been increasing calls from legislators, experts, and the public for increased efficiency and effectiveness in government. For many, this means “do more (or the same) with fewer resources.” This seems logical since much of our society is headed in the direction of eliminating waste and securing goals as effectively and efficiently as possible. It should come as no real surprise that this has included a desire that the public sector should aspire towards the

apparent efficiency experienced in the private sector. This has manifested itself in increased program evaluations, performance management, and now the rising star of evidence-based policy making. These efforts and initiatives hope to achieve the often-promised goal of effective and impactful government strategies, as well as cost savings when waste is eliminated from government agencies and programs. It's important for researchers and government administrators to understand not only these motivations but the history of these movements and the lessons learned from similar movements (Cartwright and Hardie, 2012).

Some areas of government policy are more steeped in evidence-based policymaking than others and have a long history of implementing and utilizing it, such as the medical field or education (Heinrich, 2007). However, other areas are still either in a nascent stage or have not yet evaluated their potential to implement evidence-based policymaking. Regardless of where they are now, there is more internal and external pressure for government administrators to implement evidence-based policymaking in their organizations. Government-wide, irrespective of political ideology, the promise of adding scientific rigor to government programs administration and oversight appeals to the idea of a more efficient and more effective government where programs exist, as well as rigor in determining which programs should continue to receive taxpayer money (Commission on Evidence-Based Policymaking, 2017).

The Coalition for Evidence-based Policymaking

In the United States, much work on the theoretical application of scientific evidence for policymaking as well as practical applications was conducted by the Coalition for Evidence-based Policy. Billed as a nonpartisan and nonprofit organization, the coalition worked on several important program initiatives which helped shape the discourse and understanding of the use of scientific evidence and evaluation techniques of programs to understand their effectiveness and impact, which they argued should directly be linked to their future support, funding, and adoption of similar programs in other areas (Haskins, 2018).

The Evidence-based Policymaking Commission

The work of the coalition and other government and academic proponents resulted in a bipartisan bill in March 2016 that set up the Commission on Evidence-based Policy. This Commission was charged with studying and making recommendations in several areas to inform how the federal government can move towards and further implement principles of evidence-based policymaking. The commission put out several interim reports during its tenure and published its final report in 2017 with twenty recommendations for congress on how to require

and implement evidence-based policymaking criteria and standards in the federal government. Of note, beyond the creation of evidence and data and application of these to scientific methods for evaluation, there was a substantial amount of effort focused on how to increase the transparency and sharing of government data to allow for further public-private partnerships and academic use of government data. While not all recommendations were taken up, the commission's work and the members of the commission made a lasting impression on many lawmakers, administrators, and academics who are optimistic about many of the potential benefits identified by the evidence-based policymaking movement (Cartwright & Hardie, 2012; Commission on Evidence-Based Policymaking, 2017).

Current Evidence-based policy in Federal Government

Evidence Act

The final report of the commission resulted in the drafting, debate, and eventual passing of the Foundations for Evidence-based Policymaking Act of 2018 (which was signed into law in 2019) and colloquially known as the Evidence Act (Ryan, 2019; *Three-in-One: The New Evidence Act*, 2020). The Evidence Act had three main thrusts.

First, to create data and information to be used for evidence gathering and evidence confirming activities such as program evaluation, program learning agendas, and infrastructure to nurture and guide evidence-building activities. Of note, this section required designation of evaluation officers within agencies and linkages between strategic plans and performance management activities to data creation, evidence creation, and evaluations of agency programs. It also set up OMB bodies to provide guidance, support, and oversight to these activities within agencies (Ryan, 2019; Vought, 2019). It does this by linking new requirements under existing agency strategic plans (as required by GPRAMA) to new requirements for agencies to identify policy questions related to their strategic plans and to create evaluation plans to design and conduct evaluations of their strategic priorities and policy areas. These have become known as “learning agendas.” Additionally, to help achieve these requirements, agencies are required to appoint evaluation officers and statistical officials, and the OMB is required to create an Advisory Committee on Data for Evidence Building to provide guidance for agencies and to provide a community of practice and assistance to agencies on these new activities. This section and its requirements build off of the performance management and performance measurement requirements of GPRAMA to additionally link evidence-building and evaluations to the Agency Priority Goals and Cross Agency Priorities goals

The second section was the “Open, Public, Electronic and Necessary Government Data Act, or the OPEN Government Data Act. This act is aimed at making government data available to the public as a default rather than the default, being that federal agencies do not share data unless required through legal or policy means (especially through the Freedom of Information Act process). The law requires agencies to create and publish data inventories and data catalogs of available data sets, many of which are now available on www.data.gov to the public. This section was a congressional recognition of the value of federal government data to the public, either academic institutions or businesses. Additionally, there was a recognition that agencies themselves were failing to leverage their own internal data or interagency data for the most effective policies or processes. The OPEN Government Data Act also established the roles of Chief Data Officers (CDOs) and CDO councils within agencies and within the interagency. The CDO is responsible for implementing these requirements but is also charged with several areas of internal improvement in agency creation, categorization, and use of data for policymaking and program evaluations. It required the designation of Chief Data Officers for each agency who are empowered to take ownership of agency data, requiring them to create data strategies, data inventories, and data sharing efforts. This section is aligned to the first so that these data can be leveraged for evidence-based policymaking but is also representative of the shifting understanding that data has become as important and valuable as the systems and programs aligned to these data. This section also calls for increased sharing of data within agencies, between agencies, and more towards identifying what government data can safely be shared with the public, including academic and nonprofit partnerships (Ryan, 2019).

The third section of the Evidence Act reauthorized the Confidential Information Protection and Statistical Efficiency Act of 2002, which required the protection of information collected by the government for statistical purposes as well as interagency sharing of that data for statistical purposes. The section of the Evidence Act expanded these protections but also empowered OMB to promulgate additional sharing permissions, even those with entities outside of the government, as long as the confidentiality of the information collected could be maintained. This is a furtherance of the ideals within the evidence-based policymaking movement that government data can be used by researchers to expand on the evidence of the efficacy of policies and programs, which should then be used by the government when deciding whether, how, and where to implement policies.

M-19-23

The first OMB guidance for federal agencies on the Evidence Act Requirements was M-19-23, “Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance”. It was Phase 1 of a four-phase approach to OMB implementation guidance. It focused on the establishment of the required officials (Chief Data Officer, Evaluations Official, and Statistical official), the creation of agency data governance bodies, and the creation of agency learning agendas and evaluation plans. It also defined what the federal government will consider as “evidence” for the purposes of complying with the Evidence Act, as detailed in Figure 2-11 below. These components clearly show the inclusions of performance management and evaluations as well as analyses and “fact-finding” for non-existent programs or activities.

Figure A. 1: Components of Evidence

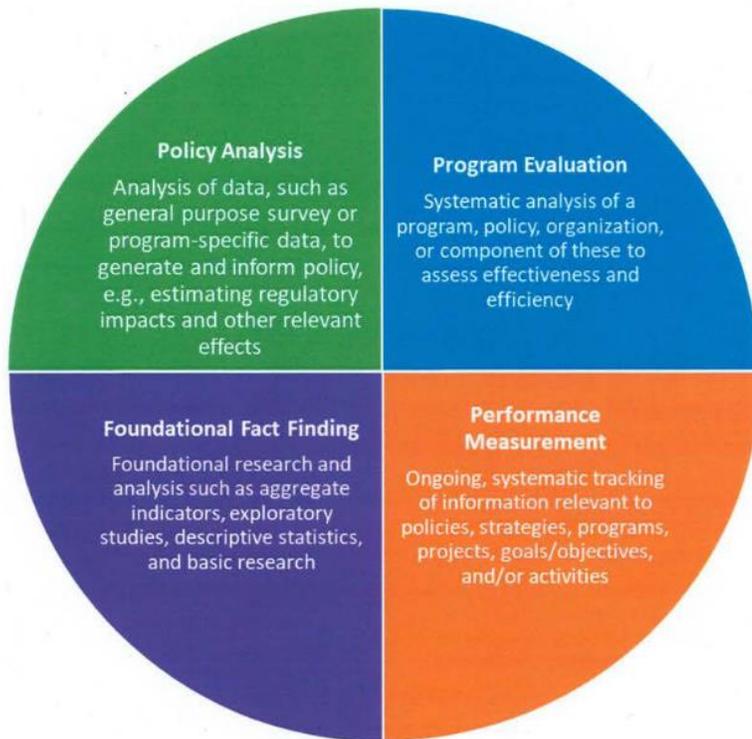


Figure 2-11: Components of Evidence (M-19-23)

The agency learning agendas are required to be linked to the agency's four-year strategic plan (as required under GPRAMA). The learning agendas should focus on building evidence that is directly linked to answering questions associated with the agency's long-term strategic

policy and operational goals (as may be codified as APGs and CAP goals). Learning agenda questions support both the use of performance management as well as set the stage and support program evaluations conducted by the agency. This guidance also focused on the establishment of annual evaluation plans. This plan must include significant evaluations in furtherance of the learning agenda, as detailed in Figure 2-12 below, as well as other evaluations required by statute. Importantly, these plans help agencies centralize the evaluation activities making it simpler for leadership to have oversight and understanding of these activities, especially as they relate to other requirements. Additionally, the Evidence Act requires the plan to be published for the public, whereas previously, many evaluation activities within the federal government were not made public, and their existence of them and their results often were not available outside the specific agency or sub-agency unit. These plans must include specific details on the question to be answered by the evaluations, information and data needed, the methods that will be used, anticipated challenges for the evaluators, and the plan to disseminate the results of the evaluations.

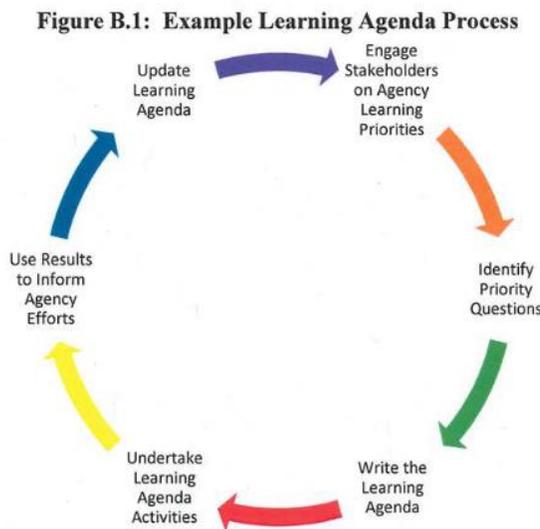


Figure 2-12: Learning Agenda Process (M-19-23)

M-20-12

In March 2020, OMB published the next Evidence Act Memorandum, “Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices”. Notably, this memorandum covering phase 4 was published before phases 2 and 3 guidance by OMB was published as they determined not to progress sequentially. Memorandum M-20-12 provides agencies' specific requirements for evaluation

standards, practices, and implementation. The required federal program evaluation standards are relevance and utility, rigor, independence and objectivity, transparency, and ethics.

<p>Relevance and Utility</p> <ul style="list-style-type: none"> • Relevancy means addressing important issues within and across agencies as linked to their strategic plans. • Utility is evidenced by moving forward the agency's mission by producing information which is clear and actionable by leadership.
<p>Rigor</p> <ul style="list-style-type: none"> • The methods of evaluation must match the evaluation goals and design and these methods must be implemented in appropriate ways to ensure the validity of the finding based on scientific theory and practice.
<p>Independence and Objectivity</p> <ul style="list-style-type: none"> • Federal Agencies must ensure inclusivity of stakeholders in evaluations, but evaluators must be appropriately separate from programmatic, policy, and budgetary offices to help ensure the objectivity of their evaluations and findings to leadership.
<p>Transparency</p> <ul style="list-style-type: none"> • This includes documenting and disseminating specific goals, methods, and findings of evaluations internal and external to the agency.
<p>Ethics</p> <ul style="list-style-type: none"> • Ensure the use of the highest standards for ethical pursuit of evaluations, especially when subjects are involved. This includes ethical standards for evaluations as well as the privacy of individuals and their data as the subjects of the evaluations.

Table 2-10: OMB Guidance on Evaluation Standards (from M-20-12)

The Memorandum also details the practices the federal agencies must follow for evaluations, which are based on discussions and reviews with the leading public sector evaluation groups and federal government resources. These practices include building and maintaining evaluation capacity in their agencies. This can be accomplished through hiring staff with evaluation expertise, leveraging inter-agency resources as well as contracting with outside expertise to plan and conduct evaluations, and providing continued professional development and training to agency staff. Effectively using expert consultations helps agencies tap into the evaluation expertise within the agency, interagency, and private sector to ensure their evaluation plans and implementation are done to the highest standards. Establish and disseminate evaluation policies and plans which also detail the methodologies and standards which will be used. Additionally, the evaluations should both plan for data collection and use for evaluations but also leverage these data for secondary uses (such as performance measurement or sharing

with the public). And lastly, the practices include steps to ensure ethical design and implementation as well as protecting the privacy of individuals and their information.

1. Build and Maintain Evaluation Capacity
2. Use Expert Consultation Effectively
3. Establish, Implement, and Widely Disseminate an Agency Evaluation Policy
4. Pre-Specify Evaluation Design and Methods
5. Engage Key Stakeholders Meaningfully
6. Plan Dissemination Strategically
7. Take Steps to Ensure Ethical Treatment of Participants
8. Foster and Steward Data Management for Evaluation
9. Make Evaluation Data Available for Secondary Use
10. Establish and Uphold Policies and Procedures to Protect Independency and Objectivity

Table 2-11: OMB Evaluation Plan Requirements (from M-20-12)

M-21-27

In June 2021, OMB issued Memorandum M-21-27, “Evidence-based Policymaking: Learning Agendas and Annual Evaluation Plans.” Notably, this OMB Memorandum provided updated guidance to agencies on topics covered in Phase 1 and Phase 4 implementation guidance, but OMB has still not issued guidance on phases 2 and 3. M-21-27 was required by the Executive Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking in January 2021 in response to many concerns the new administration was trying to address in the midst of the COVID-19 pandemic. M-21-27 reiterated much of the guidance and requirements contained in the Evidence Act and the prior OMB memorandums but an extended policy that these ideals and requirements should be championed by agency leadership and adopted by non-covered agencies as well as sub-agency bureaus and units in their work, rather than just in support of strategic goals priorities, APGs, and CAP goals. The Memorandum also extended guidance on evidence-building questions, learning agendas, and evaluation logic models.

Figure 1:

Using Evidence to Improve Agency Processes

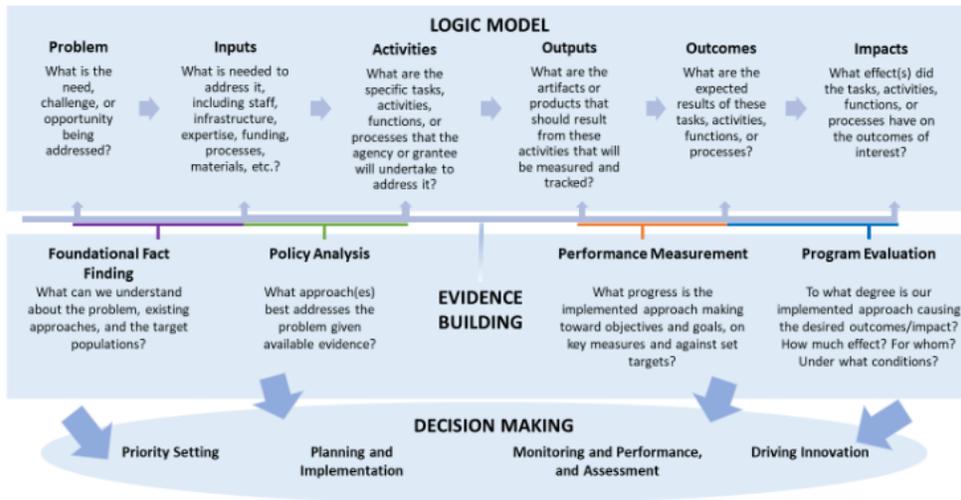


Figure 2-13: Evidence Logic Model (M-21-27)

M-21-27 also included a table of different types of evidence-building activities with definitions and methodological approaches. Of note are the definitions and methodologies listed for performance measurement as well as the different types of evaluations suggested for government agencies: Process/Implementation Evaluation, Formative Evaluation, Outcome Evaluation, and Impact Evaluation.

Table A.1: Evidence-Building Questions, Types, and Methods

For Questions Like...	Potential Evidence-Building Type	Methodological Approaches May Include, But Are Not Limited to:
<p>Did the program, policy, regulation, or organization meet its pre-established goals?</p> <p>Are program activities being effectively or efficiently performed?</p> <p>Is service delivery as effective or efficient as planned?</p>	<p>Performance Measurement</p> <p><i>Ongoing monitoring and reporting of program accomplishments, particularly progress toward pre-established goals</i></p>	<ul style="list-style-type: none"> Tracking and Reporting Key measures (often relying on administrative data) Data Dashboards Value Stream Mapping Root Cause Analysis

Figure 2-14: Performance Measurement Definition and Methodology (M-21-27)

As stated by the OMB guidance, a process or implementation evaluation is used to determine how the policy or program was implemented and if it was implemented as it was designed or required, which is sometimes known as a compliance evaluation. Formative evaluations are aimed at determining whether a policy or program is feasible as designed,

identifying those areas which do work and areas which may not work in order to provide feedback to the designers and implementors. Outcome and Impact evaluations are known as Summative Evaluations. They are aimed at measuring the results of the policy or program against the intended outcomes and impacts, respectively. An important distinction is that impact evaluations attempt to determine causality between the program and the observed changes, whereas outcome evaluations assess the effectiveness of the policy and the implementation but cannot assess causality between the policy and the outcomes.

<p>Was the program, policy, regulation, or organization implemented as intended?</p> <p>How is the program, policy, regulation, or organization operating in practice?</p>	<p>Process/Implementation Evaluation</p> <p><i>Assesses how the program, intervention, operation, regulation is implemented relative to its intended theory</i></p>	<ul style="list-style-type: none"> • Structured Observations • Qualitative Interviews and Focus Groups • Ethnography • Statistical Analysis of Program or Participant
	<p><i>of change, and often includes information on processes, content, quantity, quality, and structure of what is being assessed.</i></p>	<p>Data (Administrative Data and Surveys)</p> <ul style="list-style-type: none"> • Document Reviews • Time Studies
<p>What aspects of the program, policy, or organization do not seem to be working as intended?</p> <p>Can the program, intervention, policy be implemented as designed?</p>	<p>Formative Evaluation</p> <p><i>Typically conducted to assess whether a program, policy, or organizational approach—or some aspect of these—is feasible, appropriate, and acceptable before it is fully implemented. Can include process and outcome measures.</i></p>	<ul style="list-style-type: none"> • Pilot Projects • Structured Observations • Qualitative Interviews and Focus Groups • Case Studies • Statistical Analysis of Program or Participant Data (Administrative Data and Surveys) • Community-Based, Participatory Research
<p>Were the intended outcomes of the program, policy, regulation, or organization achieved?</p>	<p>Outcome Evaluation</p> <p><i>Measures the extent to which a program, policy, or organization has achieved its intended outcome(s), and focuses on outputs and outcomes to assess effectiveness. Cannot attribute causality.</i></p>	<ul style="list-style-type: none"> • Qualitative Interviews and Focus Groups • Statistical Analysis of Program or Participant Data (Administrative Data and Surveys), including Longitudinal Data • Data Linkages
<p>Does it (intervention, policy, program, regulation) work? Or, for whom does it work, under what conditions, and compared to the alternatives?</p> <p>Did it (intervention, policy, program, regulation) lead to the observed outcomes?</p>	<p>Impact Evaluation</p> <p><i>Estimates and compares outcomes with and without the program, policy, or organization, or aspect thereof, usually seeking to determine whether a causal relationship can be established between the activity and the observed outcomes.</i></p>	<p>Includes:</p> <p><i>Experimental Designs</i></p> <ul style="list-style-type: none"> • Randomized Controlled Trials <p><i>Quasi-Experimental Designs</i></p> <ul style="list-style-type: none"> • Difference-in-Difference • Regression Discontinuity • Propensity Score and Other Matching Approaches • Instrumental Variable Modeling <p>For each design:</p> <ul style="list-style-type: none"> • Pilot Projects

Figure 2-15: Government Evaluation Types and Methodologies (U.S. Government Accountability Office, 2021a)

Evaluation Officer Council

The Evaluation Officer Council is coordinated by the Evaluation Team at OMB and made up of Evaluation Officers from the federal agencies. The EOC exists to exchange information and best practices, coordinate federal evaluation standards and practices with OMB, coordinate and collaborate on common areas of interest between agencies, and serve as a leadership community for evaluation efforts throughout the federal government (*Evaluation.Gov*, 2022). The activities, guidance, and materials promulgated by the EOC are posted for the public on www.evaluation.gov, which also serves as a repository for OMB evaluation policy and guidance. This site also provides links to all agencies' learning agendas, evaluation plans, evaluation policies, capacity assessment, and their completed evaluations by the agency and fiscal year. Beyond these materials, as well as other evaluation-related laws and regulations, there are not many other resources made available publicly for agencies from the council.

Assessment of Evidence-Based Policymaking

Through these recent efforts, there is increased scientific information utilization by federal government agencies. However, just as successful performance management has been defined as the actual "use of performance information for decision-making," it is also important to look at how, when, and if policymakers use the results of these evidence-based methods. While the use of evidence-based methods in the government rates differs based on the knowledge, education, and experience of government employees, the strongest predictor of use has been the relevance of scientific research to the government agency and linkages from the academic community to the public administrators (Landry et al., 2003). While scientific evidence is highly valued by most policymakers and legislators, it is typically used in three different ways. First, it might be utilized to gain credibility regarding their current political position on a policy. Secondly, it might be utilized to undermine a political position of a rival position. Lastly, it might be used to inform changes to a current program or policy position. What is noticeably absent from this list is the creation of a new policy or political stance. This is not surprising when you understand that elected officials are mainly focused on responding to their constituencies, and these are rarely waving scientific evidence at their elected leaders (Bogensneider & Corbett, 2021b). Similarly, several researchers have shown that policymakers (legislators and public administrators) who do utilize scientific research view it as only one type of necessary information needed to make decisions. Policies and politics reflect

diverse values, concerns, and opinions, and not all of these can be dealt with through scientific evidence (Arinder, 2016; Fobia et al., 2019).

Through the legalistic paradigm, policymakers need to be concerned with viewing policy and programs in conjunction with their application to and compliance with existing laws. Since not all legislators consult scientific evidence when drafting legislation, it cannot be expected that these laws would comport to science. Indeed, multiple researchers have pointed out that the mere availability of scientific evidence regarding a policy or program is not correlated to whether or not policymakers avail themselves of that information (Jennings and Hall, 2011).

Several researchers have pointed to the unevenness of evidence-based policymaking across government agencies. Some are due to the nature of the agencies' work. For instance, the Food and Drug Administration has a long history of RCTs required prior to approving a new drug for the market, while other agencies, such as the postal service, are exploring how scientific methods and evidence can inform their policies and processes (Jennings and Hall, 2011). Additionally, these two examples infer other necessary conditions for the adoption and implementation of evidence-based policymaking. The FDA has numerous medical doctors, scientists, and scientifically trained staff who have knowledge, experience, and appreciation for scientific methods and evidence. The U.S. Postal Service would not, by default, have many similar employees, so they would need to proactively seek out and create an office with this level of knowledge before they could adequately implement the creation and usage of scientific knowledge (Adam et al., 2018; Fobia et al., 2019). Therefore, the importance of the infrastructure for the creation of data and evidence-based activities is well-founded but not yet adequately implemented or assessed to understand how they will impact agencies' abilities and the resulting usage of their abilities to create scientific data about their programs.

In fact, research has begun to show that the availability of scientific evidence, staff resources to ingest and utilize the evidence, and mission and mandates are necessary but not sufficient conditions (Adam et al., 2018; B. W. Head, 2010). There is some evidence that the leadership and political atmosphere of an agency will also impact the usage of evidence-based policymaking. On the one hand, if agency leadership values scientific evidence, then they are much more likely to utilize it. However, there is also evidence to show that a highly politically controversial agency mission inhibits the use of scientific evidence and increases more political and value-based policymaking instead. The causal theory behind this is that these politically controversial agencies and policies are often the battlegrounds of elections and important issues which divide voters and decide elections. In this environment, values become overriding, and at

best scientific evidence is less of a factor. At worst, society will witness instances of defamation of scientific evidence or carefully chosen scientific evidence that props up the value-based policy decisions. Many have referred to the latter issue as policy-based evidence (Cartwright and Hardie, 2012; Jennings and Hall, 2011; Heinrich, 2007).

Criticisms of Evidence-Based Policymaking

While there is almost universal acceptance of the ideals behind evidence-based policymaking, the processing of information is fraught with problems and is incongruent across federal organizations based on at least two factors that undermine evidence-based attempts. The first constraint is politics in the form of a politicized policy arena that is highly partisan. In highly political policy areas, values, negotiations, and persuasion often crowd out the use of expertise and data to guide decisions or problem discussions (B. Head et al., 2014; B. W. Head, 2016). The research argues that when politics drives policy decisions, there must be a policy resulting from conflicts, trade-offs, and compromises that inherently will not align with the data and evidentiary recommendations (Jennings and Hall, 2011). Another factor is the diversity of organizations and their leadership. Different agencies and their senior leadership will vary according to their preferences and structure. The types of organizations (such as service delivery, regulatory oversight, or policy development) and the domains (such as social policy, economic development, or environmental regulation) will also either inhibit or nurture evidence-informed policy implementation and use (Head, 2015).

Others have been quick to point out that the scientific benefits of the “gold standard” of evidence-based policy are an ideal that is often not attainable for political or practical reasons or not applicable to the policy problem being contemplated. This gold standard, of course, is a randomized controlled trial, or RCT (Khagram and Thomas, 2010). Many researchers laud the rigor of the methods of RCTs, and legislators have required agencies to only adopt programs that have withstood and been deemed successful by one or multiple RCTs. However, many have pointed out that RCTs are not always achievable or even the best method for certain research programs. Additionally, others are concerned about the foundational issues of structural changes and increases in privacy protection and data sharing capabilities as the first steps before RCTs are advisable. Until these are achieved, they say, the government should not take the next steps toward trying to mandate RCTs (Adam et al., 2018; Commission on Evidence-Based Policymaking, 2017; B. W. Head, 2016).

Other researchers want to make sure that the deification of RCTs does not blind policymakers and administrators to misinterpret what RCTs actually mean and install policies

and programs in places or situations where they will not be implementable or not effective. Just because an RCT showed that a policy or program worked for a specific instance, that does not mean that you are guaranteed to see the same or similar results in other instances (Cartwright & Stegenga, 2011; Khagram & Thomas, 2010). Instead, people must strive to understand the RCT data in a causal way so that one can evaluate not just that the program worked, but why it worked, and what the necessary and sufficient factors allowed it to be implemented successfully and experience the effects that it did. Practitioners need to invest a great amount of intellectual capital in determining whether to adopt a policy or program based on its success elsewhere in their instance.

Another viewpoint is that RCTs and quantitative studies alone cannot provide researchers or public administrators with the full toolset by which to evaluate policies and programs. The “gold standard” of RCTs and quasi-scientific methods needs to include well-designed and conducted qualitative studies to create the “platinum standard” of program evaluation. Public administrators and policymakers alike need to identify the proper research and evaluation methods for the program. Simple reliance, or insistence, on one method alone not only limits the ability to evaluate certain policies and programs but may contribute to efforts to create or stretch data which will lead to invalid results (Khagram and Thomas, 2010).

Other research has made the case that RCTs need to be augmented by early evidence about policies and programs to inform policymakers and administrators. This is because of the long time and high cost of many RCTs and the fact that no single RCT can definitively show that a policy or program will be successful or impactful. However, early evidence can show that the policy or programs will be cost-prohibitive, unethical, or politically infeasible. Therefore, policymakers’ ability to develop and utilize early evidence is necessary for the successful implementation of evidence-based policy making. Such early evidence should include estimates of program costs, evidence of processes required to maintain implementation, and quasi-experimental research regarding subgroup effects (Crowley & Scott, 2017).

Three areas of early evidence are strategic review, guiding standards, and active communication. A strategic review is a process of continually updating analysts with the current research in the field, policies, and programs. This enables policymakers and administrators to be up to date on all advances and prepared to take on new evidence and information as it may affect their program. Guiding standards are necessary to ensure that as early evidence is gathered and the organization has directed methods to signal quality and potential impacts will be evaluated and synthesized by all. This avoids potential ambiguity about what will be considered and how

it will be utilized. Lastly, active communication controls how early evidence will be communicated ahead of time, as well as clear parameters regarding the communication to internal and external stakeholders about what conclusions can and cannot be drawn from the early evidence (Crowley and Scott, 2017). I believe that this need for early evidence in the public sector is a requirement that can be met through the closer integration of performance management and evidence-based methods.

There is also criticism and concern that a focus on scientific evidence and evidence-based policy is premature. The administrators and researchers still need to lay the foundation of this movement before researchers can truly evaluate how well it is implemented or how effective it is in building successful programs or policies. The academic community needs a better consensus on what counts as “evidence” and how it is to be produced by what methods to ensure rigor and validity of measurement as well as of conclusions. If it extends beyond statistical data to qualitative assessments, then researchers need to include the perspectives and experiences of legislators and public administrators on an equal footing when considering the evidence. Therefore, our government should land closer to a working definition that evidence-based policy is about using scientific evidence to produce better public outcomes (Heinrich, 2007).

We also need to focus on the extent to which government organizations and policymakers are utilizing data and evidence in policy making and implementation decisions. As the movement has grown and with the adoption of the Evidence Act, there has been increased emphasis on the utilization of data and pilot evidence when designing and implementing programs, and democracies around the world have increasingly been focused on better design, improved effectiveness, and increased efficiency of domestic and international programs. However, while there are reports of increased data creation by government agencies, there have been few studies regarding how this data is actually used by government organizations and policymakers ((Arinder, 2016; Fobia et al., 2019; B. W. Head, 2016). There is evidence that the processing of information is fraught with problems and incongruent across federal organizations because of political values and partisan policy areas. In highly politically controversial policy areas, values, negotiations, and persuasion often crowd out the use of expertise and data to guide decisions or problem discussions. When politics is driving policy decisions, there must be result of conflicts, trade-offs, and compromises that inherently will not align with the data and evidentiary recommendations. The second factor is looking at the diversity of organizations and their leadership. Different agencies and their senior leadership will vary according to their preferences and structure. The types of organizations (such as service delivery, regulatory

oversight, or policy development) and the domains (such as social policy, economic development, or environmental regulation) will also either inhibit or embrace evidence-informed policy (Head 2015, p 473).

Community Dissonance Theory

One theory that is emblematic of the problem resulting from the disconnect between the implementation of Evidence-based approaches and the use of their results by policymakers is Community Dissonance Theory which divides stakeholders between “knowledge doers” and “policy doers.” This theory views this problem as stemming mainly from cultural differences between knowledge producers and knowledge users. Knowledge producers include researchers who are theoretical academics; applied researchers who go beyond theory to put their programs into testable action. Knowledge users are classified as either policy doers or policymakers. Policy doers are public administrators responsible for designing, implementing, managing, or applying the policies. Policymakers are those responsible for deciding to adopt a policy and setting the direction of that policy. In Community Dissonance Theory, there also exist intermediaries who attempt to bring the knowledge producers and knowledge users together. These can be members of either the knowledge producers or users or true intermediaries such as think tanks (when not solely producing knowledge), lobbyists, non-governmental organizations, and legislator staff who seek out knowledge producers on particular issues (Bogenschneider and Corbett, 2010).

These camps are important to view in their separate cultures to understand what they value and what motivates both entities. For instance, knowledge producers often are drawn to abstract knowledge topics, advancing theoretical understandings, or methodological issues. Knowledge users, on the other hand, are much more focused on applied research in topical areas. Additionally, researchers are interested in the work of other researchers, and many view their peers as their main target audience. Knowledge users are typically looking at their constituents, lobbyists, the media, and their peers as their main audiences (Bogenschneider & Corbett, 2021b).

Cognitive frameworks are also very different between the two camps. Researchers approach topics cautiously and are fairly deferential to counterfactuals, and will the null hypothesis. This typically results in conclusions that are couched in. Since so much effort is in identifying statistical significance, it is often a surprise when knowledge users do not find significance as important as the size of the effect, which potentially is only one individual. It is an important lesson about politics to understand the power of anecdotes. Timeframes are important as well. Researchers often think of their work as somewhat timeless; they are adding

their research and knowledge to a long-standing and continually improving paradigm of theory and research findings. Knowledge users, on the other hand, often are elected for very short periods of time in which they are pressured to “make progress” for their constituents in order to have a chance of being reelected. Added to that are small “policy windows” in which they have to further specific goals, many of which are opened and closed outside of their ability to predict or control. Combining all of these differences in how they approach the world and policy issues, it is no wonder that many find very little interaction between knowledge producers and knowledge users (Bogenschneider and Corbett, 2010).

Suggestions regarding creating closer ties between the two entities with the goals of establishing greater scientific research use by policymakers (Bogenschneider & Corbett, 2021a, 2021c). While this is understandably not a goal for all knowledge producers, many academics wish to use their research and knowledge to inform policy making and to help policy-makers make informed policy decisions. Beyond this, there is a belief that research institutions and institutions of higher education should do more to train researchers in how to make their knowledge more accessible to policymakers, to be able to interact more freely and focus research on areas of applicability and need of current policymakers. In order to help accomplish this, researchers should make an effort to understand the legislative and policy-making process, which is complex and very different from academics.

Additionally, researchers should become more familiar with policymakers and politicians to understand the skills and values required to be successful in that arena (Bogenschneider & Corbett, 2021a). Most researchers (and policymakers) often have negative views about each other that are most often completely dispelled after greater interactions. Researchers should also be prepared to answer questions and respond with information to policymakers in a very short timeframe. This often means having research answers prepared prior to a request being made. Researchers also need to distill their findings down into very short and plain-language explanations (Bogenschneider & Corbett, 2010). This is often difficult as academic writing is focused on explanations of methods in order to allow for repeatable research, explanations of validity and reliability, as well as literature reviews. This type of writing and presentation is typically counterproductive outside of the academic realm. Researchers also need to be careful with objectivity. They need to understand the role of values in politics and ensure they present their research in an objective manner that includes opposing views. Policymakers spend a great deal of time being lobbied on all manner of issues. Therefore if researchers are viewed to be lobbying on behalf of their research or program, the impact of their findings can be lost. Perhaps

the most important is that researchers need to appreciate the power of the constituency on policymakers. If you can present research in a way that also shows the impact and effect on the legislator's district, this will go a long way from taking the general abstract research finding and turning it into a concrete understanding of the impact on him (Bogenschneider and Corbett, 2010).

Use of Evidence by Policymakers

Despite the above criticisms and concerns, the evidence-based policy movement is viewed as a boon to our public sector, yet many legislators and public administrators are not relying on scientific evidence when making many policies and program decisions. Several researchers have begun to look at this problem in great detail. One finding is that there may be very different causes of the usage between legislators and public administrators. Let's take elected politicians first. It will likely come as a disappointment to many academics and U.S. citizens alike to find out that when asked what amount of a political policy decision was accounted for by scientific evidence or research, many answers were in the range of five to ten percent (Bogenschneider and Corbett, 2010.) To those close to legislators or who have worked in the lawmaking capacity, this might not come as much of a surprise.

There are many factors for this. First, elections demand that legislators are responsive to their constituencies, or they will cease to be elected. Another factor is that politics, elections, and preference for candidates have been linked more closely with values and value-based issues than to complex scientific evidence-based decisions (Bogenschneider and Corbett, 2010). There is also a distinct separation between the information utilized to adopt programs and policies and the types of information used to evaluate the efficiency and effectiveness of policies and programs. Unfortunately, there is often little connection between adoption and implementation by legislators, and therefore the implementation evidence does not seem to get much consideration when debating the adoption (Hall and Jennings, 2008).

Sub-legislature Evidence-based Policymaking

It is a fact that policy-making happens beyond the legislature. There is a great deal of policy-making that happens within agencies and even at the individual bureaucrat level. It is important to also focus on how evidence-based policymaking can be expanded to agencies and policymakers outside of the legislature. Similarly, one of the primary focus areas needs to be on the utilization of scientific evidence by public administrators and agencies. While some agencies utilize scientific studies, both RCTs and quasi-experiments, these are often either

mandated and conducted by the agency or funded and monitored by the agency for a specific purpose. Strong examples are the FDA's drug trials, EPA's environmental studies, and the CDC's work in combating diseases. These are well-integrated scientific programs, but they are obviously called for by the mandate and tools involved in their field. What is lacking is the same scientific rigor and evidence in non-hard science areas of the federal government. This includes the use of scientific methods and data informing the delivery of services, whether it is the IRS collecting taxes or the USPS delivering mail. There are also policy agencies that gather intelligence or conduct foreign affairs, which need to utilize the same scientific evidence and data methods and techniques to inform not only how they collect but also analyze information to report to policymakers. These are less obvious but perhaps more important areas that thus far have received little attention (Commission on Evidence-Based Policymaking, 2017; Head, 2015).

The 2017 Commission on Evidence-based Policymaking looked at the federal government and specifically focused on policymaking in federal agencies. What they found was not that surprising. They found that some agencies, and some parts of agencies, utilized internal and external scientific evidence consistently in both policymaking and program evaluations. Examples of this are the Chief Evaluation Officer at the Department of Labor and the Health and Human Service's Data Council (A. Kroll & Moynihan, 2017). However, they saw that these uses are very disparate between agencies and even within agencies. Even where it exists well, there are areas of the agency which do not engage in any evidence of usage in policy decisions. One recommendation is that OMB must mandate that each agency creates a centralized office that is responsible for the proliferation of evidence-based policymaking in the agency. This office would supply the knowledge, expertise, and coordination of data within each agency. OMB then would coordinate among these agency offices to comprise a federal government-wide network of offices where data could be shared, standardized, and culture could be created, which encouraged the use of evidence in almost all areas of agency decision making. Another issue that needs to be tackled is that as agencies implement programs, modify programs, and administer programs in new ways. There is no consistent process for making long-term plans for data collection, which will allow for program evaluations and evidence-based decisions regarding the program (Commission on Evidence-Based Policymaking, 2017).

Summary and Next Steps

It is clear that there is much focus on the promise of evidence-based policymaking in the federal government. Despite significant concerns and criticisms, most offer adjustments and

future efforts rather than concluding that this effort should be abandoned or that it will ultimately fail. One important initiative is the focus on bringing together evidence-based policymaking, evaluation, and performance management, which have very inter-related goals, though with sometimes disparate methods to create requirements and tool kits for government agencies, administrators, and policymakers. The Evidence Act has established the legal requirement for the federal government to implement several organizational structures and processes to help them establish the use of evaluations and performance measurement data in policymaking. There have not yet been many meaningful studies on the success of this. However, we can still build on these requirements to apply them to areas where we wish to have a better understanding of the impact of policy and process changes, such as our attempts to lower administrative burdens through the implementation of machine learning programs.

Chapter 3 - Research Design and Methods

The Research and Solution

In the preceding chapters, I have shown how the underlying theories of performance management, artificial intelligence, and evidence-based policymaking are important in the public sector and how they can be applied to help us identify, measure, and solve the problem of administrative burdens in federal programs. In this chapter, I will outline and explain the three frameworks that I am building in this research, the solutions they bring to my overall research focus, how they fit together but also how they could be used independently, and how they are built upon existing theory and practice in the federal government. The importance of this section lies in understanding how I will set out to achieve my research goals and what steps I will take in my research.

As a reminder, the overall focus of my research is that administrative burdens are important phenomena in government interactions when individuals or organizations are seeking benefits or services from the government (Burden et al., 2012; Herd & Moynihan, 2018). The existence of administrative burdens directly impacts the efficiency and efficacy of programs and policies. The larger the administrative burdens, the lower the efficiency and effectiveness of the programs, which directly decreases the overall program outcomes, especially for some of the neediest potential beneficiaries (Heinrich, 2007; Herd et al., 2013; Herd & Moynihan, 2018, 2020). There are many potential solutions to reduce or eliminate administrative burdens, but I will explore the use of machine learning to lower learning costs, compliance costs, and psychological costs. To accomplish this, I will show how administrative burdens can be identified and measured through existing performance management requirements and practices, and how machine learning solutions can be designed and implemented in government programs to reduce or eliminate administrative burdens. And finally, how to leverage evidence-based policymaking requirements and practices to evaluate the effectiveness of machine learning solutions in reducing administrative burdens as well as the overall policy or program outcomes.

My research starts with and extends the theoretical basis of administrative burdens as conceived and extended by Burden et al. (2012) and the seminal work by Herd and Moynihan (2018). While my research will be grounded in performance management and evidence-based policymaking, I am not seeking to recommend changes to these areas. Rather, I am acknowledging their importance and primacy in the federal government and therefore using them to ground and extend my research into administrative burdens. Additionally, my research is

focused on exploring the use of machine learning in the public sector. I will gather and combine the contemporary work and focus on this emerging issue, but I will do so through applied example solutions to reduce or eliminate administrative burdens. I believe this is important because it must explore and be guided by the important legal, policy, technical, and ethical considerations, but it will do so with a clear goal and the desired outcome clearly fixed rather than from a theoretical exploration.

Of the three frameworks, I expect Framework 1 to be the most immediately applicable since it will create a clear path towards using performance management practices and requirements to identify and measure administrative burdens. Framework 2 will explore just one potential solution to reduce or eliminate administrative burdens, machine learning. However, many government agencies are not necessarily ready to implement AI solutions as they are still building the foundations of databases, technology, and expertise before they will be ready to tackle advanced data analytics solutions. Framework 2 is still important because it will walk through potential solutions for administrative burdens and highlight the precursors and design and implementation considerations of machine learning in the public sector. Before agencies are ready to implement Framework 2, they might also consult it to better understand the steps to work on to prepare themselves for these solutions.

Framework 3 might also be less readily accessible but is important nonetheless. Framework 3 will focus on tying together the requirements of evidence-based policy and the benefits of public sector evaluation in an applied manner which can both be used to measure the effectiveness of programs but also measure the impact of changes (such as machine learning solutions on administrative burdens). With the advent of the Evidence Act and similar requirements at the state and local levels, it is important to have guidance on how to use these requirements for policy and program decision-making to leverage them towards the greatest benefit rather than potentially turning them into compliance exercises which provide little or no value or benefit to policymakers and administrators.

As stated previously, these three frameworks could be used independently of each other towards their particular solutions. This may be appropriate in certain circumstances. For instance, using Framework 1 to identify and measure administrative burdens through performance management may be enough when implementing a performance management system and wanting to be aware of administrative burdens in a program. It could be that is all that is needed because their presence is small or non-existent. Or the agency might not yet know that they want to or have the resources to reduce the administrative burdens. I expect researchers

and administrators will ultimately want to implement all three frameworks, though, which would be ideal for programs wishing to reduce administrative burdens.

The Three Frameworks

This research is about identifying and measuring administrative burdens by using performance management. It is about applying machine learning to reduce or eliminate administrative burdens. And it is about using evidence-based policy and evaluation methods to measure the impact of the machine learning techniques on administrative burdens as well as the impact of the reduction of administrative burdens on the overall outcomes of the policy or program. To accomplish these goals, this research will develop three different frameworks to cover each of these areas. These three frameworks can be used independently but will be designed to work together, building on each other to help accomplish all of these goals.

As shown below in figure 1, these frameworks are distinct but related. The frameworks build on each other. Framework 1 is foundational because it uses performance management to measure administrative burdens' three costs. Framework 2 uses the identification and measurements of administrative burdens from Framework 1 to design and implement machine learning solutions to reduce the burdens. And Framework 3 uses evidence-based policymaking and evaluation requirements and theory to measure the impact of Framework 2's solution on administrative burdens as well as the impact on the overall program or policy outcomes. These three frameworks together will provide researchers and administrators a toolkit to use when implementing machine learning to reduce administrative burdens in the government, especially one that will ensure they can measure the impact of their solutions.

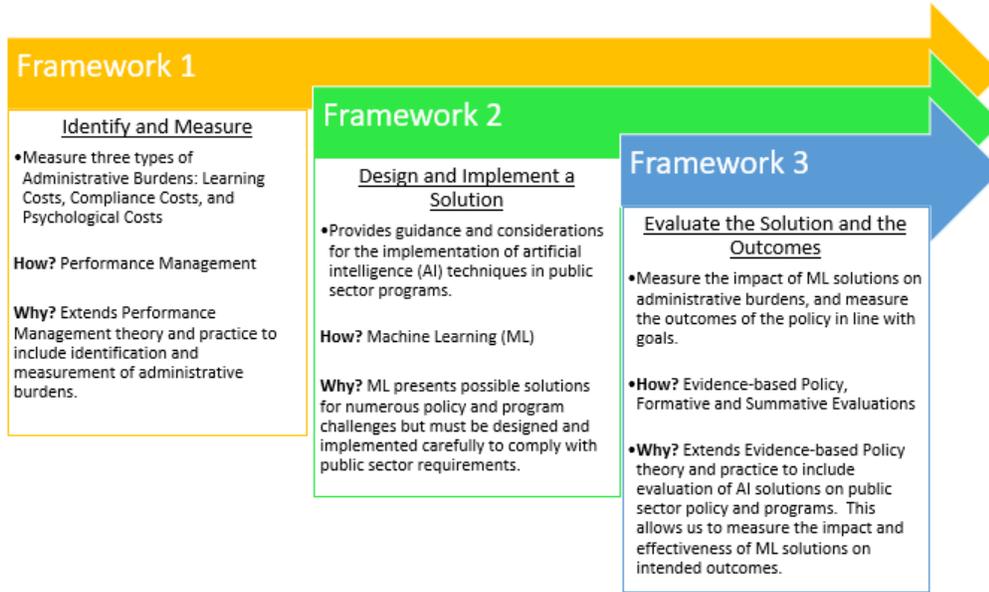


Figure 3-1: The Three Frameworks (author)

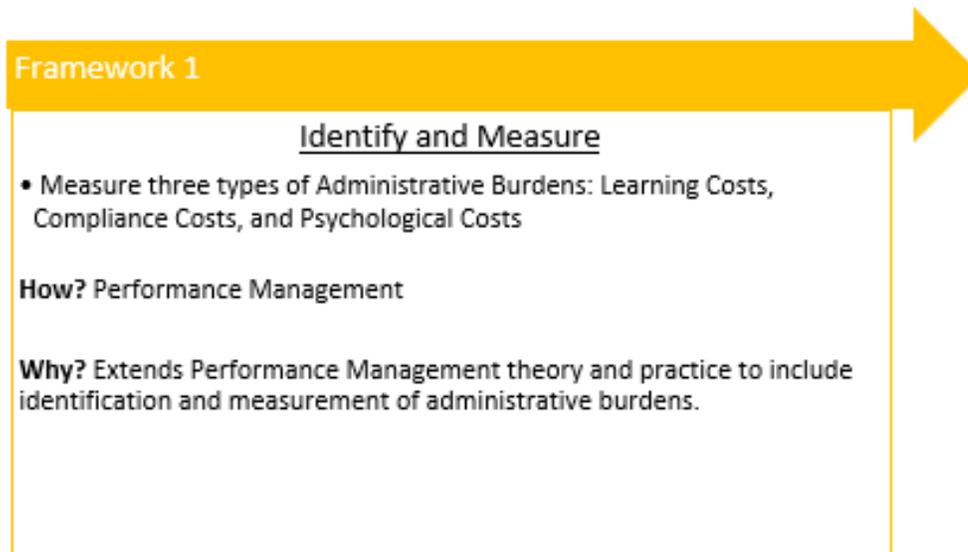
Framework 1 - Measuring Administrative Burdens within Performance Management

Figure 3-2: Framework 1 (author)

The public administration research on administrative burdens is growing rapidly. There is a focus on identifying new instances of burdens, new types of burdens, and new causes of burdens, including inadvertent burdens but also intentional burdens to achieve certain policy goals. However, what is currently lacking in the research is a mechanism to identify and measure these administrative burdens. If these phenomena are as present as claimed to be, and if they have a significant impact on policy and program outcomes as claimed, then the field of public administration needs a formal means to identify and measure them. This would help ground the theory and research of administrative burdens in an empirical epistemology as proposed by Herbert Simon (Ricucci, 2010). Framework 1 will attempt to achieve this by providing means and processes to empirically identify and measure administrative burdens in programs by placing them within a performance measurement process. This means that existing performance management systems will be extended to include key performance indicators for administrative burdens.

To do this, I'm going to extend the research and practice of performance management to include measurement and integration of administrative burdens in performance information processes. This will leverage the research and practice of performance management, which is quite robust and common in the federal government of the United States, to be inclusive in identifying and measuring these newly focused-on phenomena. I believe that this direction will lower the costs of administrators and researchers to begin to track and measure administrative

burdens because it builds on existing processes, frameworks, and performance measurement systems rather than necessitating stand-alone or new processes. It also relies on ubiquitous concepts, theories, and literature for how to best measure and report performance data which means that the government can hopefully more easily help people understand what administrative burdens are, how to measure them, and why they matter in terms of the overall program of policy.

I have already discussed the existence of administrative burdens, which are frictions in interactions between individuals and organizations when seeking services or benefits from the government (Burden et al., 2012; Herd & Moynihan, 2018). These burdens have been defined through the existence and magnitude of three types of “costs”: learning costs, compliance costs, and psychological costs (Herd & Moynihan, 2018). I have discussed the importance of understanding these costs because they directly impact the efficiency and efficiency of public programs and policies (Moynihan et al., 2015). Because of this importance, I want to be able to identify and measure these costs to understand the overall administrative burdens within a specific program. However, there is no agreed-upon framework or mechanism to measure administrative burdens within programs. Having a standard method for administrators and researchers to identify and measure administrative burdens will improve and extend the theory and research in this area. It will allow for practical application of the administrative burden research and facilitate additional research by standardizing measurement methods allowing for comparison across time and programs.

Performance management isn't the only possible solution to administrative burden measurements. I have already discussed the history of performance management in the federal government. Similar to the desire to reduce administrative burdens, performance management is born out of a desire and goal to increase the efficacy and efficiency of government programs, as well as to more accurately and transparently track and report the expenditures of tax dollars and resources in particular programs (Dooren et al., 2015; A. Kroll & Moynihan, 2017). There are several laws and policies which require government agencies to implement and use performance management processes (Gerrish, 2016; Moynihan & Kroll, 2016). And many agencies have successfully implemented but also begun to realize the benefits of performance management, especially on sub-agency projects and programs (Destler, 2016; Hassan & Hatmaker, 2015; U.S. Government Accountability Office, 2017). Therefore, rather than create new techniques and processes which require funding, resources, and attention from government agencies, I will use existing performance management requirements, processes, and information to identify and

measure administrative burdens. In some instances, this will merely require programs to clarify process steps and assign labels identifying the costs of administrative burdens, and in other cases, they may need to produce some new performance data to accomplish this. However, it will reinforce and extend their existing and required performance management processes and practices.

The result of Framework 1 will be a tool that researchers and administrators can use to implement or extend performance management processes to include the identification and measurement of administrative burdens, which are further identified as either learning costs, compliance costs, or psychological costs. The tool can be applied to many programs and policies so that administrative burden measurement is routinized as part of performance management processes to further increase transparency and public awareness. Additionally, the standardization of measurement made possible through the tool will allow further research on administrative burdens within and across programs because of the increased ease in measurement and comparability.

Framework 2 - Applying machine learning to Reduce Administrative Burdens

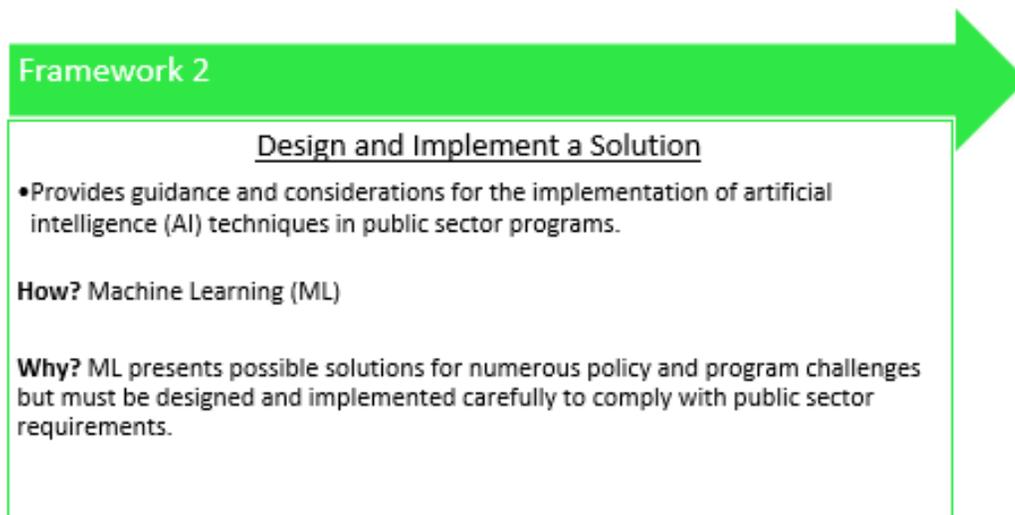


Figure 3-3: Framework 2 (author)

Once I have built Framework 1 to identify and measure administrative burdens within the performance management systems, I need to identify solutions to reduce or eliminate administrative burdens. The administrative burden research is growing with natural experiments, theories, and case studies on the impacts of decreasing administrative burdens. But there does not yet exist a practitioner’s guide to reducing or eliminating them. There have been several suggestions about how administrative burdens can be reduced, but these have not been built upon

substantially to show how these changes might be implemented, and the impact them measured thoughtfully. Some point to the possibilities of leveraging behavioral public administration, which incorporates theories and solutions from behavioral sciences to change default actions based on understandings of sociology and psychology of choices and behaviors (Herd et al., 2013; Herd & Moynihan, 2018; Sunstein, 2018). Others point to solutions based on administrative procedures, especially those designed to highlight, measure, and reduce paperwork and processing burdens on the public (Burden et al., 2012; Sunstein, 2020). Others have focused on transparency of administrative burdens in programs to engender public engagement with policymakers and administrators about the costs and impacts of administrative burdens while debating the balance between program access and needs to verify eligibility and reduce fraud (which are often the purported justifications of burdens) (Burden et al., 2012; Hamburger, 2014).

In this research, I am going to focus on a field that combines a number of the above with increased civic technology solutions, specifically the use of artificial intelligence as enshrined in machine learning. There is growing research on the public sector's use of big data and artificial intelligence techniques to help administer government programs and policies (Agrawal et al., 2018; Desouza & Jacob, 2017; Kim et al., 2014). Machine learning is a promising field where government agencies can use a growing repository of performance data, administrative data, interagency data, and open-source data to build solutions that can more accurately and efficiently automate many tasks and decisions required to be processed by the government, resulting in multiple benefits to administrators and program beneficiaries.

Therefore, Framework 2 starts with the performance information about the existence and magnitude of administrative burdens provided by Framework 1 in a government program. Framework 2 then helps the design and implementation of possible machine learning solutions to reduce or eliminate those administrative burdens. The importance of Framework 2 is the focus on purpose-driven machine learning solutions in the government sector. It focuses on the adoption of these technologies and techniques on known and measured problems. It also helps administrators and researchers methodically think through the unique challenges of artificial intelligence solutions in the public sector to help guide the solutions while addressing in a formulaic way the challenges and considerations of legal issues, ethical concerns, the questions of transparency for public programs, the requirements and needs for legal review as well as administrative and congressional oversight, as well as addresses well-known concerns about biases, and design flaws in machine learning algorithms.

Specifically, Framework 2 will be about design and implementation factors for machine learning solutions in the federal government. As mentioned, this will track closely with private sector machine learning frameworks, except that it will also need to account for the requirements and considerations of public sector solutions. These include public transparency and notification such as required in administrative law (*Administrative Procedure Act (5 U.S.C. Subchapter II)*, 2016), judicial oversight, and review considerations and processes (Busuioc, 2020; Calo, 2015; J. A. Kroll et al., 2017), and ethical design and implementation considerations (Franzke et al., 2021; O’Neil, 2016).

The creation of Framework 2 will provide researchers and administrators with a clear tool to design and implement machine learning solutions in the federal government to reduce or eliminate administrative burdens. This tool will help provide practical guidance on machine learning solutions, but more specifically aimed at the reduction of administrative burdens. This will also extend and improve research on administrative burdens because it will ensure consistent application of machine learning solutions allowing for comparability and analysis of the results between programs which will help us better understand the effects of and interactions of administrative burdens. I believe it will also extend and support the research on machine learning in the public sector by focusing on a particular solution (administrative burdens) across multiple programs and policies. This will provide increased comparable examples and data, which will further fuel public sector machine learning inquiry and research.

Framework 3 - Evidence-based Policymaking to Evaluate Outcomes After Applying machine learning to Reduce Administrative Burdens

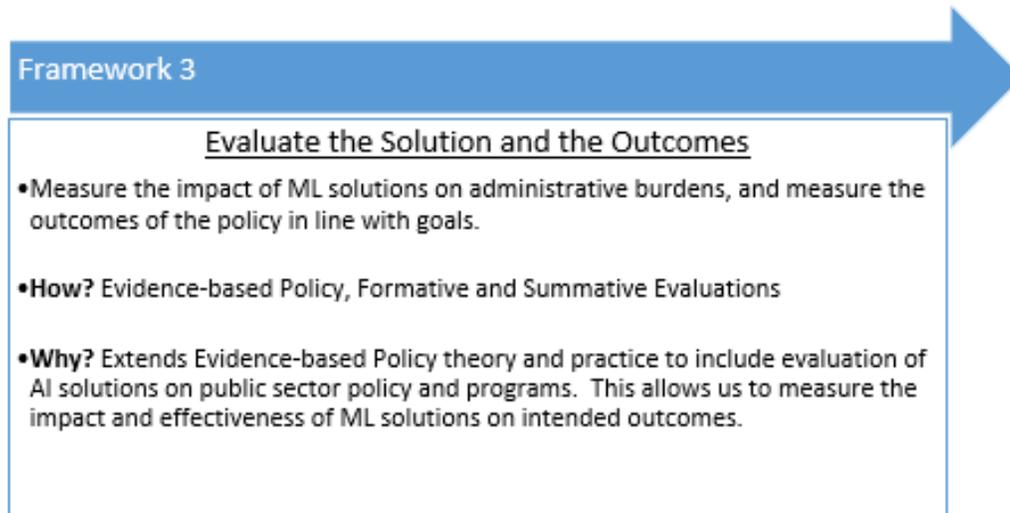


Figure 3-4: Framework 3 (author)

Thus far, I have proposed Framework 1 to help identify and measure administrative burdens as part of a performance measurement process in federal government programs. Since I am tracking them and their effect on the program, Framework 2 helps implement machine learning solutions that aim to reduce or eliminate these administrative burdens. But how do I know whether the implementation of these machine learning solutions reduces or eliminates administrative burdens in a meaningful way? This is the importance of Framework 3. Framework 3 outlines an evaluation schema based on the theories, practices, and requirements of evidence-based policymaking to providing researchers and administrators with clear guidelines to measure the impact of the machine learning solutions on administrative burdens, but also to measure the impact of the reduction of administrative burdens on the overall outcomes of the policy or program.

This is an important component because I need a clear way to determine if our machine learning solutions make a meaningful difference in the manner and direction that I aspire to. As detailed in the evidence-based policymaking research, many government programs suffer from inadequate evaluation regimes (Stack, 2018). This prohibits us from understanding whether the program is meeting its intended goals and to what extent. It also disallows us to fully understand how program changes impact the outcomes of those goals. This limits the effectiveness of policies but also cuts us off from understanding what programs are the most effective and efficient to meet the goals of taxpayers, legislators, and administrators. Without these

evaluations of programs and their outcomes, we are relegated to values-based and anecdotal understanding and discussions of policies.

Framework 3 will also draw lessons from the public and private sectors. In the private sector, companies often deploy measurement processes when testing new machine learning solutions. However, these are mostly insufficient for our purposes because private sector outcomes tend to be easily measured outputs or outcomes such as more clicks, more purchases, and less customer churn. Instead, I need to include knowledge and theories from the evidence-based policy movement, which applies scientific testing and evaluation to government programs that allow for similar measurement of outcomes and outputs but also help us draw inferences to the more complex and complicated policy goals typically found in the public sector.

Measurement of impact and evaluations are important for us to understand the full range of impacts from our program changes. As discussed, the Evidence Act was signed into law, and it has created many new requirements for federal government agencies to comply with and best practices to follow. The OMB and agencies are diligently distilling the Evidence Act requirements into memorandum and operating procedures. Framework 3 acknowledges these requirements, as well as other program evaluation requirements in the law, policy, and best practice research and oversight. Framework 3 will not attempt to create a new program evaluation standard from existing research and novel methods. Instead, I believe it is more useful to take the existing and emerging requirements which administrators must comply with and adapt them to the use cases of machine learning solutions and administrative burden measurement and change. This means that Framework 3 will be more of a theoretical case study but will provide a clear path towards leveraging requirements to inform and improve the design and implementation of machine learning in government and solutions to reduce or eliminate administrative burdens.

Similar to grounding Framework 1 in the performance measurement movement, by grounding Framework 3 in the evidence-based policymaking movement, I leverage requirements and infrastructure which already exist in the federal government and research communities. This lowers the costs of including these evaluations, both in terms of knowledge and infrastructure needs for administrators. The agency or subagency unit is already conducting audits and evaluations, which are now required through several OMB and GAO mandates (Commission on Evidence-Based Policymaking, 2017; Stack, 2018; U.S. Government Accountability Office, 2017). I believe that extending them to measure the impacts on the program, from implementing machine learning solutions to reducing administrative burdens, is not much more difficult. If an

organization has not already created this type of evaluation framework, then Framework 3 will offer a simple-to-use process to create one, and administrators will have the benefit that these efforts will easily transfer to other areas of their program and be easily understood throughout the federal government because of the current focus on these methods and theories.

In the same vein and building on the groundwork being laid by the Evidence Act requirements in the federal government, frameworks 2 and 3 focus on applying evidence-based policy theory and processes to the study of administrative burdens and machine learning solutions in the government sector. For the same reasons that I leverage Framework 1 to instill administrative burden measurement in performance management, I hope to realize the same benefits by adopting evidence-based policy theories and practice in the study of administrative burdens and machine learning solutions. As I discussed, federal agencies are being pushed through practice and regulation to adopt evidence-based practices and evaluations. Therefore I build on that structure of knowledge and requirements to focus the review on the impact of our machine learning solution on reducing administrative burdens.

Tying the Frameworks Together

The need for all three of these frameworks in my research is simple; without frameworks 1 and 2, I cannot effectively implement Framework 3. Framework 1 could stand on its own. This might be a great starting place for researchers and administrators who wish to understand if administrative burdens are present and the magnitude of those burdens. This might be sufficient for their program and purpose at some time. This could be because they look and do not find any administrative burdens. Or they determine that their program has higher levels of administrative burdens, but the levels are adequate or in line with the policy goals of the program. Or perhaps more likely, they do not have the resources or leadership focus to do anything about those burdens at the moment.

However, when there is a presence of administrative burdens in a program and a desire or motivation to reduce or eliminate them, Framework 2 guides potential machine learning solutions. I admit that there is a danger here in thinking that artificial intelligence is always, or even the best, potential solution. Rather, I want the reader to understand this is just one possible solution that I explore in this research. More broadly, I hope others can see how Framework 2 could be broadened in two important ways. First, it could be adapted to look at the implementation of machine learning or AI solutions for program issues other than administrative burdens. This would obviously necessitate focusing on ensuring these other program goals are also identified and measured. Secondly, Framework 2 could be reworked or replaced to guide

the implementation of alternative solutions to reduce administrative burdens such as behavioral public administration, decision science, shifting burdens to an agency from the beneficiaries, and a multitude of other potential solutions. In this way, it is understandable that administrators might focus on frameworks 1 and 2 but perhaps not implement Framework 3 right away.

By implementing all three frameworks, I gain the best understanding of our efforts to achieve our program and policy goals. This is the basis for the importance of the evidence-based policy movement and has long been the goal of evaluation science. Unfortunately, in both the public and private sectors, I still see too many instances of solutions and programs adopted and implemented without a framework to measure whether these are achieving the intended outcomes; or, if so, to what extent. This is a missed opportunity for many reasons because I do have a desire to understand how effective programs are; whether tax dollars are being spent in meaningful ways; how people can learn from one program to build upon our knowledge in subsequent programs; and to learn what doesn't work – or even has negative or opposite impacts from what is desired. Therefore, while you could only use Framework 1, or a combination of frameworks 1 and 2, I think it is clear why administrators should aspire to implement all three frameworks as soon as possible.

Methods

I have covered the importance of this research and how it is going to be accomplished by focusing on the three distinct frameworks. I have also covered how the three frameworks tie together and ultimately build upon each other to help accomplish the overarching goal of not only allowing us to identify and measure administrative burdens and to design and implement a machine learning solution for them but also to evaluate the impact of that solution on the overall program and policy goals and outcomes. In this section, I will explore how I am going to achieve these frameworks, the methods I will use to research and design them as well as how I propose to validate my design and research.

Research Questions

Research Question 1 - Framework 1: Identify and measure Administrative Burdens within the performance management process?

The administrative burden research has continually pointed to the importance of understanding the existence of and the impacts of administrative burdens on policy outcomes. This research has also espoused the benefits of the reduction of administrative burdens. Among

these are increased take-up and participation in programs by eligible beneficiaries and less uneven program participation because administrative burdens effects potential participants unevenly due to the lack of resources and uneven impact of administrative burdens on some of the neediest, more vulnerable, or challenged citizens (Heinrich & Brill, 2015; Herd et al., 2013; Herd & Moynihan, 2018). However, the field lacks a robust framework to both identify and measure administrative burdens. Without this, researchers and administrators are unable to measure and quantify the presence and magnitude of administrative burdens in programs. Additionally, it makes it impossible to understand how burdens change based on program or policy changes. Even if policymakers and administrators set out to reduce or eliminate them, they need tools to measure and understand the impact of their efforts on administrative burdens, to allow us to compare burdens across and between implementations of the same or similar programs, as well as within one program itself over time.

To achieve this, we need to understand the nature of learning, compliance, and psychological costs on people in the context of federal government programs and to understand how to validly measure these consistently. Ideally, this would involve existing administrative data, which is collected either for the performance measurement or some other existing purpose but can be reused for performance information to capture administrative burdens in an existing performance measurement process. Therefore, the research questions for Framework 1 are how to measure each of the administrative burden costs, as well as how to adapt performance measurement policy and processes to include each of them. Let me take them one at a time in this research and will build upon existing administrative burden research.

Research Question 2 - Framework 2: How can we implement machine learning solutions in public sector programs to reduce administrative burdens?

To design and implement machine learning solutions to reduce administrative burdens in federal government programs, I need to answer two important research questions. First, how should I design and implement machine learning solutions to include the steps and considerations which are specific to government programs? Secondly, what changes are needed most likely to reduce administrative burdens based on applied machine learning solutions? To do this, I will lean heavily on the current ideas and theories in the administrative burden research for steps and methods to reduce them to see which are most applicable to machine learning solutions. But then, I will also examine these and other possibilities based on the performance measurement process as outlined in Framework 1.

Research Question 3 - Framework 3: How do we evaluate the impact of machine learning solutions on administrative burdens and on the outcomes of the policy?

The final research questions for my dissertation are focused on the need to understand the impact of machine learning solutions on administrative burdens and the program overall. These questions are: how to evaluate the impact of machine learning solutions on the program's administrative burdens? And how to measure the impact of the change in administrative burdens on the overall outcomes and impact of the program or policy?

The need for Framework 3 is rooted solidly in the evaluation and evidence-based policymaking research and literature. Like many government programs, there is a tendency to adopt programs and policies and implement them as designed without a framework to follow up to determine if they are achieving the desired results for our country. Without a built-in mechanism to evaluate their impact and their effectiveness, governments and researchers do not have a basis for judging the resources being spent to implement them. I believe there is a similar trend going on in the artificial intelligence sphere in both the private and public sectors. This is the problem that the evaluation science and evidence-based policymaking fields are attempting to solve in many areas, which I believe can also apply to machine learning.

Innovative solutions to public administration problems are needed in many areas, and I believe that AI can help us in a multitude of ways, but AI is a tool and not a program by itself. Additionally, it is not always an effective tool, and it can create additional problems for government programs. Therefore, as part of any machine learning solution for administrative burden reduction, there needs to be a built-in framework to evaluate the impact of the solution to measure the effect on both administrative burdens and the overall program. By doing this, we are poised to accurately evaluate the changes to outputs and outcomes of our program caused by the machine learning solutions. Additionally, the more this is done, the better off researchers and administrators are when looking to adopt solutions in their programs and jurisdictions because they will have a growing repository of other program results and stories to lean on.

Research Methods

Providing answers to the above questions will necessitate the creation of three distinct frameworks. These will be created by leveraging research and theory from multiple disciplines. These will include public administration, evaluation science, data science, and behavioral economics, among others. I will accomplish this through content and conceptual analysis of academic literature study, federal government document analysis, and information mining. By

combining these different techniques, I will triangulate on facts, findings, research questions, background information, and best practices. I will be relying on several qualitative methods in this research because the data does not yet exist to research these issues quantitatively.

However, the existence of these frameworks will allow for scientific, qualitative analysis of these phenomena in the future.

As discussed previously, there are relatively few public examples of the implementation of machine learning in programs where citizens are seeking benefits. At this time, none of these instances have data collected or available in a manner that will allow us to look at the impact of machine learning on administrative burdens in the program. Additionally, while the government is producing and making available more and more performance data, there is not yet an agreed-upon method or framework for measuring administrative burdens. When completed, this research will allow practitioners and researchers to study administrative burdens and machine learning solutions in the government in a formalized and quantified manner.

To help make up for any potential shortcomings of these qualitative analysis techniques, I will employ a triangulation method to combine insights about my topic areas and research questions from different sources to synthesize findings and attempt to fill gaps. Triangulation is common in public administration research and has been found to increase validity by providing strength over potential weaknesses of using any one method alone (Ricucci, 2010). In this research, I will triangulate using literature study, document analysis, and information mining for each of the three frameworks.

Triangulation

Triangulation is the application of multiple research methods (often three or more) in a study to study the same hypothesis or phenomena. This method is popular in qualitative research but has increased with mixed-methods research focus as well. The benefits of triangulation methods are that using different methods can increase the credibility and validity of the research results (Bekhet & Zauszniewski, 2012; Carter et al., 2014; Ricucci, 2010). Triangulation gains these increases in validity through the combination of empirical evidence, theories, observations, and research perspectives to increase the sample size of data which helps overcome biases and weaknesses that single studies can create or produce (Bekhet & Zauszniewski, 2012). For each of my frameworks, I will deploy the triangulation method and use a literature study, document analysis, and information mining to build my theoretical and empirical understanding to create the three frameworks.

Framework 1 Methods

Framework 1 will build the foundation for the research and the following two frameworks. Framework 1 is focused on creating methods and a tool to identify and measure administrative burdens in policy programs with performance management techniques. While there has been a significant focus on the existence and importance of administrative burdens, there is not yet a holistic approach to measure them, especially not as part of a larger process to understand their impact on an overall policy area. There has been some focus on the effect of administrative burdens on program take-up, especially of the proportion of potentially eligible beneficiaries compared to actual beneficiaries, with the idea that administrative burdens can cause otherwise eligible individuals to not apply for or be approved to receive the program benefits (Herd et al., 2013). However, these studies have mainly focused on this aspect and the correlated impacts but not as they relate to the system as a whole.

My approach will be to leverage existing research and practice of performance management, which is a process that leverages strategic goals and performance information and measurement of outputs, inputs, and outcomes to better understand the processes and resources of a program throughout. In the current iterations, performance management is focused on providing administrators, researchers, and policymakers timely information about resource allocations, processes, and outputs as a means to benchmark and analyze policy systems to identify waste, resource mismanagement, efficiencies, and ultimately improve how programs are implemented and administered (Dooren et al., 2015).

Framework 1 will build on the performance management research, theory, and practice in the federal government to provide clear paths for the identification, measurement, and inclusion of administrative burdens of Learning Costs, Compliance Costs, and Psychological Costs as part of a program's performance management infrastructure and model. I will look at how this could be applied to programs, as well as extend the performance management research and frameworks to harmonize with administrative burden research. In some instances, administrative burdens will already be built into some performance measurement steps and performance information data collections but are not specifically labeled or disaggregated. Without this level of granularity in performance information, it does not allow me to directly understand their impact, nor does it allow for transparent and informed discussions of the policy implications of administrative burdens.

Content and Conceptual Analysis

I will review the literature and federal government documents on both performance management and administrative burden academic research and tools to combine these fields in my framework. As stated previously, I believe that the focus of administrative burden research and theory is indirectly built into the scientific management and performance management field, so this aspect of the framework will simply highlight the natural connections and explicitly interweave these related theories and schools of thought. The content analysis will also help apply the methods of performance management directly to the problem of identifying and measuring administrative burden costs. As discussed earlier in this work, the existing implementation of performance management in the federal government is not ideal, but it has improved over prior iterations. Additionally, the focus of this work is not to improve performance management writ large but instead to improve the measurement of administrative burdens as a means to understand them and reduce them. This will have the intended benefits of not only providing tested mechanisms for federal government performance measurement but also helping researchers and administrators include administrative burdens in existing methods and models, rather than requiring a separate process or separate model specifically focused on burdens. In most federal agencies, there already exist performance management processes, both because they are required by law and policy but also because some of the strongest proponents and the most valuable implementation of performance management happen through subagency processes and leadership. Therefore, it is likely that extension of the existing literature and processes to include administrative burdens will allow for increased ease, lower resource expenditure, and increased take-up of administrative burden inclusions.

For my administrative burden and performance management research, I will perform searches for current literature through the use of three research literature repositories and search engines: researchrabit.ai, EBSCO academic research database, and the Social Science Review Network (SSRN) academic research aggregator. I will use search terms such as “administrative burdens” and include modifiers such as: “public administration,” “federal government,” “public programs,” “government,” “public policy,” and “government.” An important note is that I will nearly exclusively use “administrative burden” research detailed the United States government field and exclude research from or about Europe and Canada. This is because they use “administrative burdens” with a different definition, often synonymous with what has been previously defined as red tape, which are frictions within government-to-government

interactions, rather than the definition of administrative burdens used in this research which are frictions imposed on public interactions with government.

Using the same search aggregators and repositories, I will then perform similar searches for research on performance management. I will use search terms such as: “performance management,” “performance measurement,” and performance information” with modifiers such as: “public administration,” “federal government,” “public programs,” “government,” and “public policy,” and “government.” I will find research on the use of performance management in the federal government and other parts of the public sector to inform my framework.

Based on the corpus of the resulting literature, I will use content and conceptual analysis to identify specific examples of administrative burdens in government programs and how they can be measured in a manner that would fit the requirements of performance management processes. I will exclude literature and documents which are not specific to administrative burden study in the United States federal government and those which do not contain examples of burdens in a program or how to measure them for Framework 1. While there are some resources that provide analysis of how to better design and implement for more successful performance management use - I will still exclude any resources or insights which are not applicable to the current requirements and policy of federal government performance management. These will be excluded based on whether they are implemented or can be implemented in the existing requirements and constraints of performance management in the federal government. For example, changes that would suggest changes to performance management that would require legislative or regulatory changes will be excluded from Framework 1 results, even if they have been shown to improve the overall accuracy and use of performance management based on research. I do not want to suggest performance management processes that cannot be implemented in the current state - I wish to use and extend existing requirements and practices to include administrative burdens measurement. This will be a known limitation of my research and provides an area for future research as it is likely that the current iteration of performance management is not optimal for administrative burden identification and measurement.

Similarly, there will be research on administrative burdens and theory, which will not be useful for understanding how to measure them. For the results of Framework 1, I will exclude any administrative burden research that does not contain useful insight about how to identify and measure administrative burdens. These will be excluded based on my content and conceptual analysis as I read through the work and either identify and code specific examples of

administrative burden measurement or code the work as not containing any. Based on my approach, I will include all administrative burden research which includes case studies, since these will provide data about how to identify and measure them. Therefore, while I will exclude some research, likely the number of exclusions will be low. These resources will also be out of scope for my literature review for Framework 1. Unfortunately, this is a notable gap in the existing administrative burden field - and therefore, resources are currently limited. While this is one aim of this work, I will likely need to be inclusive of administrative burden work as case studies and anecdotes of different mechanisms to identify and measure them.

I will use the results to synthesize the existing performance management requirements for research-informed features and best practices of performance management in the federal government to include tools to create performance measurement processes for each of the three administrative burden costs as they are elucidated in the research based on coding the examples from current research literature and federal government documents. This will allow me to begin to create specific methods to identify and measure each of the three costs of administrative burdens through federal government performance information. Since I am not looking at a specific program or policy, Framework 1 will be a generalized toolkit for administrators and researchers, suggesting multiple methods and techniques. As applied to a specific policy or program in the future, this will allow the selection of targeted tools to fit the requirements and aspects at that time.

Federal Document Analysis

To augment the above content analysis for the academic literature, I will also conduct content and conceptual analysis of federal government documents detailing existing best practices, frameworks, design guides, and other resources for performance management processes and systems in the federal government. This will include a search of the federal register, OMB guidance, GAO reports and materials, performance.gov, PIC.gov, and agency-specific websites to find the policy and/or procedural documents for designing and implementing “performance management,” “performance measurement,” associated GPRAMA and Evidence Act requirements such as “learning agendas,” APG planning to include the creation and monitoring of “milestones” and “indicators” processes in federal agencies and programs. To these, I will add specific carve-outs for identifying and applying these requirements to the three costs of administrative burdens in the performance measurement systems to allow for clear identification, labeling, and measurement.

As discussed, some of the existing systems and theories likely already inadvertently capture this information, but not in a sufficiently granular way to identify and label administrative burdens. In other instances, it is likely that existing documentation and systems do not adequately extend themselves to capture the totality of performance information which is necessary for accurate identification and measurement of one or more costs associated with administrative burdens. This is because the needed performance information may not be easily available to the agency or review body, or it was not determined to directly impact the performance measurement system because it was seen as an “external factor.” This is likely the case with psychological cost attributes as they may be more directly focused on the externalities of potential program beneficiaries rather than data that is readily available within the program and processes. I believe that in these cases, there may be minor changes and additional data collection, which will make it possible to easily incorporate all administrative burden costs in performance management models.

The federal government document analysis will help inform the tool created through the research literature content and conceptual analysis by further grounding it in existing federal government and agency performance management requirements and processes. This is important for my work because I am not attempting to alter the overall performance management theory or to recommend ways in which it should be modified in application in the federal government. Instead, I’m attempting to harness the resources and benefits of federal government performance management to include the measurement of administrative burden costs in policies and programs. Therefore, most of the resulting Framework 1 will be extensions of the best practices of performance management in the federal government as it currently exists, informed by the most current research and theory on performance management and administrative burdens.

Information Mining

Lastly, I will perform data mining of publicly available performance management data such as the federal register, data.gov, and performance.gov to identify possible data sources - or required modifications for data sources - to measure administrative burdens for the framework. This will include web searches to gather as wide a corpus of documents as possible. Once my corpus is collected, I will use data parsing and/or natural language processing (NLP) techniques to search for performance management and performance information examples that can serve as case studies or provide real-world data for Framework 1 to be applied. This will allow me to compare Framework 1 requirements with performance information and performance measurements already in existence to see what exists already which can be repurposed versus

where existing systems need to be extended or modified to allow for the implementation of Framework 1. In some instances, the administrative burden costs might be best compiled by existing data points. In other instances, programs may capture some but not all of the required data to accurately identify and label administrative burden costs. By providing real-world examples and linking them to Framework 1, I hope to provide concrete and actionable paths for researchers and administrators to implement Framework 1 to better focus their attention on the importance of administrative burdens in their programs and their existing performance management processes. This will also allow my research to make specific policy recommendations to the field, which will allow the use of these frameworks by researchers and public administrators.

Framework 2 Methods

Framework 2 is meant to do two things. First, it will focus on applying machine learning to federal government benefits programs to reduce or eliminate administrative burdens of learning, compliance, and psychological costs. As discussed in the prior sections, there are many potential solutions for administrative burden reduction in government programs, but this research will focus on the potential solution of machine learning. Secondly, this framework will more generally provide a guide path to implementing artificial intelligence applications in the public sector. The reason for both purposes is that we need to achieve this more general goal to achieve the first more specific goal. The world is beginning to see the promulgation of artificial intelligence frameworks, best practices, rules of ethics, and even federal government guidance. But few products also focus on applied use cases and instead mostly provide general statements and guidance. This framework will highlight the steps and challenges for a machine learning solution as experienced in nearly all implementations. Therefore, it will borrow heavily from the current academic research and private sector experience. However, because the public sector is unique and brings with it specific considerations, requirements, and challenges, it will highlight each of these and provide methods to account for each.

However, Framework 2 will be a bit more theoretical than Framework 1 since it will rely on machine learning solutions to administrative burdens as measured by Framework 1. Therefore, these burdens and solutions will be grounded in the performance measurement and performance information collections of Framework 1. Since Framework 1 will be a result of the research and not yet applied as a whole in a specific program, Framework 2 solutions will be possible machine learning techniques to reduce or eliminate administrative burdens and how these will be informed and cause changes in the performance information as measured by Framework 1. Even though Framework 2 will have to be more theoretical than Framework 1, I expect to be able to rely on the emerging growth of government agency best practices, playbooks, and frameworks as they begin to explore more use cases of AI and machine learning governance. However, since there are no centralized government requirements at the current time, I will not defer to one specific path but rather focus the framework broadly to cover the most important, but also the most complete, group on considerations and principles.

Content and Conceptual Analysis

I will extend the analysis completed for Framework 1 to include machine learning solutions to reduce the costs of the administrative burden identified and measured. This

extension of the content and conceptual analysis from Framework 1 will establish where ML solutions can potentially be used to reduce administrative burdens for Framework 2. Framework 2 content analysis will also work in concert with the document analysis to combine the identification of design and implementation concerns and solutions for machine learning in the public sections. I will explore the research within the fields of computer science, data science, artificial intelligence, and machine learning specifically focused on the design and development of ML solutions. I will exclude the myriad of academic literature regarding the specifics of different types of ML models, such as ML research where new models, algorithms, and applications are tested and reviewed, instead of focusing on the overall practice of applied ML, typically known as MLOps. This will allow me to incorporate the latest methods from this rapidly growing field. I will also incorporate the small but growing research and literature about the public sector's use of machine learning. The field of public administration has a burgeoning research and exploration of these techniques, which will form a foundation for this work. There is also significant research and exploratory work within the legal scholar community that reviews and provide potential solutions to many of the legal, ethical, and practical challenges of public sector AI. Political science also contributes work focused on the interactions, considerations, and potential solutions to unique challenges of using AI in the United States democracy, especially as it promises to interact with our societal norms and political values in challenging ways. Additionally, there is perhaps the largest body of applied research in the computer science and data science field, especially in terms of design, evaluation metrics, debiasing research, and mechanisms focused on machine learning algorithms and applications. This field must also be incorporated to ensure the public sector is taking the most recent advances in machine learning into account for public use cases.

I will perform searches on Google Scholar, EBSCO, and the Social Science Review Network (SSRN). I will use search terms such as “machine learning,” “artificial intelligence,” and machine learning” with modifiers such as: “federal government,” “government,” “framework,” “ethical considerations,” “legal implications,” “public administration,” “political science,” “bias mitigation,” “administrative procedures act,” “paperwork reduction act,” “privacy,” “transparency,” “public sector,” and others. Gathering a wide collection of research about the methods and considerations of designing and implementing machine learning in the public sector will form the basis for Framework 2 to ensure it includes input from the many sources now focused on this type of research. These methods and considerations will be developed in the framework but also applied specifically to mitigate and reduce the three costs of

administrative burdens as identified and measured in Framework 1. This means that the framework will both generally capture the varied considerations of machine learning in public programs but also be a narrow application, specific to administrative burden costs. Based on the results of the literature study, I will synthesize information to lay out a framework for designing and implementing machine learning solutions for the reduction of the three costs of administrative burdens in the government. This will be a practical framework that will look much like a project management tool.

Federal Document Analysis

As in Framework 1, I will perform a search of the federal register, OMB guidance, ai.gov, GAO reports and materials, performance.gov, and agency websites to identify existing frameworks, best practices, guiding documents, or other material about developing and implementing artificial intelligence programs in the federal government. To this, I will include internal and external reviews and research on these programs. For example, media outlets have begun reporting in-depth on some of these applications with the intent to provide transparency for the public consideration of the impact of these applications. I will also explore the private sector, which is becoming rife with best practices, frameworks, rules of practices, ethics manifestos, and consulting manuals. To do this, I will perform Google searches, as well as look specifically at trade publications and private sector websites, and identify private contractors who are soliciting federal government clients for this type of work. Additionally, I will search and review private sector machine learning frameworks, best practices, design methodology, project management techniques, codes or standards for ethics, and evaluation criteria. Like many areas, the private sector has a head-start on the applied design and implementation of machine learning solutions. The private sector has constraints and considerations that are much different than those in the public sector. However, there is still a great deal of knowledge and experience to be reviewed and which can and should contribute to this framework.

The document analysis results will help me build and refine the framework tool drafted from the literature study. It may be that eventually, OMB or even congress will mandate specific machine learning design and implementation steps and techniques in the federal government, but until then, I believe Framework 2 needs to incorporate the best insights and practices from research and applied practice in the private and public sectors. Once these are identified and incorporated, the only additional step will be to have practical steps to be followed for each of the three administrative burden costs as they may necessarily require different machine learning approaches or considerations.

Information Mining

I will perform data mining of publicly available data contained within the federal register to identify federal government uses of machine learning for consideration and comparison to Framework 2 and to see if there are any government use cases of machine learning which may be analogous or applicable to my administrative burden use case. These data include regulatory announcements and changes, Systems of Record Notifications, Privacy Impact Assessments, and evaluation reports. I believe this will allow me to capture information about existing and emerging use cases. While I believe the number of uses in the federal government is small at this time, there is growing interest and resources being applied in this area, so I want to ensure I compile as much of the contemporary universe of examples as possible to inform my research. Of keen interest will be any machine learning use cases that also contain performance information or evaluation data, which could be compared against my frameworks.

I will mine the information through API pulls to form a large corpus of potential information. Once formed, I will employ descriptive statistical analysis and NLP techniques to ascertain any possible examples of machine learning uses that may have effects on administrative burdens. Once I identify potential examples, I will analyze these manually to see if they fit my research and, if so, how I can lean on them to help inform my design of Framework 2.

Framework 3 Methods

Framework 3 is going to allow researchers and administrators to evaluate and provide evidence about the effectiveness of the machine learning solutions in reducing administrative burdens on the overall program or policy outcomes. This is an area of intense interest in evidence-based policy research and practice. Just as performance management gave rise to awareness and focus on being able to link resources with processes and outputs, evidence-based policy research has focused on measuring the impact and effectiveness of those inputs and outputs on the goals and desired outcomes of the program. The goal of the evidence-based policy movement is to adopt and implement programs that have been proven effective in ways that allow administrators, researchers, and policymakers to continue to measure the effectiveness of the outcomes. Framework 3 will accomplish this by starting with the current and emerging requirements for evaluation from the Evidence Act and adapting them to evaluate machine learning solutions and program outputs and outcomes based on solutions to reduce or eliminate administrative burdens.

Academic Literature Study

In my literature study for Framework 3, I will focus on the research and theory of evidence-based policymaking as well as evaluation science to ensure this framework is leveraging this copious work to identify models, processes, and techniques required by the federal government to measure the effect of machine learning used to reduce administrative burdens on program outcomes (as outlined in frameworks 1 and 2). I expect to focus on the public sector field as there are significant research programs to lean on. But I will also draw from research and practice literature from the private sector since there is a growing focus on the measurement of AI techniques as they impact desired outcomes for private sector uses. I believe there is value in broadly combining these resources to provide as comprehensive and nimble of a framework and trial applications for Framework 3. Partly this is because there are likely to be significant differences between future AI applications in the public sector, so one rigid framework requirement will likely not be applicable from one program to the next. However, this foundational research will be used to adapt federal government agency requirements under the Evidence Act and other legal and regulatory evaluation requirements to the specific use cases or machine learning solutions and those specific to reducing or eliminating administrative burdens. This will greatly narrow my outputs for Framework 3 but also provide a practical

framework for my use cases while helping government administrators and researchers understand how to apply Evidence Act requirements in useful ways to evaluate these use cases.

I will perform searches on Google Scholar, EBSCO, and the Social Science Review Network (SSRN). I will use search terms such as “evidence-based policy” and “evaluation” with modifiers such as: “public sector,” “government,” “machine learning,” “formative assessment,” “summative assessment,” and “artificial intelligence,” and others. I will use the outputs of the literature study to synthesize the research, requirements, and best practices from evidence-based policymaking and evaluation science to create a tool that is applicable to Framework 2 machine learning solutions. This will ensure that government administrators and researchers can use the required processes to evaluate the effectiveness of the Framework 2 solutions to determine their impact on administrative burdens and the program outcomes overall. It is important to note that I will not attempt to make modifications to evidence-based policymaking theory or practice as applied in the federal government. Rather I want to build upon these requirements and best practice research for my specific area of application to easily show how they can be used to provide a way to measure the effectiveness of Framework 2 solutions.

Federal Document Analysis

The document analysis for Framework 3 will similarly look to the federal register, OMB, GAO, agency websites, as well as evidence.gov, data.gov, performance.gov, and related repositories. I will also focus on primary source documents such as the Evidence Act and implementation guidance from OMB, GAO, and other agencies in the federal government. I’ll also include documents from agency and sub-agency organizations focused on the design and implementation of evidence-based policy and evaluations within government programs. This analysis will also extend to nonprofits, think tanks, and private consultancies, which are also key stakeholders in the federal government's evidence-based policymaking and evaluation community. These documents will help me build and refine Framework 3 by leveraging the most pertinent and practiced design and implementation tips for evaluations of federal government policies and programs. These documents will also ensure that Framework 3 will comply with requirements such as the Evidence Act and OMB policy memorandum while benefiting from the experience and expertise of the GAO and federal agencies to ensure Framework 3 is implementable and useful for the evaluation of machine learning techniques to reduce administrative burdens, and to measure the impact on the overall policy program outcomes.

Similar to the literature study for Framework 3, this document analysis will further refine and inform how to use existing evidence-based policymaking and evaluation requirements, tools, and practice to measure the impact of Framework 2 machine learning solutions on administrative burdens as a measure of performance management in Framework 1 as well as to evaluate the impact to the overall policy or program outcomes. The results of the document analysis will help ground the Framework 3 tool in the most contemporary and successful practices and requirements to ensure that it is easily incorporated by federal government agencies and not only validly evaluates as intended but adds to existing practices rather than creating new, potentially conflicting requirements. These findings will allow me to leverage and show how existing evaluation planning and materials can be used, with modifications, to evaluate the impact of machine learning solutions on administrative burdens and on the program outcomes writ large.

Chapter 4 - Framework 1 Results - Identification and Measurement of Administrative Burdens with Performance Management

Overview and Goals

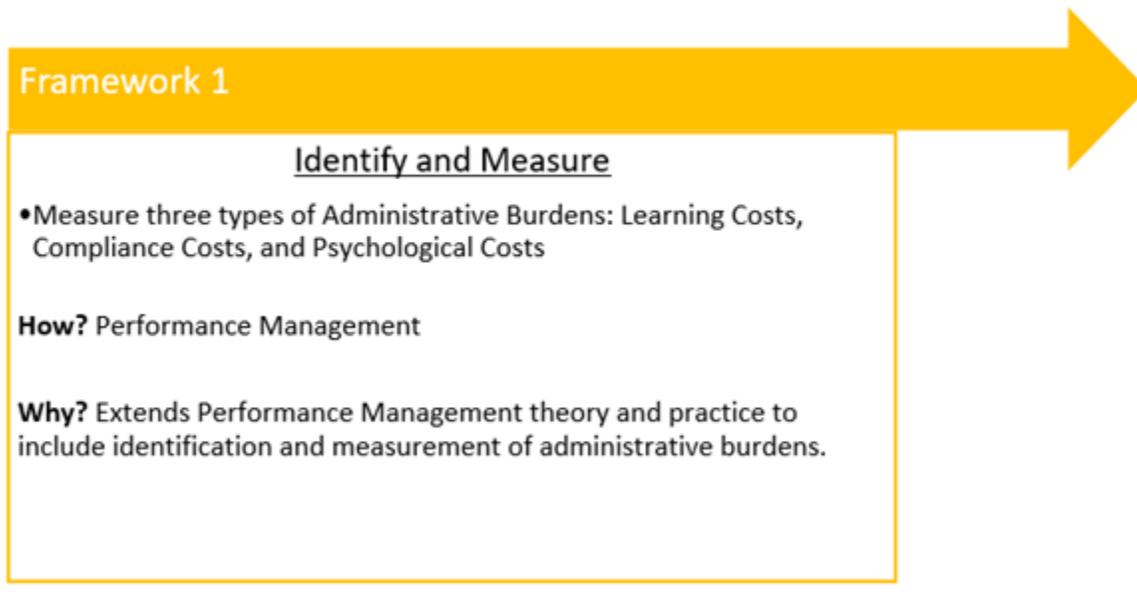


Figure 3-2: Framework 1 (author)

It's clear that administrative burdens pose impactful frictions to individuals seeking government services and benefits but that the government does not yet have a clear method to identify and measure them. It is also clear that performance management is an important requirement in the federal government, and many agencies and sub-agency units have successfully implemented and used performance management to improve their processes and strategic decision-making. As discussed, this research will not seek to modify or change performance management theory or the requirements of how it is implemented in the United States federal government. Instead, this research will rely on the performance management movement as it currently exists in the United States federal government and use it to ensure administrative burden costs can be identified and measured within performance management processes. This is a noticeable gap in existing academic research and government administration fields. I believe that the field should provide tools for researchers and administrators to identify and measure administrative burdens. This framework will allow for the existing resources, requirements, and skillsets being used for performance management in the United States federal

government to also be focused on measuring and setting goals to reduce administrative burdens within programs and policies.

Framework 1 will help ensure the continued focus on and further research of the impacts of administrative burdens in a measurable and quantifiable way. This is important because the current literature on administrative burdens has focused on their importance and provided some qualitative and specific use case quantitative measurement ideas, but the field lacks a rigorous measurement framework. Instead of providing a measurement framework that is wholly new to existing federal government methods and requires significant costs and barriers to be reckoned with to implement, I have developed a mechanism that can be implemented by government administrators and researchers under existing conditions. Ideally, grounding this in the performance management process will facilitate its use since it will be a modification of existing performance management theory and systems rather than a stand-alone administrative burden-only measurement system. I believe that it is important to create ways to identify and measure administrative burdens and build a clear path for administrators to use this information in their program design and implementation to help manage the program in a way that accounts for administrative burdens.

In this section, I will tie together the implementation guidance for performance management in the federal government with the academic research on administrative burdens, especially focusing on how to measure learning costs, compliance costs, and psychological costs in government programs. These resources will help me develop a framework to measure administrative burden costs using performance management requirements in the federal government. This framework will be supported by my literature review and my document analysis. I will also create a hypothetical performance management goal with milestones and indicators that identify performance measurement criteria based on these methods to show how to fully implement these techniques. Once the framework is built, I will conduct information mining to identify any instances of current performance management goals or measurement of administrative burdens or similar initiatives or implementation in the federal government to assess and refine my framework.

Framework Methods: Performance Measurement for Administrative Burden Costs

Framework 1 is about setting strategic goals to reduce administrative burdens, creating performance data around those goals, and then making decisions based on the data to help meet

those goals. Framework 2 will go into specific actions using machine learning to reduce administrative burden costs. However, beyond simply measuring the administrative burdens in the program, strategic goal setting needs to be specific, time-bound, measurable, and in-line with how the agency or sub-agency work unit is organized to measure their work (*PIC Resources. / PIC.Gov*, n.d.). Therefore, our performance management framework must set goals for the reduction of administrative burdens in a specific period rather than just identifying and measuring. This requires my framework to identify how to set strategic goals for the reduction of administrative burdens within the existing performance management requirements, identify and measure administrative burdens within a complimentary performance measurement program based on these goals, and identify the oversight and compliance process to ensure government leadership is engaged in the purposeful use of this performance data to make decisions aligned with meeting their strategic goals. Finally, I will perform information mining to identify existing government performance data that can be used to apply to this framework focused on the reduction of administrative burdens.

For Framework 1, the first challenge is to set strategic goals to identify and reduce administrative burdens and then to set milestones and indicators to measure progress towards that goal. I will adapt the existing performance management requirements for the federal government to show how to set strategic goals to reduce administrative burdens in programs, measure progress against these goals, and use this performance information for administrative leadership to make decisions that move agencies closer to those goals. I will also use the existing administrative burden literature to identify how to measure them within the context of a performance measurement process in the federal government.

Administrative Burden and Performance Management Literature Review and Document Analysis Methods

To identify ways to measure administrative burdens, I performed a review of the existing administrative burden research literature to identify examples of administrative burdens. As discussed previously, the administrative burden research does not yet have a methodological way to identify and measure the experienced learning, compliance, and psychological costs, which is what I am trying to solve with Framework 1. Therefore, I have created this framework to help close this gap. While there is no specific method to identify and measure administrative burdens, Herd and Moynihan (2018) proposed a series of diagnostic questions to help identify the presence of administrative burdens. They admit these are not exhaustive, but they are inductive

questions rooted in their established theory of administrative burdens but also informed by the case examples from their research (Herd & Moynihan, 2018). During my literature review, I built upon these diagnostic questions to identify specific diagnostic questions for each example of administrative burdens costs from the current academic literature on administrative burdens and identify examples of administrative burdens contained in the existing literature. This means that Herd and Moynihan's diagnostic questions were useful for some academic research examples but not others. In these other cases, I had to create my own diagnostic questions based on the examples in the research. I coded these diagnostic questions based on where they fit under three types of administrative burden costs: learning, compliance, and psychological. This categorization was based on the example of a burden as compared to the definitions of the three types of costs. In some research, these diagnostic questions were overtly mentioned, but in most of the current research, the researchers are using case studies and examples to discuss the theory, impacts, and causes of administrative burdens without specifically focusing on how to diagnose and measure these burdens, so the diagnostic questions had to be inductively applied.

Once I identified the examples and diagnostic questions from the current literature, I incorporated the current research literature and document analysis of performance management requirements and processes in the federal government to identify the primary potential cause of the administrative burden cost and the specific way to measure that cost through performance measurement theory and techniques as currently applied in the federal government. There is an important insight for administrative burden measurement and an important caveat for my data coding: there are likely to be multiple potential causes for the experienced costs since these are established as experienced by individuals, which are also disproportionately felt and impacted based on individuals' factors (Christensen et al., 2020; Masood & Nisar, 2020). Therefore, it is unlikely that I will identify single causes for them that are universal to all settings. However, I do believe it is possible and useful to identify the most likely primary causes.

From these potential causes and measurements of each example from the literature, I then identified a "refined measurement" which further grounds the measurement for each specific example into something which can easily be implemented in existing performance measurement programs in the federal government and is something which can be measured based on administrative program data or from purposefully designed information collections such as surveys of individuals and organizations involved in the program. These refined measurements are meant to be general and actionable. While they derive from the specific examples in the existing literature on administrative burdens, they are also readily applicable to government

programs and policies which are not specifically mentioned or researched in the current literature. They are also actionable as they clearly identify a potential cause of administrative burden costs and operations which can be measured by the government agency.

Once these refined measurements were created, I further coded each one of these into a more generalized “Simplified Measurements” based on general measurement categories which emerged from the data. These general simplified measurement categories are informed both by the research and by the measurement ideas proposed by Herd and Moynihan (2018) and Sunstein (2020) when suggesting ways to reduce administrative burdens. However, furthering the coding of the examples in the data set also allows me to explore the performance measurement categories as they relate to different administrative burden cost examples and approaches to show how different measurement approaches and types are related to the different types of administrative burden costs.

Building on these codified simplified measurements, which derive from the diagnostic questions, I will create a hypothetical Agency Priority Goal (APG). The hypothetical APG will be constructed to clearly articulate strategic goals set for administrative burdens, as well as milestones and indicators to create a performance measurement plan to monitor and guide leadership in pursuit of that goal within the APG time period of 24 months. In doing so, I will show a clear path for framework 1 to result in applied performance management for administrative burdens in the federal government. It is my belief that this can be applied now under current conditions without any legislative or policy changes required.

Results: Agency Priority Goal Framework

Based on the existing performance measurement requirements for the federal government, I am going to focus on a framework for an Agency Priority Goal (APG) to reduce administrative burdens in a program. The framework for an APG includes developing performance measurement criteria for the three types of costs associated with administrative burdens, then including specific milestones and review criteria. I also build this on the existing requirements for APG development and monitoring as outlined by OMB Circular A-11, Part 6. OMB recommends the strategic goal of an APG to be two sentences, where the first is an impact statement and the second an achievement statement (Office of Management and Budget, 2021). So, for Framework 1, I begin with a generalized strategic goal statement that can be modified and applied to individual policies or programs:

Improve the outcomes of [the program] by reducing administrative burdens. By [24 months from now], administrative burdens will be reduced by 20 percent overall, with at least a 5 percent reduction falling into each component of learning cost, compliance costs, and physiological costs.

Government administrators and researchers can plug in the specific instance of any federal government program, as well as adapt the specific outcomes as needed, but I use this general goal to walk through and build the framework to lay a path towards applying this to specific programs. The overall strategic goal is to reduce administrative burdens in a program by a specific amount or percentage in a two-year timeframe. This overall strategic goal will then be broken into sub-goal components and milestones based on approaches to these goals. This can be achieved by focusing on the specific legal requirements of a program, understanding the multiple ways to achieve those requirements, and then measuring or forecasting the specific learning costs, compliance costs, or psychological costs associated with each of those particular methods and then choosing the method which has lower costs but still achieves the same goals. There may be instances where there is a less costly method but at the expense of losing something which is either required or politically desirable, such as the significant decrease in the ability to keep fraudulent claims below a particular threshold. For example, particular requirements for certifications, verifications, or fraud-resistant measures may be legislated rather than open to changes at the administration or regulation level, so these will have to be accounted for, limiting what changes can be made to the program.

It is important to remember that agency administrators setting APGs do not have full control over program requirements or direction, and so the APG must still account for these requirements. As the adage goes: “this is a feature, not a bug,” because while it may seem constraining or unnecessarily political to keep these constraints in place, which induce additional administrative burdens or limit the changes that can be made to reduce burdens, the ability for elected officials to set the overarching policy goals or criteria is a key feature of representative democracy and therefore is a key foundation to performance management in the federal government (Dooren et al., 2015; Moynihan et al., 2012). However, making this an explicit decision that can also be measured increases transparency and allows it to be discussed and decided in a politically relevant way. Otherwise, as pointed out by Herd and Moynihan (2018), these administrative burden decisions are policy-making that do not happen transparently. Since they are impactful to the experiences of individuals and they have a significant impact on

program outcomes, it seems that they should be covered by administrative law mechanisms, allowing for public participation, understanding, debate, and direct feedback by citizens (Herd & Moynihan, 2018).

As required by the APG, there must be indicators and milestones designed and collected throughout the 24 months period with quarterly reviews of progress by the APG leadership group. When developing these performance measurement plans, administrators need to consider the types of administrative burdens that are experienced as it pertains to the program. Additionally, they need to consider what is measurable and reportable. Since administrative burdens are lived experiences, they are understood, experienced, and influential on individuals. Therefore, administrators will not always be able to measure the experiences of individuals directly, nor should they in certain cases of performance measurement. Instead, I focus measurements on the factors which lead to those experiences. Knowing that these causal factors will lead to differential and subjective experiences, I can ensure that my design and focused factors will most likely lead to positive experiences for the majority of individuals (Herd, 2015; Herd & Moynihan, 2018). Despite the fact that someone experienced administrative burdens to a greater extent than another person based on unique aspects of their situation, the majority of individuals should experience lower administrative burdens based on the program changes.

It is important to note that my example APG is hypothetical but realistic and conforms to the current federal government performance management criteria and guidance. While it calls for an overall reduction of administrative burdens by “20%”, this could as easily have been more specific measurement criteria for a program that sets a very specific APG, such as increased take-up rate to a proportion of eligible individuals or even a specific number of eligible individuals with that number being informed by political policy goals or a threshold of the estimated eligible population. Many APGs and performance management goals include percentage targets, however, as they allow for the aggregation of different components for the milestones and indicators. Some performance management goals, on the other hand, only set directionality and not magnitude. For example, they may call for “increased” or “decreased” administrative burdens in a program while relegating the magnitude of those directional changes to the milestone and indicators. Neither is inherently correct or incorrect, but the current guidance asks agencies to include both direction and magnitude in the APG if possible (Performance Improvement Council, 2019).

Now that I have set the APG with directionality and magnitude goals, I turn to how to create a performance measurement plan based on the traits of the program leading to

administrative burdens. Notably, the work on “sludge” focuses on measurements through existing benefit-cost analyses and paperwork reduction act procedures, but even these were not directly linked to administrative burden costs (Madsen et al., 2020; Sunstein, 2020)). Therefore, when creating my dataset, I reviewed the administrative burden academic literature and pulled out specific measurement processes for each type of administrative burden cost contained in the studies (typically based on a case study or focusing on specific causes of costs). I then analyzed these examples and categorized them into more specific buckets of measurement processes for each cost and then de-duplicated the dataset until I had unique measurement processes for each of the three costs.

I began my review of the literature by searching EBSCO, SSRN, and researchrabit.ai academic literature aggregators and repositories for search term “administrative burdens” and included the following modifiers: “public administration,” “federal government,” “public programs,” “government,” “public policy,” and “government” which resulted in 120 papers and documents. From these I excluded any which were not focused on the United States since Europe and Canada’s use of “administrative burdens” are the same as our definition of “Red Tape” which is out of scope for my research since it is government on government friction. I also excluded any research which did not contain actual examples of administrative burdens in programs or work which didn’t discuss the categorization of burdens into the three costs or the measurement of them, which resulted in 34 sources. From these 34 sources of administrative burden research that contained specific administrative burden examples, I coded 96 specific measurements of different administrative burden costs with the following breakdowns of cost measurements for all the sources as shown in Table 4-1.

<i>Type of Cost</i>	<i>Count of Cost Type</i>
Compliance Costs	51
Learning Costs	19
Psychological Costs	26
Total	96

Table 4-1: Counts of Measurement Example per Type of Cost (author)

Even though the diagnostic questions and measurement solutions were explanatory, I further coded these measurements into a “refined measurement” for each example that allowed me to summarize the measurement in a general way that was informed by, but no longer contained, the specific case example it came from. This resulted in the following cost types and measurement approaches for each of them, which is the baseline for the performance

measurement plan framework that I was able to deduplicate, which reduced my overall number of measurements from 96 to 80.

<i>Type of Cost</i>	Unique Potential Cause and Measurement
Compliance Costs	49
Learning Costs	19
Psychological Costs	23
Total	80

Table 4-2: Count of Unique Cause and Measurement by Cost types

Then I further simplified the types of refined measurements into eight “simplified measurements,” as shown in Table 4-3 below, in order to identify categories of administrative burden measurements that are generalized from the examples in the current literature. These are application measurement, assistance measurement, feedback measurement, outreach measurement, process measurement, program measurement, and tool measurement. These simplified measurements are indicative of the focus of a performance measurement indicator for each type. However, the specifics of the performance measurements still need to be tailored to the program and the processes where the information and data would need to be collected for the performance measurement system. And there are a number of simplified measurements based on the existing literature data set.

Simplified Measurement	Explanation
Application Measurement	Measures of applications, ways to apply, methods of application, the burden of application (government vs. individuals), validation and certification processes, where in-person steps are required.
Assistance Measurement	Measures of the amount and variety of methods of assistance to applicants
Feedback Measurement	Measures of direct experiences based on feedback in the form of direct observations, surveys, interviews, focus groups, etc.
Form Measurement	Measures of information collections and forms such as numbers, time to complete, and amount of data elements.
Outreach Measurement	Measurement of the types, methods, resources, and presence of outreach to individuals.
Process Measurement	Measures of variety, types, responsible entity, error rates, number of instances of processes including adjudication decisions, appeals, re-enrolment, benefit redemption, etc.

Program Measurement	Measures of the overall program such as proportions of covered compared to eligible individuals (take-up rate), measurement of attrition, etc.
Tool measurement	Measurement of the availability, variety, and usefulness of tools designed to facilitate program exploration, application, explanation, or application processes.

Table 4-3: Simplified Administrative Burden Measurements

Beyond their importance to this work, I believe that these simplified measurement categories will become the foundation of future research to help categorize measurement approaches for administrative burden costs within the performance measurement process. They will also allow us to look at and understand the interactions between these measurement approaches, the causes of administrative burdens, and attempts to reduce administrative burdens because it gives researchers and administrators a shared lexicon and set of tools to begin within program and across program measurements and analyses. I also believe that these simplified performance measurement categories for administrative burdens will allow the field to compile tools and mechanisms that can be easily adapted by government agencies to measure administrative burdens in their programs.

Exploring the breakdown of these simplified measurement approaches in the existing research helps provide an understanding of where the administrative burden field currently is in regard to identifying burdens based on how they can be measured. As shown in Table 4-4, in the existing research, feedback measurements are the most popular mechanisms to measure administrative burdens cost, likely because administrative burdens are experienced phenomena and therefore, indirect measurements are going to be important, including asking for feedback from individuals about their experiences. Beyond that, application measurement, process measurement, and program measurement are seen to be popular tools aligned with the current literature which is a positive signal since these are well established in existing performance management and performance measurement systems.

<i>Simplified Measurement</i>	Count
Application Measurement	22
Assistance Measurement	7
Feedback Measurement	26
Form Measurement	6
Outreach Measurement	2
Process Measurement	17

Program Measurement	13
Tool measurement	3
Total	96

Table 4-4: Counts of Simplified Measurement

As can be seen from the analysis of the existing literature in Figure 4-1: Count per Cost Type below, the number of measurements of compliance costs is significantly greater than the measurements of learning costs and psychological costs. This may be because the outputs and inputs associated with compliance costs are more applicable to current measurement processes in the government, as evidenced by the PRA and APA requirements. Examples of learning costs are less present in existing measurement processes which may be why the current academic literature might not focus on it as explicitly in the case of examples as they need to account for a significant number of individuals who may never begin a formal interaction with the program application process. Therefore unique measurement devices and processes need to be considered. Similarly, much work has looked at the associated stigma or stress of certain program participants based on the program requirements, but many of these studies are not specific to the administrative burden research and theory and therefore are not as directly applicable to this study. However, they do illustrate the nature of the problems and correlations between psychological costs and outcomes of programs and participation. My work creates a more easily followed path between those understandings and administrative burden measurements for future identification and measurement.

Count of Cost Types

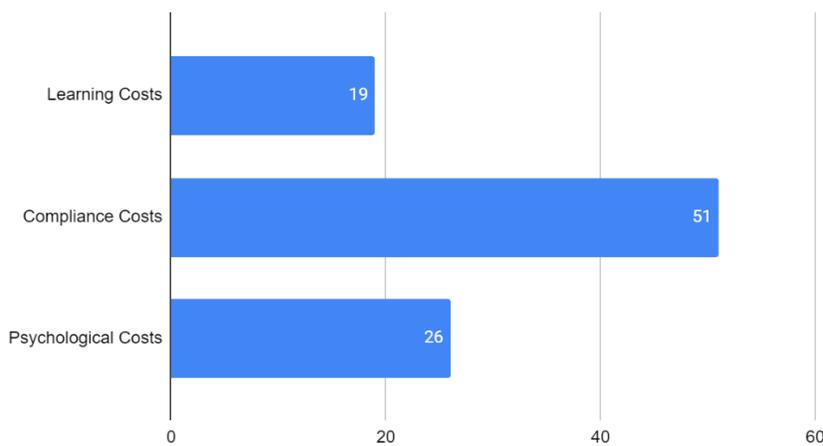


Figure 4-1: Count per Cost Type

Type of Cost	Count of Unique Refined Measurement
Compliance Costs	48

Learning Costs	19
Psychological Costs	19
Grand Total	74

Table 4-5: Refined Measurement Counts by Cost Type

From this data set, I then specified more generalized and refined measurement approaches and narrowed the list of measurement approaches to 74 refined measurements, as shown in Table 4-5. Therefore, the current academic research for measurement indicates it may be simpler to measure compliance costs or at least that these measurement processes are currently more robust. Instead, there are noticeable buckets of measurement approaches that can be tailored for specific programs. This is understandable since programs are a complex mixture of legislated requirements, regulated processes, and policies, which are further constrained by resource and technological realities. More important than a specific list of performance measures that can be applied to every program for administrative burdens, this approach identifies many potential approaches which result in measurements for some more generalized and standardized administrative burdens causal factors which can be measured to find an overall measurement of amounts of learning costs, compliance costs, and physiological costs. I believe that by focusing on these already identified administrative burden performance measures and continuing to grow and enhance the focus on administrative burden measurement, I can begin to create best practices and standards for different types of programs and different types of administrative burdens costs.

What this shows us is that even in the current implementations of performance measurement in the federal government, which is quite robust throughout programs and agencies, there are not any “one size fits all” approaches to measurement. This is because performance management relies on the specific strategic goals, combined with indicators and milestones specific to these goals and to the programs being measured. Therefore, administrators and researchers must always use an informed program theory of change and the specific of a policy and program to design our measurement plan (Gao, 2015; Hatry, 2002; *P3 Playbook / PIC.Gov*, n.d.). For example, if a program uses call data to measure compliance costs experienced by individuals and administrators, create a specific performance measurement plan for compliance costs associated with calls (e.g., answer rates, time spent on hold, the proportion of problems resolved on the first call, etc.) then they cannot apply this same performance measurement plan to a program which does not offer to call as a method for interactions with individuals.

When I examine the simplified measurements disaggregated by types of administrative costs in Table 4-6, I see patterns whereby compliance costs are more likely to be measured through application or process measurements and psychological costs are most likely to be measured through feedback measurements. There is no clear pattern of measurement types for learning costs in the existing literature, perhaps because there are wide varieties of ways to learn about programs, as well as a variety of program parts for learning costs (such as benefit eligibility, benefit types, application processes, redemption processes, etc.).

<i>Type of Cost</i>	<i>Simplified Measurement</i>	Count
Compliance Costs	Application Measurement	17
	Assistance Measurement	6
	Feedback Measurement	2
	Form Measurement	6
	Process Measurement	13
	Program Measurement	7
Compliance Costs Total		51
Learning Costs	Application Measurement	3
	Assistance Measurement	1
	Feedback Measurement	5
	Outreach Measurement	2
	Process Measurement	2
	Program Measurement	3
	Tool measurement	3
Learning Costs Total		19
Psychological Costs	Application Measurement	2
	Feedback Measurement	19
	Process Measurement	2
	Program Measurement	3
Psychological Costs Total		26
Grand Total		96

Table 4-6: Count of Simplified Measurements by Cost Type

In the following sections, I will focus on a more detailed understanding of the potential measurement approaches for each of the types of administrative burdens cost as applied to an APG. Typically, administrators are going to focus on these as specific milestones and indicators which will roll up into the larger APG measurements. As previously stated, I am not going to

determine a specific list of measurement approaches because each program is unique and has its own constraints and specific design and implementation features. However, these lists of approaches will provide a basis for program administrators to choose and apply specific measures for each learning cost that are specific to their programs, as well as provide achievable measurements within performance management goals and milestones. These will be the components of the administrative burden costs. Once rolled up to this level, I will begin to have comparable measurements between program implementations over time for the same program or between programs. For example, I will be able to compare learning costs between two programs even if those two programs are administered very differently, and the component of learning costs are not alike. Additionally, for programs administered similarly, I will be able to compare learning costs but also compare learning costs as it relates to application measurements and process measurements. This will undoubtedly open up an additional world of comparative research on administrative burdens.

Learning Costs Measurement

The APG strategic goal has laid out what the hypothetical administrator intends to accomplish within the specified time of 24 months, and now I seek to identify the specific milestones which will allow the administrators to reach that goal. The milestones and indicators are specific to the three types of administrative burden costs. Just as I looked at the overall reduction of administrative burdens, I will identify targets for reducing the three types of costs beginning with learning costs. In order to do this, administrators and researchers must understand both the intricacies of the particular program, as well as the make-up of learning costs that are experienced by individuals wishing to know and understand how to qualify, apply for, and remain eligible for our program. Learning costs are also made up of the costs associated with understanding the program benefits and rules associated with it (Herd & Moynihan, 2018; Moynihan et al., 2015). In the below table are the specific examples that I have identified learning costs and associated measurement ideas and techniques from the literature. As discussed, these are not exhaustive of the causes or learning costs or how they can be measured, nor are they applicable to every program. Rather, this list is a beginning place for researchers and administrators seeking to build their performance management milestones and indicators to create baseline measurements of the learning costs in their program and to provide performance measurement plans as they take actions to lower their learning costs. The full data set is available in the appendix.

As specified in our strategic APG, I want to lower each type of administrative burden costs by five percent, with an overall goal of reducing administrative burdens measured in our program by 20 percent within the specified two-year period. Therefore, for our milestones, I identify the learning costs indicators which are applicable to our program, as well as the areas in which we believe our program changes (discussed in the next chapter) can have the most impact on reducing administrative burdens. I do not want to choose milestones and indicators for things that are not within our control to change or that are not applicable to our program. In-person wait times for a program that does not offer in-person education resources wouldn't make much sense for a learning costs indicator. Nor would choosing a milestone measurement about shifting learning costs burdens from the individuals to the government agency if it was somehow prohibited by legislation for the government to proactively advertise or educate individuals about the program. Not because this wouldn't be an important program area to focus on overall, but because something which is legislatively required or prohibited is outside of the control of agency administrators and, therefore, shouldn't be the focus of a performance management plan (although this doesn't mean it should not become an agency legislative priority for engagement with Congress) (Choi & Moynihan, 2019; Dooren et al., 2015).

Table 4-7 below contains possible performance measurement techniques for learning costs as identified through a review of the current administrative burden literature. Interestingly there are no clear patterns in the identification and measurement of learning costs as in the other types of costs discussed in Table 4-6. I believe this is because of the diversity of what is covered in learning costs, such as learning about the program, learning about how to apply, learning about eligibility criteria, and learning about benefits and their redemption. Additionally, there are several ways to view these learning processes, such as learning by the government organization reaching out, needing to seek information directly from the government, and learning from interactions with a third-party organization or group.

Diagnostic Question	Potential Cause and Measurement	Refined Measurement	Simplified Measurement	Source
How does the required task or process differ based on the executive functioning of individuals?	<i>First-time applicant vs. repeat applicants (more or less administrative experience); health and executive functioning; scarcity will decrease human capital</i>	<i>Program application process data disaggregated by participant executive function variables</i>	<i>Application Measurement</i>	<i>Christensen et al., 2020</i>
Are program take-up rates different based	<i>Potential indicators of biases or administrative</i>	<i>Application completion rates, disaggregated by</i>	<i>Application Measurement</i>	<i>Moreno and Mullins, 2017</i>

on protected categories?	<i>exclusion based on protected attributes (Sex, race, gender, age, etc.)</i>	<i>protected category factors</i>		
Measures of burden based on PRA estimates	<i>Higher estimates of PRA burdens are associated with more difficult requirements to learn and understand, as well as more difficult program eligibility requirements</i>	<i>PRA Measures</i>	<i>Application Measurement</i>	<i>Sunstein, 2018</i>
Increased access of groups or organizations to navigate the application process, which benefits individuals	<i>Reliance on groups and organization to comply with program benefits are associated with decreased compliance costs</i>	<i>The proportion of third-party entities performing compliance processes</i>	<i>Assistance Measurement</i>	<i>Shybalkina, 2020</i>
Does distrust of government or organization make it more difficult to learn about the program?	<i>Distrust of the State can lead to more information gathering about the program because of distrust of government-sourced information.</i>	<i>Customer feedback about a government entity</i>	<i>Feedback Measurement</i>	<i>Ali & Altaf, 2021</i>
Do individuals consider the government a reliable source of information?	<i>Unreliable state services require greater diligence and alternative sources of information</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Ali & Altaf, 2021</i>
Do individuals have to locate other sources of information?	<i>Prevalence of government information-seeking vs. outside information seeking</i>	<i>Customer feedback about government vs. non-government sources of information</i>	<i>Feedback Measurement</i>	<i>Ali & Altaf, 2021</i>
What are the causal mechanisms of burdens within a specific program?	<i>"Causal Process tracing" of a program allows a within-program case study association of causes of burdens.</i>	<i>Analysis of Survey responses</i>	<i>Feedback Measurement</i>	<i>Camillo, 2020</i>
Is Non-take up due to lack of knowledge?	<i>(1) Non-knowledge. Do welfare clients and ex-clients know that the program exists? Do potential claimants understand the program's eligibility rules? What learning costs do they face? Do service providers actively propose and "sell" the program to potential claimants?</i>	<i>Survey of program participants and eligible non-participants</i>	<i>Feedback Measurement</i>	<i>Daigneault and Mace, 2020</i>
How do participants find out about the program?	<i>Amount of outreach about the program, as</i>	<i>Outreach Measurement</i>	<i>Outreach Measurement</i>	<i>Herd & Moynihan, 2018</i>

	<i>measured by efforts, budget, and reach.</i>			
Shifted burdens to state vs. individual?	<i>Does the state take responsibility for finding potentially eligible individuals?</i>	<i>Measures of program learning based on responsibility (government versus individuals)</i>	<i>Outreach Measurement</i>	<i>Herd et al., 2013</i>
Are there more or fewer rules and requirements about how program benefits are redeemed?	<i>Redemption of benefits with particular rules and requirements can increase compliance costs</i>	<i>Amount of rules regarding program use</i>	<i>Process Measurement</i>	<i>Barnes, 2020</i>
Benefits Bundling?	<i>Does one application review eligibility for multiple benefits?</i>	<i>Number of program eligibility determinations per application</i>	<i>Process Measurement</i>	<i>Herd et al., 2013</i>
What is the take-up rate based on learning costs?	<i>Learning costs can also impact the take-up rate</i>	<i>Program take-up rate</i>	<i>Program Measurement</i>	<i>Bhargava and Manoli, 2015</i>
Early Nudging about program information/eligibility?	<i>Providing behavioral nudges early in program learning can reduce learning costs</i>	<i>Measure of program up-take based on nudges</i>	<i>Program Measurement</i>	<i>Linos et al., 2020</i>
Does the program account easily for non-traditional applicant situations?	<i>Individuals who do not neatly fit into program rules or processes increase administrative burdens</i>	<i>Program participant edge cases, take-up rate</i>	<i>Program Measurement</i>	<i>Nisar, 2018</i>
How do individuals establish eligibility?	<i>Amount of information and tools that help individuals understand eligibility requirements.</i>	<i>Availability and use of eligibility tools</i>	<i>Tool measurement</i>	<i>Herd & Moynihan, 2018</i>
How do individuals understand the benefits?	<i>Information and tools to provide information about benefits, including resources for specific and individual situational understanding.</i>	<i>Availability and use of program and benefit tools</i>	<i>Tool measurement</i>	<i>Herd & Moynihan, 2018</i>
How do individuals learn about the application process?	<i>Resources and services focused on explaining and assisting with the application process.</i>	<i>Availability and use of application tools and services</i>	<i>Tool measurement</i>	<i>Herd & Moynihan, 2018</i>

Table 4-7: Learning Cost Measurements from Current Research

I believe that the implementation of learning costs measurement and focus will help increase awareness of learning costs and their causes, as well as how to measure them. More than compliance costs and psychological costs, I believe the examples and types of learning costs are underrepresented in the existing academic literature because the idea is somewhat newer and less studied.

Compliance Cost Measurement

Similar to learning costs, Table 4-8 contains possible performance measurement criteria for compliance costs as identified in the existing administrative burden literature. Program administrators should identify areas that are specific and measurable to their program that will allow them to build a baseline measurement of current compliance costs associated with their program as well as to monitor the progress of those compliance costs after the changes to their program are made which will be detailed in the following chapter.

Diagnostic Question	Potential Cause and Measurement	Refined Measurement	Simplified Measurement	Source
How do individuals submit the forms (e.g., in-person, by mail, online)?	<i>Scores for less onerous processes.</i>	<i>Variety of application processes and measurement of ease of use</i>	<i>Application Measurement</i>	<i>Herd & Moynihan, 2018</i>
Is information collected from individuals available through government administrative data?	<i>A review of existing information collection requirements against what is available through government data sharing.</i>	<i>The amount of information collected from applicants was available through administrative data sources</i>	<i>Application Measurement</i>	<i>Herd & Moynihan, 2018</i>
How frequent is re-enrolment?	<i>The multiplier effect of costs is based on the frequency of re-enrolment or verification requirements.</i>	<i>How often do individuals need to re-apply, and what proportion of initial requirements must be resubmitted</i>	<i>Application Measurement</i>	<i>Herd & Moynihan, 2018</i>
How much does an individual's inquiry for the application process cost them?	<i>Distributive costs are based on requirements, allowable formats, and missed opportunity costs associated with missed employment time.</i>	<i>Amount of time spent on the application process</i>	<i>Application Measurement</i>	<i>Herd & Moynihan, 2018</i>
Enrollment Ease	<i>Factors that make enrollment easier decrease administrative burdens</i>	<i>Measure of application "accelerators" - which are processes that speed up or make burdensome the application process for individuals</i>	<i>Application Measurement</i>	<i>Fox et al. 2020</i>
How easy is it to renew program eligibility?	<i>Enrollment ease is associated with administrative burdens</i>	<i>Measures of the application process for renewals/extensions of the program</i>	<i>Application Measurement</i>	<i>Fox et al. 2020</i>
Time Series of compliance requirements	<i>Increased or decreased compliance costs caused changed impact levels</i>	<i>Measurements of application processes over time</i>	<i>Application Measurement</i>	<i>Heinrich and Brill, 2015</i>

can identify increased or decreased levels of compliance costs				
Shifted burdens to state vs. individual	<i>Does the state take responsibility for determining eligibility?</i>	<i>Measures of application requirements based on responsibility (government versus individuals)</i>	<i>Application Measurement</i>	<i>Herd et al., 2013</i>
Rate of program application start as compared to those that complete program application	<i>Higher rates of program application completion can be representative of lower compliance costs when controlling for other variables</i>	<i>Application completion rates</i>	<i>Application Measurement</i>	<i>Masood and Nisar, 2021</i>
How do individual agencies modify experienced burdens over time?	<i>Individuals can react and respond to burdens overtime to modify their effects</i>	<i>Compliance cost measures disaggregated by participant administrative experience (first-time applicant versus returning applicant)</i>	<i>Application Measurement</i>	<i>Peeters and Campos, 2021</i>
Time spent dealing with application and benefits procedures	<i>Polling information about time spent seeking to comply with program requirements; turning time into economic costs</i>	<i>Amount of application and compliance forms</i>	<i>Application Measurement</i>	<i>Pfeffer et al., 2020</i>
Measures of burden based on PRA estimates	<i>Higher PRA burdens are associated with higher compliance costs for program applications</i>	<i>PRA Measures</i>	<i>Application Measurement</i>	<i>Sunstein, 2018</i>
Measures of costs to individuals multiplied by all individuals compared to savings of the outcomes or processes	<i>Sludge Audits should be conducted on administrative burdens associated with impacts on individuals</i>	<i>PRA Measures and benefits-costs analysis</i>	<i>Application Measurement</i>	<i>Sunstein, 2020</i>
Application Completion Rates	<i>Ability to begin and complete the applications</i>	<i>Application Completion Rates</i>	<i>Application Measurement</i>	<i>Code for America, 2021</i>
Application times	<i>Time spent accessing the application</i>	<i>Application times</i>	<i>Application Measurement</i>	<i>Code for America, 2021</i>
Renewal Rates	<i>ability to reapply and recertify after already being eligible</i>	<i>Renewal Rates</i>	<i>Application Measurement</i>	<i>Code for America, 2021</i>
The churn rate of participants, especially	<i>Ability to remain eligible/compliant</i>	<i>The churn rate of participants, especially eligible participants.</i>	<i>Application Measurement</i>	<i>Code for America, 2021</i>

eligible participants.				
Do applicants have help in completing the application processes?	<i>Scores for resources to assist individuals, including varying levels of assistance provided based on individual needs.</i>	<i>Measurement of the level of assistance available, disaggregated by individual needs</i>	<i>Assistance Measurement</i>	<i>Herd & Moynihan, 2018</i>
How long does the individual wait to speak to someone?	<i>Wait-time for applicants (either in person at interviews or on the phone)</i>	<i>Wait times for phone and in-person conversations, email response times</i>	<i>Assistance Measurement</i>	<i>Deshpande and Li, 2019</i>
How far long do individuals need to travel to locations?	<i>Where in-person events are required or used, and how do the locations associate with individuals who are applying (or are eligible)?</i>	<i>Measurement of distance from individuals to application/program center</i>	<i>Assistance Measurement</i>	<i>Deshpande and Li, 2019</i>
Increased access of groups or organizations to navigate application process which benefits individuals	<i>Reliance on groups and organization to comply with program benefits are associated with decreased compliance costs</i>	<i>Compliance steps available and taken by for third-party on individuals' behalf</i>	<i>Assistance Measurement</i>	<i>Shybalkina, 2020</i>
Call abandonment rate, call answered rates	<i>Answer times increase compliance costs; measurements of time spent attempting to access</i>	<i>Call abandonment rate, call answered rates</i>	<i>Assistance Measurement</i>	<i>Code for America, 2021</i>
First call resolution rates	<i>Answer times increase compliance costs; measurements of time spent attempting to access</i>	<i>First call resolution rates</i>	<i>Assistance Measurement</i>	<i>Code for America, 2021</i>
What are the causal mechanisms of burdens within a specific program?	<i>"Causal Process tracing" of a program allows a within-program case study association of causes of burdens.</i>	<i>Analysis of Survey responses</i>	<i>Feedback Measurement</i>	<i>Camillo, 2020</i>
Is non-take-up due to being aware by choosing not to apply?	<i>(1) Non-knowledge. Do welfare clients and ex-clients know that the program exists? Do potential claimants understand the program's eligibility rules? What learning costs do they face? Do service providers actively propose and "sell" the program to potential claimants?</i>	<i>Survey of program participants and eligible non-participants</i>	<i>Feedback Measurement</i>	<i>Daigneault and Mace, 2020</i>
How many forms must applicants complete?	<i>All forms and supporting documentation requirements.</i>	<i>Amount of application and compliance forms</i>	<i>Form Measurement</i>	<i>Herd & Moynihan, 2018</i>
How many questions are on the forms?	<i>PRA estimates for forms.</i>	<i>Time spent on forms</i>	<i>Form Measurement</i>	<i>Herd & Moynihan, 2018</i>

Do individuals have to input data multiple times?	<i>Review of data collected more than once.</i>	<i>Duplicity in information collections</i>	<i>Form Measurement</i>	<i>Herd & Moynihan, 2018</i>
How much documentation is required?	<i>Supporting documentation and validation requirements.</i>	<i>Number of forms, time spent on forms</i>	<i>Form Measurement</i>	<i>Herd & Moynihan, 2018</i>
Do individuals require an interview, consultation, etc.? If so, is that available in person, by phone, or online?	<i>Scores for validation, consultations, or interview processes are based on the results of those processes in terms of information collection.</i>	<i>Number of forms, time spent on forms</i>	<i>Form Measurement</i>	<i>Herd & Moynihan, 2018</i>
How much time do individuals commit to the process?	<i>PRA estimates for forms but also other procedural requirements.</i>	<i>Time spent on forms</i>	<i>Form Measurement</i>	<i>Herd & Moynihan, 2018</i>
Costs of Private healthcare when government services are not trusted	<i>Preferred private options are too expensive to participate</i>	<i>Comparison of government and non-government options</i>	<i>Process Measurement</i>	<i>Ali & Altaf, 2021</i>
What is the ratio of burden or costs applied to the government agency versus the individual?	<i>Shifting the burden or costs to the agency will reduce them from the individuals.</i>	<i>Program requirements for individuals versus organization</i>	<i>Process Measurement</i>	<i>Burden et al., 2012</i>
Are eligible people applying but not receiving the benefits due to organizational error or decisions?	<i>Non-demand. Does the program answer the needs of long-term welfare clients and clients (program relevance)? What are the psychological and compliance costs experienced by potential and actual participants? Is the claiming process sufficiently simple to encourage individuals to apply?</i>	<i>Rates of false negatives in program adjudication</i>	<i>Process Measurement</i>	<i>Daigneault and Mace, 2020</i>
Digital Access	<i>Increased digital access to application/compliance increases program up-take and reduces administrative burdens</i>	<i>Are digital application and compliance options available, variety, and use</i>	<i>Process Measurement</i>	<i>Fox et al. 2020</i>
What level of administrative discretion is there in program adjudication?	<i>The scale of discretion of program adjudicators</i>	<i>Measurement of program administrator discretion</i>	<i>Process Measurement</i>	<i>Heinrich, 2018</i>
Aged individuals have lower	<i>Measuring the age and health factors of individuals in the</i>	<i>Program application process data</i>	<i>Process Measurement</i>	<i>Herd, 2015</i>

administrative capital and therefore experience higher administrative burdens with program applications and renewal	<i>program can determine levels of administrative burdens</i>	<i>disaggregated by participant age/health factor variables</i>		
Benefits Bundling	<i>Does requested information or adjudication determine eligibility for multiple benefits?</i>	<i>Number of program eligibility determinations per application</i>	<i>Process Measurement</i>	<i>Herd et al., 2013</i>
Increased red tape is correlated with increased compliance costs, especially when this is shifted to the individuals rather than the government	<i>Internal bureaucratic processes can correlate with increased compliance costs to individuals (unless all red tape costs are born by the government with no noticeable delay in adjudication or administration)</i>	<i>Measures of compliance costs compared to red tape measurements</i>	<i>Process Measurement</i>	<i>Selin, 2019</i>
Do third-party organizations absorb frictions or costs rather than passing them onto individuals?	<i>How much of the administrative burdens are experienced by individuals rather than the government or third parties?</i>	<i>Compliance measures borne by third parties versus individuals (when third parties are present)</i>	<i>Process Measurement</i>	<i>Wiley and Berry, 2018</i>
Procedural denials (missed interview, document, requirements)	<i>The amount of compliance that determines intelligibility</i>	<i>Procedural denials (missed interview, document, requirements)</i>	<i>Process Measurement</i>	<i>Code for America, 2021</i>
Accuracy Rates of Benefits	<i>Administrative accuracy and inaccuracy can increase compliance costs</i>	<i>Accuracy Rates of Benefits</i>	<i>Process Measurement</i>	<i>Code for America, 2021</i>
Appeals resulting in decision reversal	<i>False negatives of adjudication</i>	<i>Appeals resulting in decision reversal</i>	<i>Process Measurement</i>	<i>Code for America, 2021</i>
False positives and negatives for benefits decisions	<i>False negatives of adjudication</i>	<i>False positives and negatives for benefits decisions</i>	<i>Process Measurement</i>	<i>Code for America, 2021</i>
Take-up rate of a program based on compliance costs	<i>Compliance costs can impact the take-up rate of a program</i>	<i>Program take-up rate</i>	<i>Program Measurement</i>	<i>Bhargava and Manoli, 2015</i>
What is the level of	<i>Does the level of caseworker/adjudicator</i>	<i>Program take-up rate compared to the</i>	<i>Program Measurement</i>	<i>Brodkin and</i>

"administrative exclusion"?	<i>discretion result in more or less program uptake?</i>	<i>evaluation of program staff</i>		<i>Majmunda r, 2010</i>
How does the required task or process differ based on the executive functioning of individuals?	<i>First-time applicant vs. repeat applicants (more or less administrative experience); health and executive functioning; scarcity will decrease human capital</i>	<i>Program compliance data disaggregated by participant executive function variables</i>	<i>Program Measurement</i>	<i>Christensen et al., 2020</i>
Why do eligible people exit the program?	<i>Some may no longer be eligible, or they can not comply with compliance sots, or they are deemed too onerous to continue to comply</i>	<i>Measurement of program exits when individuals are still eligible</i>	<i>Program Measurement</i>	<i>Heinrich, 2016</i>
Are program take-up rates different based on protected categories?	<i>Potential indicators of biases or administrative exclusion based on protected attributes (Sex, race, gender, age, etc.)</i>	<i>Program take-up rates disaggregated by protected category factors</i>	<i>Program Measurement</i>	<i>Moreno and Mullins, 2017</i>
Does the program account easily for non-traditional applicant situations?	<i>Individuals who do not neatly fit into program rules or processes increase administrative burdens</i>	<i>Program participant edge cases, take-up rate</i>	<i>Program Measurement</i>	<i>Nisar, 2018</i>
The proportion of eligible individuals not accessing the program or not applying for the program	<i>Burdens are harmful or too onerous if they are screening out eligible individuals from the program</i>	<i>Program take-up rates</i>	<i>Program Measurement</i>	<i>Sunstein and Gosset, 2020</i>

Table 4-8: Compliance Cost Measurements from Current Research

Not surprisingly, compliance cost measurements focus on the measurement of applications, processes, and assistance within programs. These fit within the current implementations of performance management programs well, as government agencies are used to incorporating measurement of their processes based on the focus of many iterations of performance management over the years. Typically, it has been focused on resource planning, and the process and timeline focus on administrative procedure requirements such as time estimates and cost-benefit analyses. These will not be completely new performance measurement focuses for government administrators, but seeing them as a component of the larger concept of administrative burdens will help focus on the potential negative impacts and outcomes of leaving them unchanged or allowing them to increase. Therefore, I believe it will become easier when government administrators are developing their strategic plans and goals since they will have a shared basis and intended outcome for their reduction efforts.

Psychological Cost Measurement

And lastly, in Table 4-9 below are possible performance measurement criteria for psychological costs within a particular program as described in the existing administrative burden literature. Psychological costs are likely to be measured in different ways than the other costs as they are personal and internally experienced by other factors which can't always be measured directly. Therefore, the data in Table 4-9 argues for the design and implementation of feedback measurement mechanisms for performance measurement, which captures psychological costs experienced.

Diagnostic Question	Potential Cause and Measurement	Refined Measurement	Simplified Measurement	Source
Loss of Autonomy through interaction with the program	<i>Resentment and fear of the state and its representatives as a repressive, controlling, or extractive entity caused by Waiting times and spaces communicating the state's (dis)regard of its citizens</i>	<i>Measurement of application step wait times, customer feedback about experiences</i>	<i>Application Measurement</i>	<i>Ali & Altaf, 2021</i>
Measures of burden based on PRA estimates	<i>Higher PRA estimates are associated with increased feelings of privacy impositions and increased judgment of "worthiness."</i>	<i>PRA Measures</i>	<i>Application Measurement</i>	<i>Sunstein, 2018</i>
Are interactions with the application process stressful?	<i>Customer feedback survey</i>	<i>Customer feedback about the application process</i>	<i>Feedback Measurement</i>	<i>Herd & Moynihan, 2018</i>
Do people receive respectful treatment?	<i>Customer feedback and employee evaluation.</i>	<i>Customer feedback about the application process</i>	<i>Feedback Measurement</i>	<i>Herd & Moynihan, 2018</i>
Do people enjoy some autonomy in the interaction?	<i>Qualitative assessment of the application and benefit redemption process.</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Herd & Moynihan, 2018</i>
Degradation, disempowerment, and frustration at intrusive, directive, or judgmental bureaucratic encounters	<i>Negative opinions or prior experience with the state produce negative associations or feelings about the program or about worthiness for the program.</i>	<i>Customer feedback about prior government interactions</i>	<i>Feedback Measurement</i>	<i>Ali & Altaf, 2021</i>
The stigma associated with the program or having to interact with the government for assistance	<i>The stigma of associating with the state because of negative associations with government services</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Ali & Altaf, 2021</i>
Stress is associated with distrust of the	<i>The stress of greater diligence required to determine the reliability</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Ali & Altaf, 2021</i>

state and having to rely on it	<i>and safety of services (vaccines)</i>			
Does the program offer benefits based on qualifications that form a label about the recipient?	<i>Certain means-tested programs potentially identify applicants within a category or label which has stigma within society.</i>	<i>Population attitude towards program and recipients</i>	<i>Feedback Measurement</i>	<i>Baekgaard et al., 2021</i>
Does complying increase stress or stigma specific to the compliance process?	<i>Increased compliance demands are correlated with increases in psychological costs of stigma, loss of autonomy, and stress.</i>	<i>The measure of compliance costs as compared to customer feedback about the program</i>	<i>Feedback Measurement</i>	<i>Baekgaard et al., 2021</i>
Survey questions about stress	<i>Stress experienced based on program participation</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Baekgaard et al., 2021</i>
Survey questions about stigma	<i>Stress experienced based on program participation</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Baekgaard et al., 2021</i>
Survey questions about Autonomy Loss	<i>Qualitative assessment of the application and benefit redemption process.</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Baekgaard et al., 2021</i>
What is the overall balance of clients' evaluations of their experiences (positive, negative, or neutral)?	<i>Attributions of control over burdens (individual vs. government actor) can impact the associated magnitude of the burden.</i>	<i>Customer feedback about program staff</i>	<i>Feedback Measurement</i>	<i>Barnes and Henly, 2018</i>
What are the causal mechanisms of burdens within a specific program?	<i>"Causal Process tracing" of a program allows a within-program case study association of causes of burdens.</i>	<i>Analysis of Survey responses</i>	<i>Feedback Measurement</i>	<i>Camillo, 2020</i>
How does the required task or process differ based on the executive functioning of individuals?	<i>First-time applicant vs. repeat applicants (more or less administrative experience); health and executive functioning; scarcity will decrease human capital</i>	<i>Program survey feedback disaggregated by participant executive function variables</i>	<i>Feedback Measurement</i>	<i>Christensen et al., 2020</i>
Perceptions of programs and participants?	<i>The attitude of individuals participating in the program?</i>	<i>Customer feedback about program use</i>	<i>Feedback Measurement</i>	<i>Haeder et al., 2021</i>
Measuring emotions through physiological measurements (facial coding, electrodermal activity, heart rate) to determine psychological costs	<i>The physical manifestation of stress and physiological feels which can be measured</i>	<i>Direct measurements of program participant's feelings</i>	<i>Feedback Measurement</i>	<i>Hatke et al., 2020</i>
What are the levels of compliance costs in the programs?	<i>Reducing compliance costs lowers psychological costs</i>	<i>Participant feedback on the program as compared to</i>	<i>Feedback Measurement</i>	<i>Baekgaard et al., 2021</i>

	<i>individuals compliance costs</i>			
Associations between public support for administrative burdens can increase psychological costs on potential program participants	<i>High learning and compliance costs can increase psychological costs</i>	<i>Measures of learning and compliance costs compared to applicant feedback about the program</i>	<i>Feedback Measurement</i>	<i>Nicholson-Crotty et al., 2021</i>
Customer satisfaction rates	<i>rates of individuals' experiences with the program</i>	<i>Customer satisfaction rates</i>	<i>Feedback Measurement</i>	<i>Code for America, 2021</i>
Increased red tape is correlated with increased compliance costs, especially when this is shifted to the individuals rather than the government	<i>Increased compliance demands are correlated with increases in psychological costs of stigma, loss of autonomy, and stress.</i>	<i>Measures of compliance costs borne by participants versus government</i>	<i>Process Measurement</i>	<i>Selin, 2019</i>
Appeals resulting in decision reversal	<i>False negatives of adjudication</i>	<i>Appeals resulting in decision reversal</i>	<i>Process Measurement</i>	<i>Code for America, 2021</i>
What is the rate of take-up of program participation as compared to the eligible or potentially eligible population?	<i>Psychological costs can attribute to lower program take-up because of the associated frictions.</i>	<i>Program take-up rate</i>	<i>Program Measurement</i>	<i>Bhargava and Manoli, 2015</i>
Are program take-up rates different based on protected categories?	<i>A potential indicator of biases or administrative exclusion based on protected attributes (Sex, race, gender, age, etc.)</i>	<i>Program take-up rates disaggregated by protected category factors and compared to participant feedback on the program</i>	<i>Program Measurement</i>	<i>Moreno and Mullins, 2017</i>
Does the program account easily for non-traditional applicant situations?	<i>Individuals who do not neatly fit into program rules or processes increase administrative burdens</i>	<i>Program participant edge cases, take-up rate</i>	<i>Program Measurement</i>	<i>Nisar, 2018</i>

Table 4-9: Psychological Cost Measurement from Current Research

Feedback mechanisms such as surveys and comparisons of which entities (individuals or government organizations) bear the administrative burdens will not be difficult to capture and measure. However, they do need to be purposefully designed. Despite some focus on the customer experiences when interacting with the federal government, rarely are these

measurements used as a key factor for program changes or as a judgment on the program itself. Rather these are typical indicators for only the customer service component of the programs and are generally used for contract management or for the rating of individual customer service agents (*Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government*, 2021). Highlighting these measurements and feedback should be used as indicators of psychological costs to give researchers and administrators a new perspective on their impacts and that addressing them can impact the overall program and policy goals.

Detailed APG for Administrative Burdens

Now that I have identified simplified measurement processes from the current administrative burden research, I will revisit the example APG to delve into the deeper milestones and indicators to help guide administrators and researchers on the components of the administrative burden measurement and reductions. As a reminder, our examples of overall APG are:

Strategic goal: Improve the outcomes of [the program] by reducing administrative burdens. By [24 months from now], administrative burdens will be reduced by 20 percent overall, with at least a 5 percent reduction falling into each component of learning cost, compliance costs, and physiological costs.

As can be seen in the overall strategic goal, administrators want to ensure they have reductions of administrative burdens within each of the three types of costs to ensure they are evenly identifying and measuring administrative burdens throughout the costs and focusing on reductions within all types. Otherwise, it might be simple to find that reduction in one area is likely one of the largest areas to change within the control and authority of program administrators while potentially losing focus on other important areas. As an old adage goes, “what is measured is managed,” and there is a risk in measuring and focusing on one area because it often comes to the exclusion of other areas.

In addition to the overall strategic goal of administrative burden reductions, agency administrators need to set milestones to reduce individual costs by at least 5 percent each, and these milestones should have shorter timeframes than the 24 months allotted to the overall goal (Office of Management and Budget, 2021). Now that there is a better understanding of the causes of administrative burden costs, simplified measurement approaches, and how these relate to the type of costs, they can be associated with APG milestones for strategic planning. In order to set these milestones, agencies should identify the likely types of costs in their programs and

the causes in order to create measurement criteria for each cost specific to their program. This means that they can use the examples of types of causes for each cost and the measurement plans to create a component for each type of cost, just like the example I created below in Figure 4-2.

To set milestones, agencies will need to identify priority areas to focus on during the 24 months APG time period. Notably, while it would be ideal for agencies to take the time to identify and baseline all administrative burdens in their programs, I do not believe that is practicable for every agency and program. However, as agencies get more involved in identifying and measuring administrative burdens, especially in regard to program evaluations this will become more feasible, and agencies should have more information to allow them to more easily keep track of the majority of administrative burden causes and measurements within their programs. However, for the initial APG milestones, I recommend that agencies review their programs and processes to identify key components of each administrative burden cost that exists in their programs and develop baselines and milestones to reduce based on these priority components. Additionally, by identifying these priority areas, agency leadership will also identify the areas that are likely most ripe for reduction within the twenty-four-month period. These are likely to be areas that agencies are already aware of as problematic and potentially already areas of focus, either successful or unsuccessful. Since the example APG calls for an overall 20% reduction of administrative burdens in 24 months, with at least 5% in each of the three areas, this allows the agency to identify those high-impact areas to be responsible for more than 5% to make the overall 20% goal. For example, an agency that has onerous application processes and re-enrollment processes and they know that many of these can be solved through changes in the process may be an area where they set targets to reduce compliance costs by 10% instead of 5%.

In Figure 4-2, I show an example of how an agency might set milestone targets for their priority components of administrative burdens by costs. Based on these components' targets, I will then create an example milestone tracking schedule as recommended by performance management guidance for federal agencies based on quarterly check-ins for measuring and reviewing the component milestones. The logic behind this is that if these milestones are met

Reduce Learning costs by five percent within the 24-month period.

- **Component: Eligibility and benefit tools.** Increase the availability, use, and usefulness of tools to help individuals learn about eligibility and benefits.
- **Component: Application Bundling.** Eligibility and benefit information for multiple programs based on information about individuals.
- **Component: Redemption Rule Simplification.** Simplify the rules and requirements to redeem benefits.

Reduce Compliance costs by at least 10 percent with the 24-month period.

- **Component: Data Collection Reduction.** Reduce the amount of data collected, validated, and resubmitted from individuals for enrollment and re-enrollment.
- **Component: Program Bundling.** Increase the number of programs and services which are applied for, adjudicated, and enrolled in based on single processes.
- **Component: Reduce Enrollment Requirements.** Reduce the required steps, forms, touchpoints, and decisions processes associated with program application and enrollment experienced by individuals.
- **Component: Customer service and Outreach Time.** Reduce the amount of time individuals spend trying to contact, needing to contact, or interacting with agents while maintaining or increasing the application and enrollment rates.

Reduce Psychological costs by at least 5 percent during the 24-month period.

- **Component: Perceived Stress Reduction.** Reduce the amount of experienced stress during the application and enrollment processes based on survey responses.
- **Component: Program Take-up Rates.** Increase the program take-up rate proportions of eligible individuals as compared to enrolled individuals. Hold rates steady when accounting for individuals with attributes associated with lower administrative capital and cognitive capacity.
- **Component: Perceived Autonomy and Control.** Increase the levels of perceived autonomy and attributions of control experienced by individuals interacting with the program based on survey data.

Figure 4-2: Example Detailed APG

each quarter, then the agency will meet the overall APG goal, but the three-month review mark of the milestones will indicate corrective implementation and design changes that need to be made if the milestone targets are not met. This will also allow time for additional efforts if

milestone targets need to be increased because of missed milestones early which would impact the overall target (Office of Management and Budget, 2021; Performance Improvement Council, 2019).

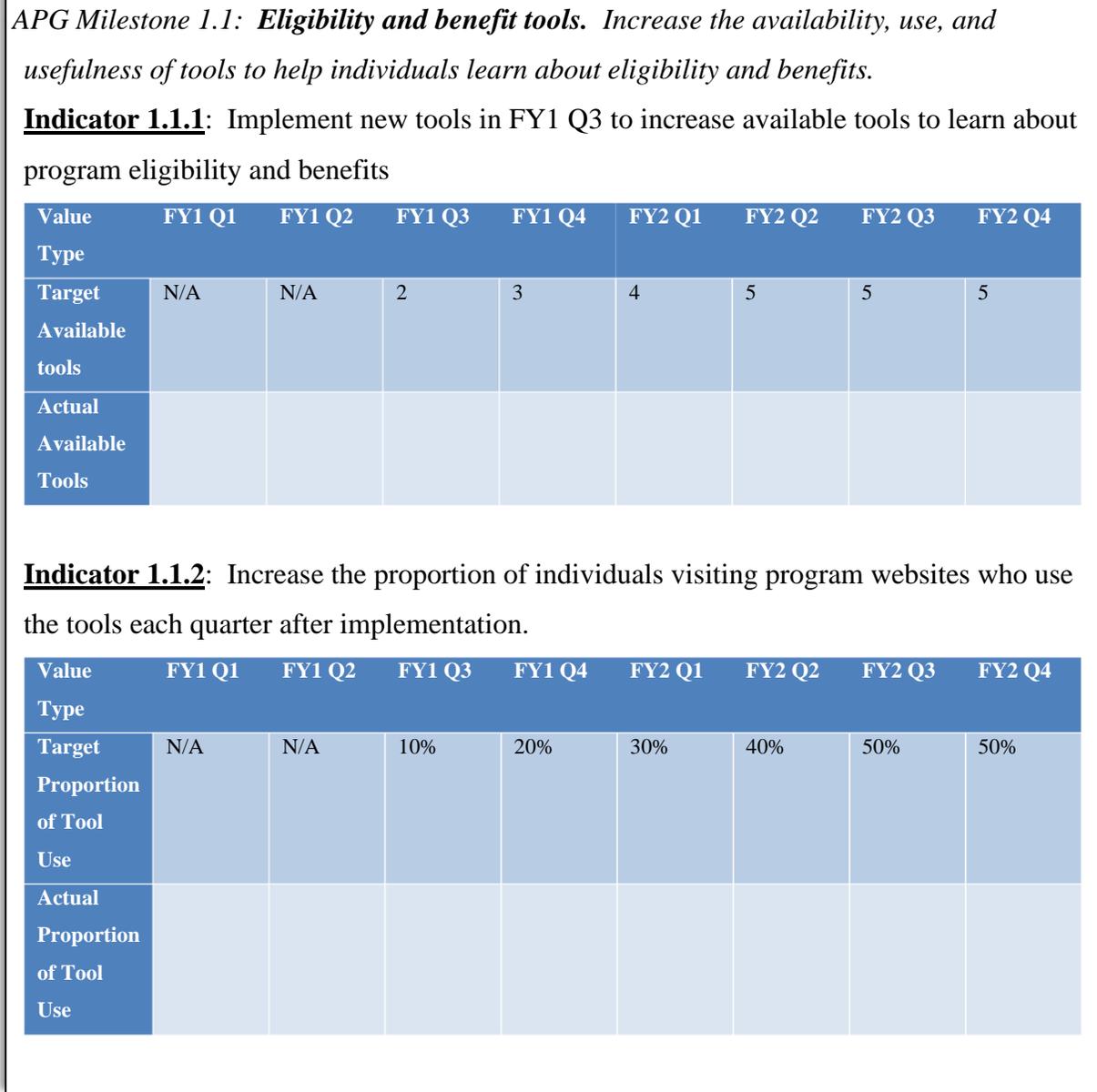


Figure 4-3: Example APG Milestone/Indicator Tracker

From these sub-component milestones, the administrator can easily create required APG indicators for each quarter within the two-year strategic goal dates which will allow agency administrators to layout their performance management plan as required, publishing the indicator targets and filling in actuals as the program progresses and APG review sessions are held with agency leadership and the performance team as seen in Figure 4-3 (Ayers et al., 2014; Office of Management and Budget, 2021; Performance Improvement Council, 2022). As can be seen with

this example milestone and indicator, I am setting strategic milestone goals based on the observation and cause of increased learning costs but not actually attempting at this stage to measure the impact on the learning costs. Rather the APG milestones and associated indicators work independently to provide specific targets for the agency leadership to achieve objective results, which can later be measured by learning agendas and evaluation techniques which I will cover in Framework 3.

Based on baselining these indicators and milestones prior to Q1 in the tracker allows agencies to understand where they are beginning with their efforts under the APG. The specific indicators targets allow the agency to aggregate the measurements into the overall milestones and even to aggregate them into the overall goal if desired. Focusing on overall proportion changes allows the agencies to combine disparate units of measurement. For example, phone call wait times (measured in minutes), numbers of website visitors (in the thousands), and numbers of website tools and users (likely in the single digits) are not immediately comparable unless you are looking for percentage change which is why that is the measurement goal in the APG. While it is important to ensure that the proportions link to meaningful goals and changes, this method allows for flexibility based on the variety of causes and measurements of administrative burdens while still providing agency leadership with simplified APG measurements which allow them to easily understand the progress towards the goals while still providing indicators for when the investigations and conversation must go into greater detail than the overall proportion changes.

Performance Measurement and Administrative Burden Information Mining

In this section, I investigate existing data and information that is publicly available and could be used to track agencies' progress towards reduced administrative burdens if part of an APG or other performance management goal or that could be used by researchers to understand the current levels of administrative burdens in programs or changes in those burdens overtime. Importantly but not surprisingly, I find very limited publicly available information that could be used for these purposes. This indicates that agencies are not focused on their current performance measurement and reporting efforts on areas that may impact the experience of administrative burdens the most. This is concerning because of the negative effects on programs and policies caused by high levels of administrative burdens.

To identify existing performance measurement data that could be used for identification and measurement of administrative burdens in programs, I performed searches of existing published information on several websites and repositories of information to make certain

information transparent from the federal government, including the specific repository of performance management information. I reviewed [performance.gov](https://www.performance.gov), which is an aggregator of federal agency performance plans and data. I also reviewed individual agency websites since they are required under GPRAMA to make this information publicly available (Moynihan & Kroll, 2016). I also reviewed and searched [data.gov](https://www.data.gov), which is a GSA-maintained catalog and aggregator of government data sets that are either available directly through the site or with instructions about how to request the data from agencies or data stewards. Of note, these are not just data from the federal agencies but also data that is made available to the public because it was collected or created by researchers, state and local government, and organizations based on government grants. [Data.gov](https://www.data.gov) was created in 2009 but was further expanded and empowered based on the OPEN Data Act requirements in the Evidence Act (*Data.Gov*, n.d.; Ryan, 2019). Finally, I also reviewed the federal register to search for APA and PRA-required information about the burdens of forms, data collections, and adjudications which are made available through federal register notices of public notices of changes as well as annual statements required to revalidate these government processes (*Paperwork Reduction Act Guide*, 2017; Sunstein, 2020).

Performance.gov

Specifically, I reviewed and searched [performance.gov](https://www.performance.gov), which is the Federal government's repository for performance management information and data (see box XX). However, beyond several blog posts regarding performance management and performance examples (many from the private sector), no repository of performance management plans and performance management data was found for the current iteration of the Biden administration. Also, many agency websites specifically link to [performance.gov](https://www.performance.gov) as the repository for their performance management plans and data. However, there is a page that aggregates and links to individual agency websites, many of which point to performance data and plans for those agencies. Prior administrations' versions of [performance.gov](https://www.performance.gov) had more available information. Under the Trump administration agency, CAP goals, APGs, and strategic goals. Additionally, performance measurement data were available (and accessible through dashboards, data downloads, and PDFs) about high-level indicators for CAP goals. While promising, only three agencies (General Services Administration, Small Business Administration, and the Nuclear Regulatory Commission) also provided detailed data about one program each. Of note, the CAP goals for most agencies were also available via dashboard and data download, but these indicators were rolled up to an aggregate level where detailed measurement data could not be tracked or inferred. Additionally, this data was only available through 2018 and 2019, so it was

not present for the final two years of the administration

(<https://trumpadministration.archives.performance.gov/data/>, accessed March 2022). The Obama administration archived the version of performance.gov contained a PDF version of all agency performance management plans, CAP goals, and administration initiatives. The site also offered an Application Programmer Interface (API) to access these plans and the subsequent reports in addition to the PDFs. Unfortunately, most attempts to access the APIs returned no data, and many of the PDFs were no longer available when accessing the site in March 2022 (<https://obamaadministration.archives.performance.gov/index.html>, accessed March 2022). Finally, the George W. Bush administration archives “Results.gov” was an overview of the PMA initiative rather than the GPRA information and data. Additionally, when accessed in March 2022, the historical scorecard data, performance plans, and results were no longer accessible (<https://georgewbush-whitehouse.archives.gov/results/index.html>, accessed March 2022).

Data.gov

Data.gov was created in 2009 by the GSA to host data sets made available to the public by the federal government. It received a significant boost from the requirements of the OPEN Data Act as part of the Evidence Act and now contains more than 300,000 federal government data sets, as well as data from State and local government, academic institutions, and public organizations (many of which share data obtained as part of grant receipt from the federal government.) Data.gov is significant because it manages data standards for federal agencies sharing their data publicly, as well as crawls agency website attempting to determine compliance with OPEN Data requirements to make data available and machine-readable. Data sets on Data.gov can be downloaded in multiple formats (e.g., .csv, excel, JSON, HTML) and are available and accessed via APIs through developer resources.

I searched data.gov based on administrative burden “simplified measurement” terms, as well as for terms such as “customer service,” “processing times,” “benefit adjudications,” and “wait times.” Notably, none of the simplified measurement terms resulted in any data which was applicable to federal government programs, confirming that these categories are not inherent in available federal government datasets. The additional search terms were identified to result in existing data on federal government performance measurement, but less than 90 results were returned for each search term. A detailed qualitative review of them resulted in the analysis that few returned any data that was applicable. The only usable results were agency “processing times” for applications for services and benefits. However, a review of these seven data sets revealed that most have aggregate time from filing to a final decision simply listed in weeks or

months rather than any granular data about the sub-processing steps and where the government agency had action versus individual applicants.

Based on this comprehensive search of data.gov, it is clear that if the performance data is currently available to apply to agency performance measurement of administrative burdens in programs, it is not being made available to the public in identifiable ways. This is not a surprising finding since most agencies do not make publicly available processing metrics or customer service data that may be collected and used for contract management or workforce management unless it is tied to a larger CAP goal or APG. Even the fairly common publication of processing times for applications and forms is not sufficiently granular to link them to specific administrative burden costs other than overall compliance costs of application processing times.

Agency Websites

I also reviewed major cabinet-level agencies, which are required to post their performance management plans, APG, CAP goals, and milestone and indicator results. What I found through this search was that most agencies only listed their current iteration of goals rather than the archived version of these documents and data. Additionally, many agencies pointed to “performance.gov” for repositories of their performance plans and data. However, as I already pointed out, much of this is not currently available and incomplete for prior administrations. Additionally, I located some data sets using the same search terms for agency websites that I used for data.gov, but not unexpectedly, I found the same data sets available on data.gov (and fewer of them indicating that agencies may not also make this data available through their website if they post it through data.gov).

Of note, I did also find evaluation and audit results on some agency websites. While I didn’t do a comprehensive review based on a qualitative review of several of these aimed at specific federal government benefits programs, there are some data that could be used to support the measurement and analysis of administrative burdens. This was especially true for evaluations that collected feedback from program applicants and participants, as often this included collecting feedback on their experiences and opinions about the application, compliance, and program use. However, these evaluations were often in PDF format with few data tables which could be repurposed by the public for a specific measurement or review of administrative burden costs. However, this does indicate that agencies possess some of the data they would need or have collected in the past to begin to compile performance measurements and indicators of their programs’ administrative burdens.

Federal Register

As discussed previously and noted in Sunstein’s work on “Sludge” and “Sludge Audits,” the federal register contains information about federal agency forms and information collections, including estimated burdens as measured by hours to complete forms which are renewed and validated by agencies when forms change or expire (Madsen et al., 2020; Sunstein, 2018, 2020). The federal register site allows advanced searches of documents through the web interface, as well as API searches through developer tools to search, download, and extract federal register information (as .csv or JSON files). It contains more than 885,000 documents as of March 2022, of which more than 25,000 contain the phrase “estimated burden,” indicating information about the time needed by an individual to complete a form or information request. As required by the PRA, these estimates are achieved by agencies by multiplying the number of respondents, the frequency of responses, and the estimated time for the response. Typically agencies detail these data in their burden statements (*Estimating Burden / A Guide to the Paperwork Reduction Act*, 2022b).

It would be possible to pull this information for specific forms and agency programs from the federal register for all records where they are available. While this is only one small component of compliance costs and associated psychological costs, this is a robust data source for point-in-time measurement and estimates, as well as a time series analysis of change over time for forms and programs. This most important limiting factor, though, is that these are estimates and therefore not associated with actual performance measurement data, so they could not be validated or disaggregated by factors that would allow me to look at how these factors apply to differently situated individuals (such as different levels of administrative and cognitive capacity, socioeconomic factors, or different iteration so policy or process implementation).

Summary

Overall, this search of existing data that can be used to measure administrative burdens as part of performance management processes makes it clear that while some data is available, the federal government is still a long way from providing the necessary and sufficient data to the public to clearly measure administrative burdens. It also has shown me that there may be additional data within government agencies that are not yet shared and not yet used in this manner. However, this is a notable highlight because it indicates more of this data may actually be created but that it simply needs to be identified and used in this manner. Hopefully, with the increased focus on administrative burdens and with the organizing definitions I’m providing in

Framework 1, the government will make much more of this data available to the public through these fora.

Information Source	Explanation	Results
<i>Performance.gov</i>	Performance management plans and measurement data	Some data available from prior administrations
<i>Data.gov</i>	Measurement data associated with programs and services	Some processing time data is available
<i>Agency/Program Websites</i>	Performance measurement, customer service, and program data	Evaluation data may be available for the agency
<i>Federal Register</i>	PRA Estimates of Time to complete forms	Estimated burden data available

Table 4-10: Summary of Opensource Results

Framework 1 Summary

This chapter has brought together the theory, practice, and requirements for performance management in the federal government as well as the theory of administrative burden to help understand how to set strategic goals to reduce administrative burdens within the context of the GPRAMA APG methods. I have also identified the most important focuses of administrative burden costs and identified specific criteria which can be used to build a performance measurement baseline understanding of the current levels of administrative burden costs within programs. In the examples for Framework 1, these performance measurement criteria became our APG milestones and indicators for us, as in the next chapter, I will show how to apply machine learning techniques to our programs to reduce administrative burdens. Having the existing performance measurement baselines and criteria will allow us to continue to measure the impacts of the ML processes and monitor our progress towards our overall APG of reducing administrative burdens by 20 percent in a two-year period.

I have also shown a path forward for agencies and researchers to begin to identify administrative burdens costs and measurement categories and criteria which will help build repositories of performance measurement data that can easily be applied to administrative burden

performance management plans and research to help us understand the existence of, and the impact of different levels of administrative burdens. I believe that using well-defined measurement categories and applying them to different programs and measurement methods for those programs will contribute significantly to our overall understanding of burdens.

As stated previously, these criteria outlined in Framework 1 build a foundation for identifying ML solutions (as well as non-ML solutions) to reduce administrative burdens in government programs that I will cover in Framework 2. Additionally, these definitions and associated measurement plans and resulting data will build a foundation for the evaluations of administrative burdens in programs and the impacts of reducing levels of burdens on the outcomes of the overall program that I will cover in Framework 3.

Chapter 5 - Framework 2 Results - Reducing or Eliminating Administrative Burdens with Machine Learning

Overview and Goals



Figure 5-1: Framework 2 (author)

In the past chapter, I created a framework to help researchers and administrators identify and measure administrative burdens in federal government programs using the theory and requirements of performance management. In this chapter, I will extend that work to create a framework that facilitates the development and implementation of machine learning (ML) solutions that will reduce or eliminate the identified administrative burden from programs. This will include the considerations of ML design and implementation in the public sector based on federal government requirements and specific considerations interwoven into the best practices from industry and research.

As discussed already in this paper, it is likely that few ML programs are publicly known or available for my research that specifically states their goal of reducing administrative burdens, but I do now have the Framework 1 outline of the types of administrative burden costs to be baselined and monitored which will allow us to cross-reference any existing ML models and their intended goals to determine if they are directly or indirectly effecting administrative burdens based on the measurement criteria I outlined.

Framework Methods: Machine Learning Solutions for Administrative

Burdens

As discussed previously, there are many types of solutions for reducing administrative burdens, and I have discussed some of these, ranging from behavioral economics, decisions support, administrative law changes, and machine learning (ML) (Carrigan et al., 2020; Herd & Moynihan, 2018). In this paper, I am focusing solely on the use of ML techniques to reduce and eliminate administrative burdens. However, by doing so, I am expanding the research on administrative burdens by developing clear paths to focus on the design and implementation of solutions. I am also furthering the research on the development and implementation of applied machine learning in the federal government. In this framework, just as I did for Framework 1, I am beginning with the work explored in the administrative burden academic literature review and document analysis. Specifically, I will build on the development of the diagnostic questions to identify administrative burdens by cost types developed in Framework 1, which included specific performance measurement criteria to develop a performance management framework for identifying and reducing administrative burdens. In this framework, I will extend this work to include proposed ML solutions for each of the administrative burden costs being measured in a program. As described in Chapter 4, these are the specific milestones and indicators that the agency leadership will focus on to meet the overall strategic goal of reducing administrative burdens. Therefore, these ML solutions become the mechanisms and changes to the program which will allow the strategic goal to be reached.

To build this framework, I use the literature review of ML to identify specific types of ML solutions and categories from which to explore how each one might fit into an ML solution for each administrative burden cost identified in my Framework 1 data set. Just as the performance measurement methods identified in Framework 1 were not the only possible method, I will not be proposing every single possible ML method as a solution. Rather, I will develop and code the most probable solutions based on the specific challenge and the performance data available as identified in chapter 4. I will rely on the considerations and constraints I have identified and discussed through the literature review and document analysis in chapter 2 for ML solutions in the public sector to inform the ML solutions. Once these potential ML solutions for each identified cost are developed, I will further code them based on the specific type of ML and subtype of ML to help explore the types and methods of ML that are most applicable to the identified administrative burden challenges. In addition to this coding, I

will also generalize the ML solutions into a “Simplified ML Solution,” which is a generalized category of ML solutions for administrative burdens influenced by Herd and Moynihan’s (2018) work proposing general solutions, as shown in Table 2-4. These will be extended to administrative burdens and informed by the data set built from the administrative burden literature to explore and develop ML solutions approaches for future research and agency work on administrative burdens in the federal government.

This extension and coding of my data set will allow a further exploration and analysis of ML solutions in the federal government, which can be used to reduce administrative burdens but also potentially in other areas and for other purposes. I will then provide an example of an applied MLOps model as described in chapter 2, which allows the agency administrators to design, develop, and implement ML solutions in line with federal government requirements and considerations while also leveraging the industry and research best practices and requirements for developing effective and responsible ML solutions.

Finally, I will perform information mining of the federal register documents through querying for specific terms via an API and then using Natural Language Processing (NLP) to explore those documents to identify possible examples of applied ML in the federal government and to analyze which federal agencies are doing work on ML and for what purposes. This information mining will allow me to identify ML solutions in the federal government which has been designed and implemented in government programs to reduce administrative burdens. Identifying actual use cases will allow me to explore the requirements and considerations of Framework 2 against actual use cases in the federal government. This analysis will serve three purposes. First, it will allow me to examine my framework against an actual use case to determine any gaps or areas of weakness in my framework that should be addressed. Secondly, it will allow me to analyze the specific use case against the framework to identify any potential areas for improvement in the federal government implementation. Lastly, it will allow me to give an example to future researchers and administrators about how to use this framework to approach these tasks in the future.

Results: Machine Learning Framework for Administrative Burden

Reductions

I have reviewed the current state of ML in the federal government in chapter 2, including the requirements, principles, and best practices. I built on these findings to explore practical implementations of ML techniques and programs to reduce administrative burdens in

government programs. Just as I showed in chapter 4, I am unable to provide an exhaustive list of all potential administrative burden costs, but Framework 1 laid out a number of potential costs from the existing literature as well as measurement criteria to help researchers and administrators understand the existence and magnitude of administrative burdens in a particular program. In this section, I will extend those performance measurement criteria to suggest ML applications that can reduce the levels of administrative burden frictions for each type of cost.

<i>Type of ML</i>	<i>Sub-type</i>	<i>Definition</i>
<i>Supervised Learning</i>	Classification	ML program draws conclusions from observed values to determine a category for new observations.
<i>Supervised Learning</i>	Regression	The ML program must estimate the relationships between variables and make predictions on the independent variable.
<i>Supervised Learning</i>	Forecasting	ML program learns trends from historical data and variables and applies this knowledge to predict future trends in data.
<i>Unsupervised Learning</i>	Clustering	ML program clusters by grouping sets of similar data (based on defined criteria) from a larger set of data.
<i>Unsupervised Learning</i>	Dimension Reduction	The ML program reduces the number of variables being considered based on variables that have no impact on the target variable or removing variables that covary with each other.
<i>Reinforcement Learning</i>	Positive Reinforcement	An event occurs because of specific behavior that is desirable, and therefore the algorithm reinforces this behavior.
<i>Reinforcement Learning</i>	Negative Reinforcement	Strengthens behavior that occurs because of a negative condition or the absence of something which should be stopped or avoided.

Table 5-1: Types of ML Models (Burkov, 2019; Raschka & Mirjalili, 2019)

The first step in this extension of my data is to define and clarify the types and subtypes of ML from the current literature, as I have done in the above table. This can be a confusing space because there is often a conflation between specific definitions of ML, their algorithms, and the specific use cases. For instance, many will speak about “deep learning-enabled computer vision” since it has grown in popularity as many researchers work on autonomous vehicles, but “deep learning” is an algorithmic approach of unsupervised or reinforcement learning which contains several “hidden layers” within the neural net (Baylor et al., 2017; Raschka & Mirjalili,

2019). Therefore, for Framework 2, I want to focus on the three types of ML, supervised learning, unsupervised learning, and reinforcement learning, along with their subtypes, and stay away from the myriad iterations of specific algorithms and development approaches. As I discussed in chapter 2, these would be chosen as the ML model was designed, tuned, and tested in the development phase of the MLOps approach, and model choice is dependent on the ML solution design, the underlying data, and the choices about accuracy and transparency (Treveil et al., 2020). Furthermore, the specific algorithmic and development approaches are less useful at this stage since I am not using actual data to develop and test an ML solution yet. However, I am focused on pointing administrators and researchers toward the identification of ML approaches such as those identified in Figure 5-2 for the reduction of administrative burdens.

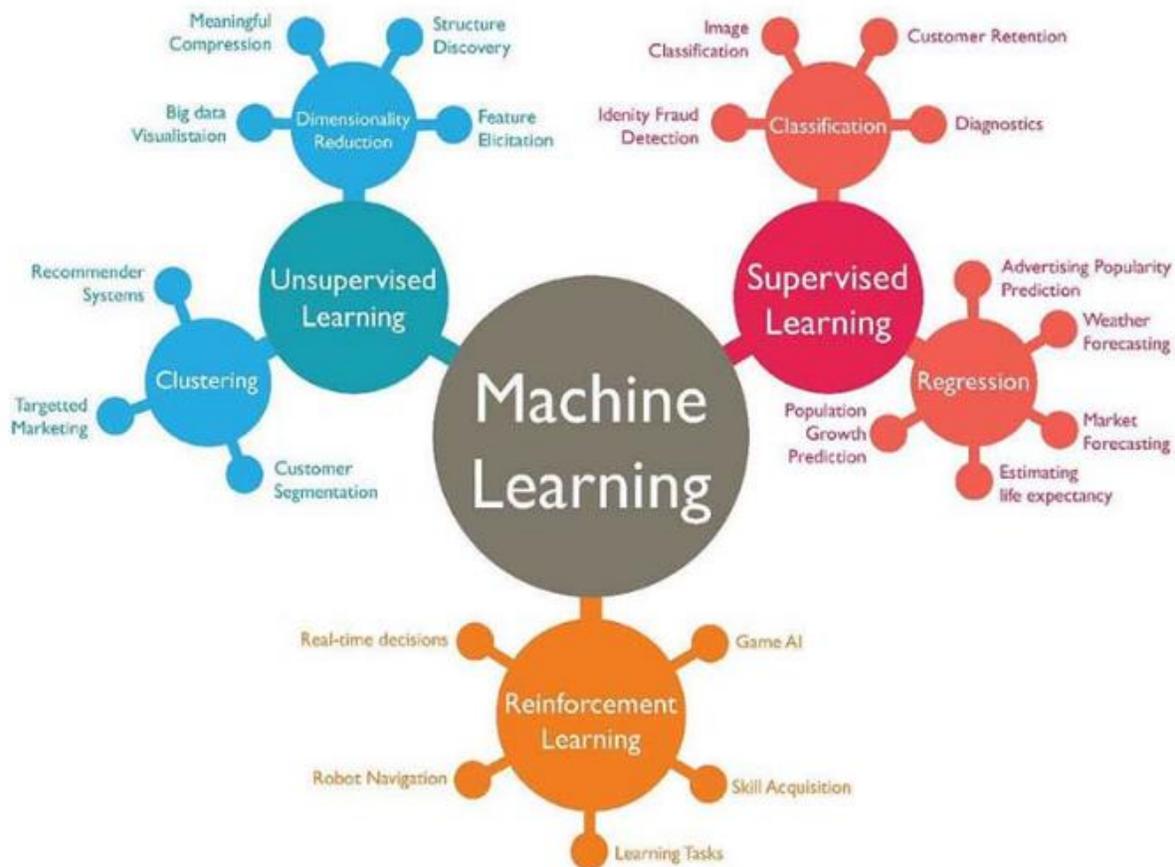


Figure 5-2: Machine Learning Types and Examples (Granville, 2017)

Leveraging the administrative burden performance measurement concepts, I assigned each a most likely ML solution from the available solutions from the three different types of ML, supervised, unsupervised, and reinforcement learning. These “Potential ML Solution” assignments, as seen in Table 5-4, Table 5-5, and Table 5-6, will not be the only possible ML solution but will leverage the existing literature and understanding of proven ML solutions, as

discussed in chapter 2 that are most likely to meet the specific needs of public sector requirements and provide a useful solution to the administrative burden costs as understood. This will help extrapolate the types of ML solutions most likely to help reduce administrative burdens but also provide some context about applied ML in the federal government.

Based on my existing literature data set of the types of administrative burden costs and performance measurements, I extended this data set (the full data set can be found in the appendix) by identifying an appropriate ML solution for each administrative burden cost example. From this ML solution, I created an explanatory solution of the ML program and then categorized these into seven distinct simplified types of ML solutions for administrative burdens, as well as identified the type of ML and subtype of ML that each solution would require. The simplified types of ML solutions for administrative burdens are assistance optimization, auto adjudication, autocompletion, autoenrollment, causal inference, location optimization, and virtual assistance. Of note, several administrative burden costs did not have a potential ML solution because the costs are most likely experienced because of indirect factors which could not be solved directly. Examples of these typically fall into psychological cost causes that cannot be solved except through indirect program changes or the experiences of other costs which are covered under other ML solutions directly. For example, increased psychological costs based on the stress experienced when participating in a program are tangential to aspects of the program and opinions and stigmas associated with program participation. Therefore this stress is not directly solvable through an ML solution (or other direct solutions). It should still be identified and measured as an aspect of psychological costs, but solutions to reduce this aspect likely come through other direct levels associated with participation and stigma causes. These received a “N/A” coding to indicate that I do not believe an applicable ML solution exists for that cost.

<i>Simplified ML Solution</i>	Count
Assistance Optimization	22
Autoadjudication	31
Autocompletion	9
Autoenrollment	18
Causal Inference	3
Location Optimization	2
N/A	7
Virtual Assistant	4

Total 96

Table 5-2: Count of Simplified ML Solutions

These different ML solution types can help us understand how to map ML solutions to the causes of administrative burden costs. It’s important to note that these are not the only ML solutions that might be applied or a definitive categorization of ML solutions. Several ML approaches to solutions could be devised for each cause. However, these proposed solutions are straightforward and avoid techniques that are more complex or complicated at the current time. This is a useful categorization of types of ML solutions specifically designed to address the costs of administrative burdens experienced in government programs. Of note, there are a few types that could be found in other domain areas. For example, I didn’t identify any types of ML computer vision despite this being one of the fastest-growing areas of ML application and research because these are not ML solutions that are likely to provide valuable assistance and reduction to administrative burden costs in government programs at this time. Potentially in the future, autonomous vehicles will be a cost-effective solution for helping transport individuals to in-person applications or benefit redemption processes or locations, but this is not an area of focus in most programs currently not a practical, cost-effective ML solution at this time.

Simplified ML Solution	Explanation
Assistance Optimization	ML solutions that predict and recommend methods, types, groups, or individuals to receive assistance, usually proactively before they request it.
Autoadjudication	ML solutions adjudicate decisions about eligibility, benefit types, amounts, and benefit redemption decisions.
Autocompletion	ML models autocomplete forms and other information collections from administrative data, user-submitted data, and open data sources through a variety of entity resolution techniques.
Autoenrollment	ML solutions automatically enroll individuals or groups into government services or benefits programs based on entity resolution and autoadjudication across single or multiple programs.
Causal Inference	ML models use administrative data, program data, and theoretical and logic models to predict causality between input variables and outcomes.

Location Optimization	ML solutions that predict the optimal types and placements of application assistance, benefit redemption, or other in-person offices involved in the program.
Virtual Assistant	ML solution that uses natural language processing to interact with individuals to provide assistance, information, and answer questions.

Table 5-3: Simplified ML Solution Definitions

Assistance Optimization is a category of ML solutions for administrative burdens which would harness the ability of ML solutions to provide predictions and optimization of groups and individuals who are most likely to need and benefit from additional assistance in their learning or compliance processes. This would likely look like proactively identifying individuals either through their interactions with web services, phone services, or in-person services or through the identification via administrative data (e.g., interacting with similar services, indications of medical statuses, employment statuses, disability services, etc.) that they have lower administrative capital or cognitive capacity or otherwise disparately impacted by administrative burdens in ways that practice assistance to them would be beneficial to assist them in navigating learning and compliance phases and steps for government programs.

Autoadjudication is ML solutions that are designed and implemented to make more accurate adjudication decisions from fewer data and more efficiently as compared to more manual adjudication methods. These solutions would be aimed at decreasing decision times, removing administrative and procedural errors, and more efficiently determining eligibility for program phases or directly for benefits, as well as adjudicating benefit redemption determinations more accurately and efficiently. For example, rather than requiring individuals and business owners to stay compliant with SNAP benefit redemption rules and requirements, these ML solutions could provide tools to determine whether certain products or services are compliant with any SNAP program rules, and similar solutions could improve the decision processes for benefit adjudications of Medicaid and other health benefit determinations to decrease associated administrative burdens experienced with redemption.

Autocompletion ML solutions would reduce administrative burdens associated with data collection, validation, and re-enrollment by using ML-based entity resolution to accurately predict from administrative data (including prior applications and program enrollment) data to auto-complete program forms and information required for applications and benefit redemptions. These could even be used to proactively complete information when individuals do not even

know they may be eligible for a program and present the completed forms to individuals for verification and “signature” with an explanation of programs and likely benefits amounts rather than requiring individuals to take these actions.

Autoenrollment builds on autocompletion and autoadjudication to finalize application, enrollment, and redemption steps to proactively determine eligibility and enroll individuals in programs based on administrative data. This could look like a process whereby required wages and employment information reported by a company indicates someone was laid off or otherwise meets requirements for unemployment insurance. Based on this prediction, an ML solution automatically compiles information required, adjudicates eligibility and benefit amounts, and begins the transfer of money to the individual. Additionally, the individuals’ information would be used to determine eligibility for other benefit programs such as SNAP benefits and Medicaid insurance for the individuals and their beneficiaries. This is potentially a more extreme solution than some may be comfortable with and would be more likely to require regulatory, policy, and potentially legal changes but could be a more impactful option to reduce administrative burdens. It is also informed by behavioral economics and behavioral public administration research that has shown the positive impacts on program participation and outcomes when program enrollment is the default (Bhanot & Linos, 2020; Congdon & Shankar, 2018).

Causal Inference would use ML solutions to help provide researchers and program administrators with more accurate and timely information about the specific causes of administrative burdens and associated negative impacts within programs allowing the targeting or policy and process changes to change outputs and outcomes as part of their performance management approach to program review to meet strategic goals.

Location Optimization is an ML solution to make predictions about the more effective, efficient types of locations for in-person services using open source and administrative program data. Often used by the private sector to optimize profits and other business factors, these ML solutions can help public organizations optimize for other outputs and outcomes which are most likely to reduce experienced administrative burdens and make the most positive impacts on program goals by reducing associated administrative burdens experienced by applicants.

Finally, Virtual Assistants are an ML solution that has become quite popular, such as Apple’s Siri, Google Assistant, and Amazon’s Alexa, but also the myriad chatbots, virtual operators, and computer-based services that we can interact with to gather information and make changes to digital experiences more effectively and efficiently than other manual processes. These have been beneficial because they have allowed the organization to save money when

compared to hiring people for these basic information and processing services when deployed well. There have also been studies showing that they are simpler to interact with than more manual digital and web-based programs for some individuals, including older individuals and individuals with certain types of disabilities. Public programs and services could more effectively help individuals understand program eligibility, application processes, benefits, and details about their own applications and accounts and be interacted with in many ways allowing for individuals to have the interactions and experiences they will find most useful. It would also help the government organization reserve more expensive assistance methods for more complex interactions, thereby reducing wait times and processing times across the board.

In the following sections, I will review the specific ML solutions, categories of ML, and Simplified ML Solutions for each administrative burden cause and measurement in my dataset.

ML for Learning Cost Solutions

The below table shows the ML solutions as well as the types and subtypes of ML models used for each performance measurement of learning costs experienced in the current administrative burden literature and ML literature. Extending the original measurement data set to provide ML solutions begin to make clear how administrators can seek to meet the strategic goals set out in Framework 1.

Refined Measurement	Simplified Measurement	Potential ML Solution	Simplified ML Solution	Type of ML	Subtype of ML
Customer feedback about a government entity	<i>Feedback Measurement</i>	<i>Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about government vs. non-government sources of information	<i>Feedback Measurement</i>	<i>Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rate	<i>Program Measurement</i>	<i>Targeted, proactive outreach to educate likely eligible individuals about the program.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program application process data disaggregated by participant	<i>Application Measurement</i>	<i>Proactive, differentiated outreach based on predictions of executive functioning levels.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>

executive function variables					
Survey of program participants and eligible non-participants	<i>Feedback Measurement</i>	<i>Practice targeted outreach based on likely eligible individuals with high non-take-up rates.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measure of program up-take based on nudges	<i>Program Measurement</i>	<i>Targeted, proactive outreach based on predicted lower take-up rates of individuals.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
The proportion of third-party entities performing compliance processes	<i>Assistance Measurement</i>	<i>Proactive assignment to a third-party group to assist with application based on predicted need/benefit.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Availability and use of eligibility tools	<i>Tool measurement</i>	<i>ML categorization model that predicts eligibility and benefits levels benefits based on minimum hypothetical data entered.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Availability and use of program and benefit tools	<i>Tool measurement</i>	<i>ML categorization model that predicts eligibility and benefits levels benefits based on minimum hypothetical data entered.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Amount of rules regarding program use	<i>Process Measurement</i>	<i>The burden for benefits redemption is automated and removed from the participant's responsibility.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measures of program learning based on responsibility (government versus individuals)	<i>Outreach Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Number of program eligibility determinations per application	<i>Process Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals, For multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoadjudication</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>

PRA Measures	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Outreach Measurement	<i>Outreach Measurement</i>	<i>Automatic proactive eligibility determination and targeted outreach about the program.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Application completion rates, disaggregated by protected category factors	<i>Application Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Program participant edge cases, take-up rate	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Analysis of Survey responses	<i>Feedback Measurement</i>	<i>The automated root-cause analysis process increases the accuracy of causal process tracing.</i>	<i>Causal Inference</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Increased opinions and experiences of government reliability based on increased accuracy to benefit adjudications and program administration (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Availability and use of application tools and services	<i>Tool measurement</i>	<i>Chat Bot is based on natural language processing that can answer questions about the application process via web or phone.</i>	<i>Virtual Assistant</i>	<i>Supervised Learning</i>	<i>Classification</i>

Table 5-4: ML Solutions for Leaning Costs (author)

ML for Compliance Cost Solutions

The below table shows the ML solutions as well as the types and subtypes of ML models used for each performance measurement of compliance costs experienced in the current administrative burden literature. Notably, some of the most common and obvious ML solutions for many of the compliance costs are “autocompletion, autoenrollment, and autoadjudication.” These are techniques that are looked at in much of the administrative burden research as well as behavioral economics because they change the default from having to enroll to having to unenroll. This is associated with increased levels of participation in many types of programs it has been tested (Bhanot & Linos, 2020; Congdon & Shankar, 2018; Grimmelikhuijsen et al., 2017). Additionally, combining administrative data with ML techniques is well suited to providing solutions and allowing for accurate processes which resolve entities, recommend categorization, and auto-process applications and benefits.

Refined Measurement	Simplified Measurement	Potential ML Solution	Simplified ML Solution	Type of ML	Subtype of ML
Measurement of the level of assistance available, disaggregated by individual needs	<i>Assistance Measurement</i>	<i>Prediction of individuals who will require additional assistance based on submitted information for targeted, proactive outreach. Success is measured by the proportion of the additional non-predicted population which requests additional assistance.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program compliance data disaggregated by participant executive function variables	<i>Program Measurement</i>	<i>Proactive, differentiated outreach based on predictions of executive functioning levels.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Survey of program participants and eligible non-participants	<i>Feedback Measurement</i>	<i>Proactive program outreach of likely eligible individuals</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>

Program application process data disaggregated by participant age/health factor variables	<i>Process Measurement</i>	<i>Targeted, Practice assistance to individuals predicted to need it based on age and health factors.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Compliance cost measures disaggregated by participant administrative experience (first-time applicant versus returning applicant)	<i>Application Measurement</i>	<i>Proactive, targeted outreach and program assistance based on a prediction of the level of individual agency of program experience.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Compliance steps available and taken by for third-party on individuals' behalf	<i>Assistance Measurement</i>	<i>Proactive assignment to the third-party group to assist with application based on predicted need/benefit.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Compliance measures borne by third parties versus individuals (when third parties are present)	<i>Process Measurement</i>	<i>Proactive assignment to a third-party group to assist with application based on predicted need/benefit.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Time spent on forms	<i>Form Measurement</i>	<i>Reduction of data needed based on a minimum amount of data to generate threshold accurate adjudication predictions/decisions</i>	<i>Autoadjudication</i>	<i>Unsupervised Learning</i>	<i>Dimensionality Reduction</i>
Amount of forms, time spent on forms	<i>Form Measurement</i>	<i>Reduction of data needed based on a minimum amount of data to generate threshold accurate adjudication predictions/decisions</i>	<i>Autoadjudication</i>	<i>Unsupervised Learning</i>	<i>Dimensionality Reduction</i>
Number of forms, time spent on forms	<i>Form Measurement</i>	<i>Prediction for adjudication of eligibility determinations based on submitted and administrative data to minimize those</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>

<p>Program take-up rate compared to the evaluation of program staff</p>	<p><i>Program Measurement</i></p>	<p><i>who require any additional verification or information submission. Automation of eligibility determinations at multiple steps in the program to increase the accuracy of eligibility determinations and reduce administrative discretion in a majority of cases.</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Program requirements for individuals versus organization</p>	<p><i>Process Measurement</i></p>	<p><i>Increased automated processes developed by the government reduce necessary inputs and participation from individuals.</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Rates of false negatives in program adjudication</p>	<p><i>Process Measurement</i></p>	<p><i>Automated program application and eligibility determinations.</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Measurement of program exits was still eligible</p>	<p><i>Program Measurement</i></p>	<p><i>Automated eligibility determination through administrative data and eligibility determinations, automated re-enrollment to decrease eligible non-take up rates.</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Measurement of program administrator discretion</p>	<p><i>Process Measurement</i></p>	<p><i>Automated information collection and eligibility determination for most individuals, reduced discretionary decisions by program staff.</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Measures of application requirements based on responsibility (government versus individuals)</p>	<p><i>Application Measurement</i></p>	<p><i>Automated application processes and eligibility determination shift the burden to the government from individuals.</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Amount of application</p>	<p><i>Application Measurement</i></p>	<p><i>Automated program application and</i></p>	<p><i>Autoadjudication</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>

and compliance forms		<i>eligibility determinations.</i>			
Measures of compliance costs compared to red tape measurement	<i>Process Measurement</i>	<i>Automated, more accurate processes based on entity resolution and eligibility predictions.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
PRA Measures	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
PRA Measures and benefits-costs analysis	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rates	<i>Program Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Application Completion Rates	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Application times	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Procedural denials (missed interview, document, requirements)	<i>Process Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>

Renewal Rates	<i>Application Measurement</i>	<i>Automated re-enrollment based on administrative data and adjudication predictions.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Accuracy Rates of Benefits	<i>Process Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Appeals resulting in decision reversal	<i>Process Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
False positives and negatives for benefits decisions	<i>Process Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
The churn rate of participants, especially eligible participants.	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Amount of application and compliance forms	<i>Form Measurement</i>	<i>Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Duplicity in information collections	<i>Form Measurement</i>	<i>Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>

		<i>then presented for correction or validation before completion.</i>			
The amount of information collected from applicants was available through administrative data sources	<i>Application Measurement</i>	<i>Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.</i>	<i>Autocompl etion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Time spent on forms	<i>Form Measurement</i>	<i>Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.</i>	<i>Autocompl etion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Are digital application and compliance options available, variety, and use	<i>Process Measurement</i>	<i>Automatic application completion through the use of administrative data via entity resolution minimized data collection from individuals.</i>	<i>Autocompl etion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measure of application "accelerators" - which are processes that speed up or make burdensome the application process for individuals	<i>Application Measurement</i>	<i>Increased application automation</i>	<i>Autocompl etion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measurements of application processes over time	<i>Application Measurement</i>	<i>Automated application completion through administrative data and entity resolution.</i>	<i>Autocompl etion</i>	<i>Supervised Learning</i>	<i>Classification</i>

How often do individuals need to re-apply, and what proportion of initial requirements must be resubmitted	<i>Application Measurement</i>	<i>Automatic re-enrollment based on predicted eligibility determination.</i>	<i>Autoenrollment</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Amount of time spent on the application process	<i>Application Measurement</i>	<i>Automatic enrollment and proactive request for validation of information/eligibility status based on administrative data.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rate	<i>Program Measurement</i>	<i>Automatic enrollment of predicted eligible individuals.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measures of the application process for renewals/extensions of the program	<i>Application Measurement</i>	<i>Automatic re-enrollment</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Number of program eligibility determinations per application	<i>Process Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Application completion rates	<i>Application Measurement</i>	<i>Automated application process based on entity resolution and use of administrative data.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rates disaggregated by protected category factors	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs</i>	<i>Autoenrollment</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>

<p>Program participant edge cases, take-up rate</p>	<p><i>Program Measurement</i></p>	<p><i>needed and eligible for. Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i></p>	<p><i>Autoenrollment</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Analysis of Survey responses</p>	<p><i>Feedback Measurement</i></p>	<p><i>The automated root-cause analysis process increases the accuracy of causal process tracing.</i></p>	<p><i>Causal Inference</i></p>	<p><i>Unsupervised Learning</i></p>	<p><i>Clustering</i></p>
<p>Variety of application processes and measurement of ease of use</p>	<p><i>Application Measurement</i></p>	<p><i>Predicting optimal placement of needed application support centers based on historical data, predictions of how individuals will wish to apply (online, by mail, via phone, in-person)</i></p>	<p><i>Location Optimization</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Measurement of distance from individuals to application/program center</p>	<p><i>Assistance Measurement</i></p>	<p><i>More efficient application center location for individuals.</i></p>	<p><i>Location Optimization</i></p>	<p><i>Unsupervised Learning</i></p>	<p><i>Clustering</i></p>
<p>Comparison of government and non-government options</p>	<p><i>Process Measurement</i></p>	<p><i>Increased trust ineffectiveness of government services (auxiliary effect).</i></p>	<p><i>N/A</i></p>	<p><i>N/A</i></p>	<p><i>N/A</i></p>
<p>Wait times for phone and in-person conversations, email response times</p>	<p><i>Assistance Measurement</i></p>	<p><i>Chatbots are accessible via web, phone, or email and can accurately answer questions and maximize personal phone operators for more advanced issues.</i></p>	<p><i>Virtual Assistant</i></p>		
<p>Call abandonment rate, call answered rates</p>	<p><i>Assistance Measurement</i></p>	<p><i>Chatbot to increase call assistance and first-time resolution rates.</i></p>	<p><i>Virtual Assistant</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>First call resolution rates</p>	<p><i>Assistance Measurement</i></p>	<p><i>Chatbot to increase call assistance and</i></p>	<p><i>Virtual Assistant</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>

first-time resolution rates.

Table 5-5: ML Solutions for Compliance Costs (author)

ML for Psychological Cost Solutions

The below table shows the ML solutions as well as the types and subtypes of ML models used for each performance measurement of psychological costs experienced in the current administrative burden literature. As discussed previously, some causes and measurements of psychological costs are tangential or indirect because they involve experienced stress or stigma that is associated with other phenomena that can not be solved directly with ML solutions. Therefore, I have coded them as “N/A” in my data set. The belief and the hope are that these measured psychological costs will also decrease through the implementation of other ML solutions aimed at the component costs causes, and they should still be identified and measured in Framework 1 even though they cannot be directly impacted.

Refined Measurement	Simplified Measurement	Potential ML Solution	Simplified ML Solution	Type of ML	Subtype of ML
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about prior government interactions	<i>Feedback Measurement</i>	<i>Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Measurement of application step wait times, customer	<i>Application Measurement</i>	<i>Automatic data completion, enrollment, re-enrollment from administrative</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>classification</i>

feedback about experiences		<i>data, and proactive benefits eligibility determinations and requests for verification to individuals.</i>			
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Population attitude towards program and recipients	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in an understanding of individuals' eligibility, or making program participation opaquer.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
The measure of compliance costs as compared to customer feedback about the program	<i>Feedback Measurement</i>	<i>Decrease compliance requirements through automatic form completion and automatic enrollment/re-enrollment.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Increased sense of autonomy based on program changes and reduced compliance costs (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>

Customer feedback about program staff	<i>Feedback Measurement</i>	<i>Potential auxiliary impact of other program changes.</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Program take-up rate	<i>Program Measurement</i>	<i>Targeted proactive outreach changes for individuals less likely to successfully access the program.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Analysis of Survey responses	<i>Feedback Measurement</i>	<i>The automated root-cause analysis process increases the accuracy of causal process tracing.</i>	<i>Causal Inference</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Program survey feedback disaggregated by participant executive function variables	<i>Feedback Measurement</i>	<i>Proactive, differentiated outreach based on predictions of executive functioning levels.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in the understanding of individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>Autoenrollment</i>	<i>N/A</i>	<i>N/A</i>
Direct measurements of program participant's feelings	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>Autoenrollment</i>	<i>N/A</i>	<i>N/A</i>
Participant feedback on the program	<i>Feedback Measurement</i>	<i>Automated application completion and</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>

as compared to individuals compliance costs	<i>program adjudication.</i>				
Program take-up rates disaggregated by protected category factors and compared to participant feedback on the program	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Measures of learning and compliance costs compared to applicant feedback about the program	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>Autoenrollment</i>	<i>N/A</i>	<i>N/A</i>
Program participant edge cases, take-up rate	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measures of compliance costs borne by participants versus government	<i>Process Measurement</i>	<i>Automated, more accurate processes based on entity resolution and eligibility predictions.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
PRA Measures	<i>Application Measurement</i>	<i>Automated application completion and eligibility determination based on the use of</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>

Customer satisfaction rates	<i>Feedback Measurement</i>	<i>administrative data and adjudication. Auxiliary effect based on increased accuracy of program administration and decreased experienced burdens.</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Appeals resulting in decision reversal	<i>Process Measurement</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about prior government interactions	<i>Feedback Measurement</i>	<i>Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rate	<i>Program Measurement</i>	<i>Targeted proactive outreach changes for individuals less</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>

		<i>likely to successfully access the program.</i>			
Program survey feedback disaggregated by participant executive function variables	<i>Feedback Measurement</i>	<i>Proactive, differentiated outreach based on predictions of executive functioning levels.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about prior government interactions	<i>Feedback Measurement</i>	<i>Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rate	<i>Program Measurement</i>	<i>Targeted proactive outreach changes for individuals less likely to successfully access the program.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program survey feedback disaggregated	<i>Feedback Measurement</i>	<i>Proactive, differentiated outreach based on predictions of</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>

by participant executive function variables		<i>executive functioning levels.</i>			
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about the application process	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about prior government interactions	<i>Feedback Measurement</i>	<i>Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Targeted proactive outreach to individuals based on predicted psychological costs experienced.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program take-up rate	<i>Program Measurement</i>	<i>Targeted proactive outreach changes for individuals less likely to successfully access the program.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Program survey feedback disaggregated by participant executive function variables	<i>Feedback Measurement</i>	<i>Proactive, differentiated outreach based on predictions of executive functioning levels.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measures of compliance	<i>Process Measurement</i>	<i>Automated, more accurate processes based on entity</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>

<p>costs borne by participants versus government</p>	<p><i>resolution and eligibility predictions.</i></p>
<p>PRA Measures</p>	<p><i>Application Measurement Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>Appeals resulting in decision reversal</p>	<p><i>Process Measurement Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>Measures of compliance costs borne by participants versus government</p>	<p><i>Process Measurement Automated, more accurate processes based on entity resolution and eligibility predictions.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>PRA Measures</p>	<p><i>Application Measurement Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>Appeals resulting in decision reversal</p>	<p><i>Process Measurement Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>Measures of compliance costs borne by participants versus government</p>	<p><i>Process Measurement Automated, more accurate processes based on entity resolution and eligibility predictions.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>PRA Measures</p>	<p><i>Application Measurement Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>
<p>Appeals resulting in</p>	<p><i>Process Measurement Automated application</i></p> <p><i>Autoadjudication</i></p> <p><i>Supervised Learning</i></p> <p><i>Classification</i></p>

decision reversal		<i>completion and eligibility determination based on the use of administrative data and adjudication.</i>			
A measure of compliance costs as compared to customer feedback about the program	<i>Feedback Measurement</i>	<i>Decrease compliance requirements through automatic form completion and automatic enrollment/re-enrollment.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Participant feedback on the program as compared to individuals compliance costs	<i>Feedback Measurement</i>	<i>Automated application completion and program adjudication.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
A measure of compliance costs as compared to customer feedback about the program	<i>Feedback Measurement</i>	<i>Decrease compliance requirements through automatic form completion and automatic enrollment/re-enrollment.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Participant feedback on the program as compared to individuals compliance costs	<i>Feedback Measurement</i>	<i>Automated application completion and program adjudication.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
A measure of compliance costs as compared to customer feedback about the program	<i>Feedback Measurement</i>	<i>Decrease compliance requirements through automatic form completion and automatic enrollment/re-enrollment.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Participant feedback on the program as compared to individuals compliance costs	<i>Feedback Measurement</i>	<i>Automated application completion and program adjudication.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measurement of application step wait times, customer	<i>Application Measurement</i>	<i>Automatic data completion, enrollment, re-enrollment from administrative</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>classification</i>

feedback about experiences		<i>data, and proactive benefits eligibility determinations and requests for verification to individuals.</i>			
Population attitude towards program and recipients	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in an understanding of individuals' eligibility, or making program participation opaquer.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>Autoenrollment</i>	<i>N/A</i>	<i>N/A</i>
Direct measurements of program participant's feelings	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>Autoenrollment</i>	<i>N/A</i>	<i>N/A</i>
Program take-up rates disaggregated by protected category factors and compared to participant	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for</i>	<i>Autoenrollment</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>

feedback on the program		<i>multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>			
Measures of learning and compliance costs compared to applicant feedback about the program	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>Autoenrollment</i>	<i>N/A</i>	<i>N/A</i>
Program participant edge cases, take-up rate	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Measurement of application step wait times, customer feedback about experiences	<i>Application Measurement</i>	<i>Automatic data completion, enrollment, re-enrollment from administrative data, and proactive benefits eligibility determinations and requests for verification to individuals.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>classification</i>
Population attitude towards program and recipients	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>

<p>Program take-up rates disaggregated by protected category factors and compared to participant feedback on the program</p>	<p><i>Program Measurement</i></p>	<p><i>participation opaquer. Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i></p>	<p><i>Autoenrollment</i></p>	<p><i>Unsupervised Learning</i></p>	<p><i>Clustering</i></p>
<p>Program participant edge cases, take-up rate</p>	<p><i>Program Measurement</i></p>	<p><i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i></p>	<p><i>Autoenrollment</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Measurement of application step wait times, customer feedback about experiences</p>	<p><i>Application Measurement</i></p>	<p><i>Automatic data completion, enrollment, re-enrollment from administrative data, and proactive benefits eligibility determinations and requests for verification to individuals.</i></p>	<p><i>Autoenrollment</i></p>	<p><i>Supervised Learning</i></p>	<p><i>classification</i></p>
<p>Population attitude towards program and recipients</p>	<p><i>Feedback Measurement</i></p>	<p><i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation opaquer.</i></p>	<p><i>Autoenrollment</i></p>	<p><i>Supervised Learning</i></p>	<p><i>Classification</i></p>
<p>Program take-up rates disaggregated by protected</p>	<p><i>Program Measurement</i></p>	<p><i>Automated application processes and eligibility</i></p>	<p><i>Autoenrollment</i></p>	<p><i>Unsupervised Learning</i></p>	<p><i>Clustering</i></p>

category factors and compared to participant feedback on the program		<i>determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>			
Program participant edge cases, take-up rate	<i>Program Measurement</i>	<i>Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
Analysis of Survey responses	<i>Feedback Measurement</i>	<i>The automated root-cause analysis process increases the accuracy of causal process tracing.</i>	<i>Causal Inference</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Analysis of Survey responses	<i>Feedback Measurement</i>	<i>The automated root-cause analysis process increases the accuracy of causal process tracing.</i>	<i>Causal Inference</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Analysis of Survey responses	<i>Feedback Measurement</i>	<i>The automated root-cause analysis process increases the accuracy of causal process tracing.</i>	<i>Causal Inference</i>	<i>Unsupervised Learning</i>	<i>Clustering</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Increased sense of autonomy based on program changes and reduced compliance costs (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>

Customer feedback about program staff	<i>Feedback Measurement</i>	<i>Potential auxiliary impact of other program changes.</i>	N/A	N/A	N/A
Customer satisfaction rates	<i>Feedback Measurement</i>	<i>Auxiliary effect based on increased accuracy of program administration and decreased experienced burdens.</i>	N/A	N/A	N/A
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	N/A	N/A	N/A
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	N/A	N/A	N/A
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Increased sense of autonomy based on program changes and reduced compliance costs (auxiliary effect).</i>	N/A	N/A	N/A
Customer feedback about program staff	<i>Feedback Measurement</i>	<i>Potential auxiliary impact of other program changes.</i>	N/A	N/A	N/A
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	N/A	N/A	N/A
Direct measurements of program participant's feelings	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or</i>	N/A	N/A	N/A

		<i>making program participation more opaque (auxiliary effect).</i>			
Measures of learning and compliance costs compared to applicant feedback about the program	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer satisfaction rates	<i>Feedback Measurement</i>	<i>Auxiliary effect based on increased accuracy of program administration and decreased experienced burdens.</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Direct measurements of program participant's feelings	<i>Feedback Measurement</i>	<i>Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Measures of learning and	<i>Feedback Measurement</i>	<i>Automatic enrollments and</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>

compliance costs compared to applicant feedback about the program		<i>re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).</i>			
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Decreased experienced stress based on program changes (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program use	<i>Feedback Measurement</i>	<i>Increased sense of autonomy based on program changes and reduced compliance costs (auxiliary effect).</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer feedback about program staff	<i>Feedback Measurement</i>	<i>Potential auxiliary impact of other program changes.</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Customer satisfaction rates	<i>Feedback Measurement</i>	<i>Auxiliary effect based on increased accuracy of program administration and decreased experienced burdens.</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>

Table 5-6: ML Solutions for Psychological Costs (author)

In Table 5-7 below, I show the types of ML solutions based on the three types of administrative burden costs. From this analysis, several insights emerge. First, and perhaps not surprisingly, autoadjudication, autoenrollment, assistance optimization, and autocompletion are the most popular ML solutions for experienced compliance costs. This is because many aspects of the compliance processes in programs could be automated through ML, thereby taking most of the burden off individuals and placing it on the government organization through the ML solution. Assistance optimization is the most likely ML solution for learning and psychological costs experienced. This is because proactively engaging with individuals and pushing program

and benefit information is likely to help alleviate many of the causes of experienced learning costs and the psychological costs and stigma experienced by individuals. Psychological costs are more likely to not have an ML solution (coded as “N/A”) because they are indirect or auxiliary-caused experiences that cannot be addressed directly.

<i>Type of Cost</i>	<i>Simplified ML Solution</i>	<i>Count</i>
Compliance Costs	Assistance Optimization	7
	Autoadjudication	22
	Autocompletion	7
	Autoenrollment	8
	Causal Inference	1
	Location Optimization	2
	N/A	1
	Virtual Assistant	3
Compliance Costs Total		51
Learning Costs	Assistance Optimization	7
	Autoadjudication	6
	Autoenrollment	3
	Causal Inference	1
	N/A	1
	Virtual Assistant	1
Learning Costs Total		19
Psychological Costs	Assistance Optimization	8
	Autoadjudication	3
	Autocompletion	2
	Autoenrollment	7
	Causal Inference	1
	N/A	5
Psychological Costs Total		26
Grand Total		96

Table 5-7: Count of Simplified ML Solution per Cost Types

Also of note is that seventy-four, as shown in Table 5-8, which is the majority by far, of the identified ML solutions, are based on classification, which is a type of supervised learning, and there are no ML solutions that would use reinforcement learning. Only twelve of the ML solutions use unsupervised learning, and only two of those would be based on dimensionality

reductions. This shows us that the currently less complex ML solutions are viable and might be best for public sector use to reduce administrative burdens. This is good news considering the current state of ML work and use in the public sector, including the public sector’s delays in information technology infrastructure and competition for a skilled workforce that I discussed in chapter 2. This indicates that there are many potential opportunities to reduce administrative burdens through ML techniques that are accessible to the federal government.

<i>Counts of ML types and Subtypes</i>				
	Supervised Learning	Unsupervised Learning	Unsupervised Learning	Grand Total
<i>Simplified ML Solution</i>	Classification	Clustering	Dimensionality Reduction	
Assistance Optimization	22			22
Autoadjudication	28	1	2	31
Autocompletion	9			9
Autoenrollment	10	5		18
Causal Inference		3		3
Location Optimization	1	1		2
N/A				7
Virtual Assistant	4			4
Grand Total	74	10	2	96

Table 5-8: Counts of ML Solution Types per ML Solution Categories (author)

Results: MLOps Example for ML Solutions

As discussed previously, MLOps is an approach to the development, deployment, and monitoring (operations) of ML solutions in specific use cases that are based on leading industry and research standards, best practices, and norms (Treveil et al., 2020). These phases of ML solutions are important to ensure that the ML use case is appropriately identified and designed to solve the problem at hand, the ML model is accurately trained, tuned, and reviewed according to statistical requirements as well as legal and ethical requirements, and then to ensure the model deployed (in operations) is appropriately monitored to make sure that it continues to be accurate and appropriate for the particular use case.

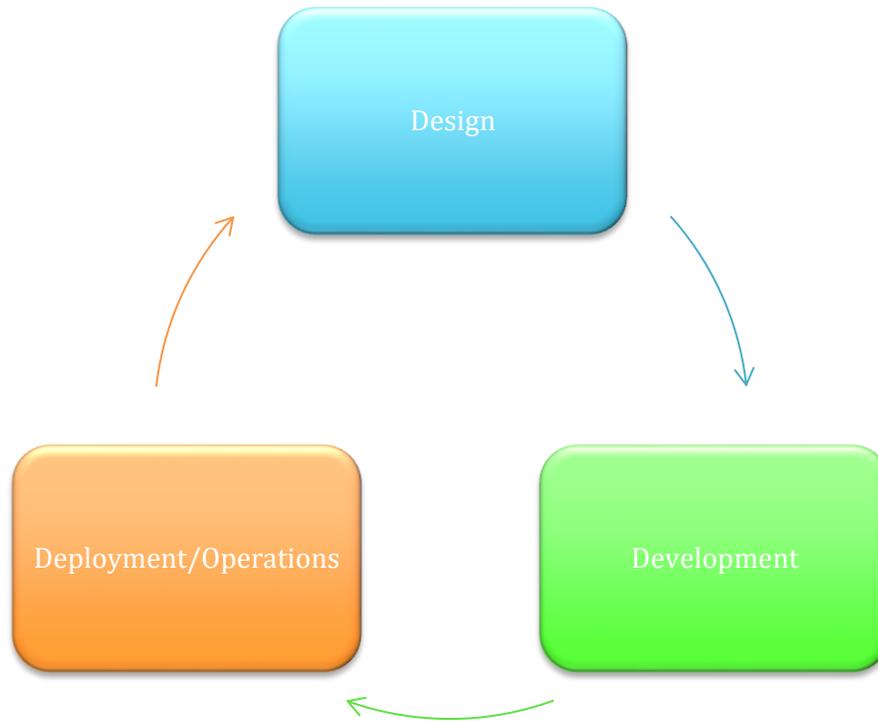


Figure 5-3: MLOps Cycle

These MLOps phases should be applied to each ML solution for administrative burden causes and reduction. Specific concerns need to be addressed as applied to the use case, the data available for the potential ML solution, how the ML model is designed and developed, and the consideration of evaluating the model for accuracy, validity, and biases. For example, when approaching an ML solution that will provide autoadjudication, practitioners will need to ensure that the use case makes sense for the ML approach. As discussed in chapter 2, there are also principles from E.O. 13960 to comply with for federal government ML solutions, many of which map to existing MLOps requirements as I have laid out in Table 5-9, but also include the requirements for transparency which can be addressed through administrative procedure act requirements such as specifying the use of ML solution in program notices, information collections, and adjudication and oversight requirements to ensure the public and individuals understand how their data is being used and how the government is using ML solutions to make program adjudications and decisions.

E.O. Principle	MLOps Mapping
Lawful and Respectful of our Nation’s Values	Model Development: Initial concept and model research and design

Purposeful and performance-driven	Model Development and Model Monitoring: Evaluating the outputs of the outcomes of the ML implementation.
Accurate, reliable, and effective.	Model Development: Model training and assessment; Model Monitoring: model accuracy and drift assessments, as well as model evaluations.
Safe, secure, and resilient.	Model Monitoring: Assessment of deployment model, including model logging and cyber security monitoring.
Understandable	Model Development: When developing and testing the model, can the results of the model be explained and transparent.
Responsible and Traceable	Model Development and Model Deployment: Documentation for oversight and reproducibility. Also, the deployment processes and standard operating procedures within the programs they are being deployed.
Regularly Monitored	Model Monitoring: As part of the ML monitoring plan, assessing the model against E.O. 13960 requirements (along with any additional legal or regulatory requirements).
Transparent	Not specifically addressed in MLOps, but this would be the process of including documentation, compliance checks, and other pertinent information in public and oversight workstreams to ensure there are no “hidden ML projects” within the federal government.
Accountable	MLOps Governance: This could be the agencies detailing their MLOps processes and having an outside entity monitor or audit them, or this could be judicial or congressional oversight of agency MLOps processes.

Table 5-9: Responsible AI Principles Mapped to MLOps Process

Table 5-9 shows how following current industry best practices of MLOps can facilitate agencies' compliance with the responsible AI principles as required by E.O. 13960. Beyond the compliance requirements, this will ensure agency administrators are adequately considering and documenting important decisions and processes during the design, development, and implementation of ML solutions in the government sector. In the next section, I will go through

an MLOps checklist with one of the ML solution's general types to show how this would be applied.

MLOps Checklist Example

For this example, MLOps walk-through, I will use the simplified ML solution of autoadjudication. As discussed previously, autoadjudication is an ML solution that is designed and implemented to make more accurate adjudication decisions from fewer data and to do so more efficiently as compared to manual adjudication methods. For my example, autoadjudication is an ML solution that uses supervised learning to categorize individuals as “eligible,” “not eligible,” or “needs review” based on limited application data, as well as administrative data about the individual that the government has access to. Below I will go through an example MLOps review based on an autoadjudication ML model for determining benefit eligibility based on predicting eligibility from the data through the categorization technique.

Design

Problem Development

- **What is the problem trying to be solved?** Determining effectively and efficiently if an individual is eligible or not eligible for program benefits.
- **What is the definition of success? What are the acceptance tests?** If the ML solution can determine eligibility at least as accurately as manual adjudication rates and more efficiently based on time and resources, then it should be used.
- **What are the performance indicators that would be used to measure success?** Accuracy rates of eligibility predictions across all individual types. The efficiency of predictions as compared to manual adjudications.

Data

- **Do you have access to the data needed?** The agency has access to more than ten years of application and program data; through data, sharing has access to individual data from multiple records systems for identity resolution, including socioeconomic indicators.
- **Can the data be used for this purpose?** The agency has updated our regulations and information collections to notify the public about this use of their data, how these adjudications will be made, and other pertinent information about the ML process for oversight.

- **Is the data up-to-date and accurate?** Data sharing agreements and data pipelines provide real-time access to all data sources.
- **Is the data representative of the population of interest?** New groups of individuals who never applied or were found eligible before are under-represented in the data. As program changes are made which create this gap, data transformation will be used to provide sufficient representation in training data.
- **Could using this data lead to biased results?** Because prior adjudication results will be used, and there is a potential for biased adjudications previously made by human adjudicators, the data and the model results will be tested for biases. If found, data will be transformed to create unbiased training and testing data to reduce or eliminate training bias.

Baselining

- **What is the baseline to measure the model against?** The baseline for comparison of the ML model is the results and indicators of performance for manual adjudications being replaced by the ML autoadjudication model.

Development

Preprocessing

- **What are the results from the exploratory data analysis?** EDA results indicate no biases in the underlying training data, and the data is sufficiently clean for ML model development.
- **Are the assumptions about the model documented and translated into automated checks?** Assumptions about how the model will be documented and automated checks are coded into the procedure, which will notify of any actions which fail these checks.
- **Are the data cleaning, preprocessing, and feature tuning steps documented for replication?** All preprocessing, training, and tuning steps taken are documented to allow for reproducibility.
- **How is missing data handled, and why?** No missing data is imputed into any of the data sets because we have multiple data sets to perform entity resolution, and the model has been tested and remains accurate without any missing data.

Model Evaluation

- **Does the ML solution work based on predetermined criteria?** The model is accurate based on predetermined criteria (accuracy, precision, sensitivity, F1 scores, specificity, AUC, etc.)
- **Do you have documented model performance metrics and acceptance ranges?** Model performance parameters with associated tolerance ranges are documented and coded to trigger model stoppage if thresholds are exceeded.
- **Check for overfitting, biases, and data leaks?** Model and data are evaluated for overfitting and bias.
- **Manually check misclassified examples. When does the model make mistakes?** There are no detectible patterns or causes for misclassifications.
- **Are the code, dependencies, and technical requirements documented?** All code and technical environment requirements and dependencies are documented and reproducible.

*Deployment/Operations**Production Technical Requirements*

- **Will the ML model run in real-time or in batches?** Autoadjudication will run in batch once daily for all applications meeting requisite data requirements and will generate results and notifications the following business day.
- **Will there be additional processes or reviews for anomalous predictions?** Any anomalous results, as defined by the model and data thresholds and evaluation metrics, will be held in abeyance and reviewed the following business day.
- **What metadata, artifacts, and audit logs will be collected? When will they be reviewed?** All model metrics, adjudication and prediction metrics, and results will be logged. Representative samples of logs will be reviewed for accuracy and compliance once per week.
- **How often does the model need to be retrained?** The model must be retrained when policy or program changes are made regarding eligibility.

Monitoring

- **What data quality metrics need to be monitored in production?** Model and prediction accuracy, computational resources, and runtimes will be monitored.

- **What model criteria need to be monitored?** Accuracy, specificity, and F1 scores appropriate for the classification model will be monitored as well as data and model bias.
- **How will input drift and model degradation be monitored?** ML model predictions will be compared to eligibility criteria and determinations based on random sampling to determine model drift (becoming less accurate).
- **What feedback loops need attention?** Predictions of “eligible” where individuals decline benefits; predictions on “not eligible” where appeals are filed to determine if eligibility was wrongly predicted.

As can be seen from the above example, there are many important considerations and factors which go into ML design, development, and deployment in the public sector, and the government will need to focus on foundational issues when taking these approaches, such as data sharing, data quality, workforce training and culture, and IT infrastructure. However, these are achievable goals, and these considerations, while important, are standard across many instances of ML implementations in the private and public sectors.

Results: Information Mining for Applied ML Solutions

As stated, I wanted to see what examples of applied ML I could find in the federal government, so I searched through federal register documents based on key terms to find all documents which might be indicative of the federal government announcing the use of, development of, or the results of applied ML. Form this corpus, I will further refine through NLP techniques to determine if any of the examples are in a program or implementation which is likely to reduce administrative burdens experienced by individuals.

Count of Serch Terms in Federal Register

Search Term	
Artificial Intelligence	360
Machine Learning	180
Natural Language Process..	47
Reinforcement Learning	6
Supervised Learning	4
Unsupervised Learning	2

Table 5-10:Count of Search Terms

As can be seen in Table 5-10, there are more instances of broad categories of “Artificial Intelligence” and “Machine Learning” than there are the specific subtypes of machine learning which indicates that if there is information in the federal register about the use of ML in the federal government, then it is not specific about the ML approach. There are only eleven instances of the use of “Supervised Learning,” “Unsupervised Learning,” and “Reinforcement Learning” combined in the federal register. Additionally, and not surprisingly, there were few (less than ten per year) instances of documents in the federal

register containing these terms until 2016, when the prevalence of these terms increased precipitously. Additionally, as can be seen in Table 5-11, most agencies have less than ten documents in the federal registers that contain any of the terms.

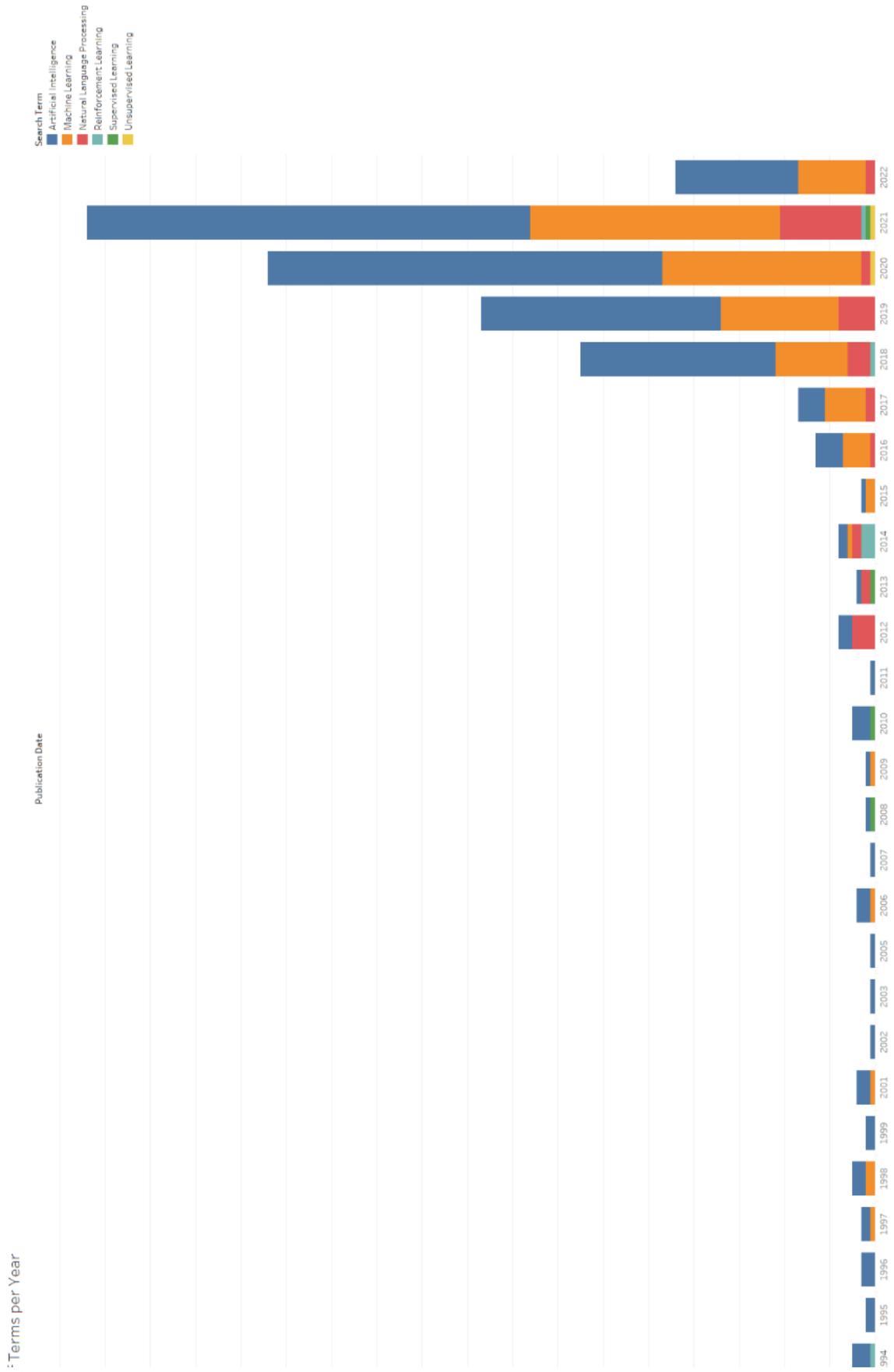


Table 5-11: Search Terms Found by Year of Publication

Count of Terms per Year Since 2010

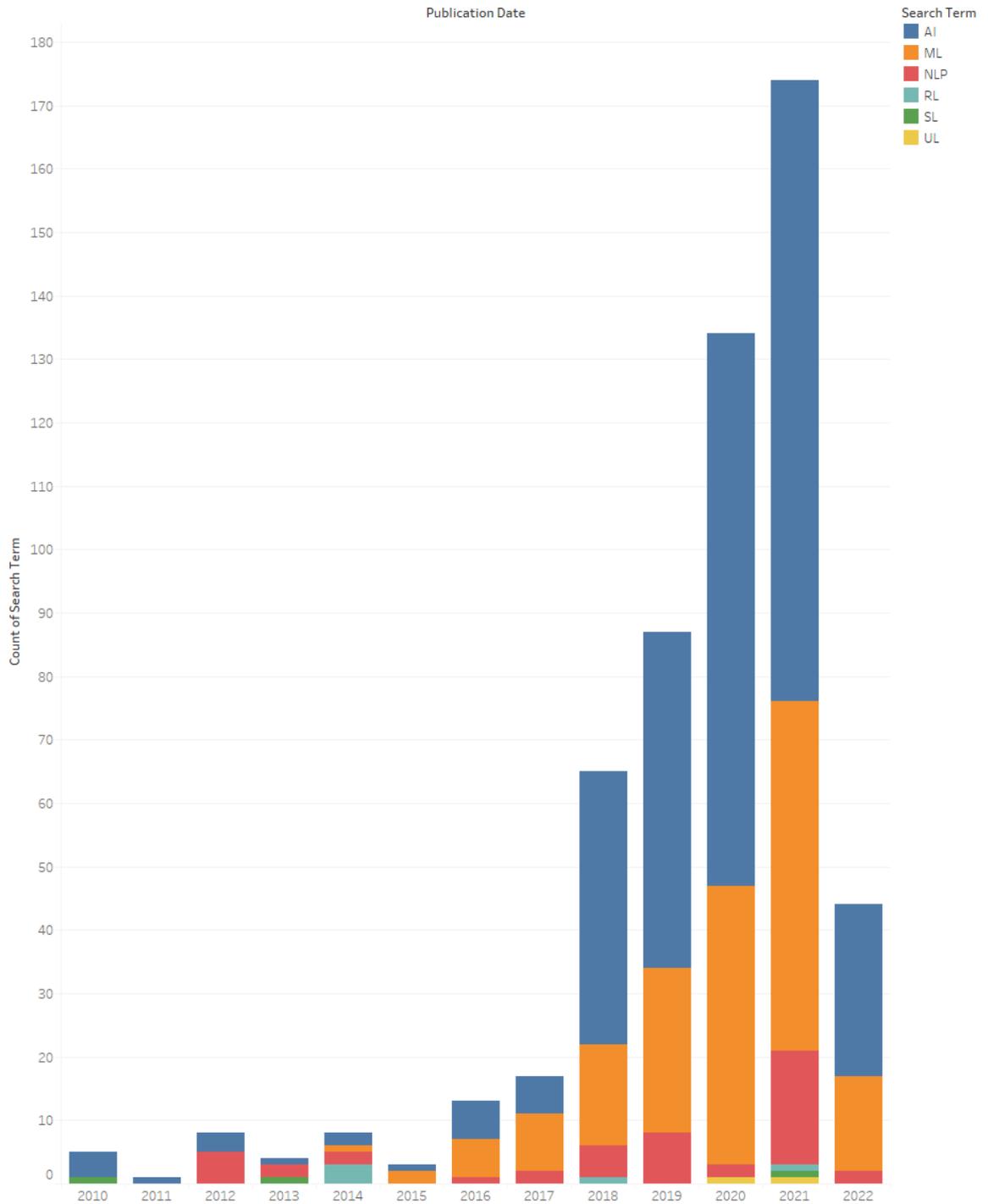


Table 5-12: Search Terms by Year 2010-2022 (YTD)

Terms by Agency

Agency Names	Search Term						Grand ..
	AI	ML	NLP	RL	SL	UL	
Health and Human Services Department	58	38	29		1	1	127
Commerce Department	59	22	2	1			84
Securities and Exchange Commission	18	13	1	3	1	1	37
National Science Foundation	27	6					33
Education Department	27	1	2				30
Energy Department	17	9					26
Consumer Financial Protection Bureau	11	13					24
Executive Office of the President	17	5					22
Defense Department	15	5		1			21
Treasury Department	10	7	3				20
Federal Communications Commission	12	6	1				19
National Security Commission on Artificial Int..	9	8					17
Homeland Security Department	11	6					17
Justice Department	11	4	1				16
Transportation Department	7	5					12
Science and Technology Policy Office	5	4					9
Federal Deposit Insurance Corporation	4	4	1				9
Administrative Conference of the United States	5	3	1				9
Veterans Affairs Department	2	1	2				5
National Credit Union Administration	2	2	1				5
Labor Department	2	1			2		5
Consumer Product Safety Commission	3	2					5
Nuclear Regulatory Commission	2	1	1				4
Library of Congress	2	2					4
Environmental Protection Agency	1	2	1				4
State Department	3						3
Housing and Urban Development Department	1	2					3
Federal Trade Commission	2	1					3
Federal Reserve System	1	1	1				3
Management and Budget Office	2						2
Legal Services Corporation	1	1					2
Export Import Bank	2						2
Commodity Futures Trading Commission	1	1					2
Regulatory Information Service Center	1						1
Privacy and Civil Liberties Oversight Board	1						1
Personnel Management Office	1						1
National Foundation on the Arts and the Huma..		1					1
National Archives and Records Administration	1						1
Interior Department		1					1
Final Rules	1						1

Table 5-13: Results of Search Terms by Agency from all years

The types and subtypes of documents with the search terms in the top fifteen agencies are in Table 5-14 below. The difference between a notice and a rule is important for our analysis. Notices are documents other than rules and proposed rules which are often required to be posted by law. Notices can be hearings and investigations, committee meetings, agency decisions, delegations of authority, issuances or revocations of licenses, grant application information, environmental impact statements, and agency organizational or structural changes. Rules and proposed rules, on the other hand, impact program rules, regulations, and processes. The rules

Top 15 Agencies by Total Terms Disaggregated by Document Types and Subtypes

Agency Names	Type / Subtype							Grand Total
	Notice	Presidential Document			Proposed Rule	Rule	Uncategorized Document	
		Executive Order	Memorandum	Proclamation				
Health and Human Services Department	61				29	35		125
Commerce Department	70				8	5		83
National Science Foundation	33							33
Securities and Exchange Commission	21				8	2	1	32
Education Department	24				3	3		30
Energy Department	25				1			26
Consumer Financial Protection Bureau	17				5	2		24
Executive Office of the President	4	6	4	8				22
Treasury Department	9				7	4		20
Defense Department	17				2		1	20
Federal Communications Commission	14				2	3		19
National Security Commission on Artificial Intelligence	17							17
Homeland Security Department	7				4	6		17
Justice Department	16							16
Transportation Department	7				2	2	1	12

Table 5-14: Top 15 Agency Terms by Document Types

and proposed rules documents are where I would expect to find agencies notifying the public of the use of ML techniques for government program and service applications, decisions, and processing. Since proposed rules are posted for public comments that must be addressed and modifications made (if applicable) before posting the finalized rule, I will filter down to show the rule titles for agencies where the terms exist. In Table 5-15 below, reading the titles of the rules posted, very few are likely to relate to the use in government programs and services and therefore not likely to impact administrative burdens in either direction. The most promising rules were those from HHS regarding Medicare processing changes each year. However, a manual review of the documents identified the terms were most often used in the rules when responding to public comments, and the responses noted that HSS was not employing ML techniques in their fee payments, nor were they regulating medical technologies using ML as a contingency regarding Medicare coverage or payments.

To be clear, this federal register document analysis is not necessarily evidence that the federal government is not using ML in programs in ways that could impact administrative burdens. However, it does mean that if federal agencies are using ML, they are not being specific about the use in the one place where they would most likely meet several of the

principles for responsible AI as required by E.O. 13960, especially the principle of transparency. Additionally, while there is not yet any legal case law making this explicit, it is likely that federal agencies will need to update their regulations via rulemaking when using data collections from individuals as part of an ML process for government programs simply to comply with the APA. As a researcher, the federal register is one location to keep monitoring for the federal use cases of ML in program administration, especially those that would impact the level of administrative burdens experienced. However, there are likely to be many use cases of ML technology that are being used by the federal government, which has not met the current criteria and interpretations of being posted in the federal register. As discussed in chapter 2, these use cases are most likely to be those where federal agencies have acquired ML technology and processes through contracting with private companies, in which the technology is considered proprietary, and it is not being used specifically in benefits or services adjudications. Researchers should keep an eye on any government agencies posting catalogs of internal ML applications, especially since this is required by E.O. 13960 Section 5 if OMB catches up with the requirements of the existing E.O.

Rule Titles by Agency where Document Contain Terms

Agency Names	Title	Type
Commerce Department	Securing the Information and Communications Technology and Services Supply Chain	2
	Setting and Adjusting Patent Fees During Fiscal Year 2020	2
Consumer Financial Protection Bureau	Taking and Importing Marine Mammals; Taking Marine Mammals Incidental to Southwest Fisheries Science Center Fisheries Research	1
	Policy on No-Action Letters	2
Education Department	Distance Education and Innovation	1
	Final Priorities, Requirements, and Definitions-Fund for the Improvement of Postsecondary Education-Open Textbooks Pilot-Program	1
Federal Communications Commission	Final Priority; National Institute on Disability and Rehabilitation Research-Disability and Rehabilitation Research Projects-Inclusive Cloud and Web Computing	1
	Earth Stations in Motion	1
Health and Human Services Department	Restoring Internet Freedom	2
	21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program	1
Homeland Security Department	Medicare and Medicaid Programs; CY 2020 Home Health Prospective Payment System Rate Update; Home Health Value-Based Purchasing Model; Home Health Quality Reporting Requirements; and Home Infusion Therapy Requirements	1
	Medicare and Medicaid Programs; Electronic Health Record Incentive Program-Stage 2	1
	Medicare and State Health Care Programs: Fraud and Abuse; Revisions to Safe Harbors Under the Anti-Kickback Statute, and Civil Monetary Penalty Rules Regarding Beneficiary Inducements	1
	Medicare Program; CY 2021 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment Policies; Medicare Shared Savings Program Requirements; Medicaid Promoting Interoperability Program Requirements	1
	Medicare Program; CY 2022 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment Policies; Medicare Shared Savings Program Requirements; Provider Enrollment Regulation Updates; and Provider an...	3
	Medicare Program; End-Stage Renal Disease Prospective Payment System; Payment for Renal Dialysis Services Furnished to Individuals With Acute Kidney Injury, End-Stage Renal Disease Quality Incentive Program, and End-Stage Rena...	2
	Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Final Policy Changes and Fiscal Year 2021 Rates; Quality Reporting and M...	1
	Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Fiscal Year 2015 Rates; Quality Reporting Requirements for Specific Provi...	1
	Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Fiscal Year 2019 Rates; Quality Reporting Requirements	2
	Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Policy Changes and Fiscal Year 2020 Rates; Quality Reporting Requirements	2
	Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Policy Changes and Fiscal Year 2022 Rates; Quality Programs and Medicar...	3
	Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Policy Changes and Fiscal Year 2020 and Updates to the IRF Quality Reporting Program	1
	Medicare Program; Prospective Payment System and Consolidated Billing for Skilled Nursing Facilities; Updates to the Quality Reporting Program and Value-Based Purchasing Program for Federal Fiscal Year 2020	1
	Medicare Program; Revisions to Payment Policies Under the Physician Fee Schedule and Other Revisions to Part B for CY 2018; Medicare Shared Savings Program Requirements; and Medicare Diabetes Prevention Program	1
	Medicare Program; Revisions to Payment Policies Under the Physician Fee Schedule and Other Revisions to Part B for CY 2019; Medicare Shared Savings Program Requirements; Quality Payment Program; Medicaid Promoting Interopera...	1
	Medicare Program; Changes to Hospital Outpatient Prospective Payment and Ambulatory Surgical Center Payment Systems and Quality Reporting Programs; Revisions of Organ Procurement Organizations Conditions of Coverage; Prior...	1
	Medicare Program; Hospital Outpatient Prospective Payment and Ambulatory Surgical Center Payment Systems and Quality Reporting Programs; New Categories for Hospital Outpatient Department Prior Authorization Process; Clinica...	1
	Medicare Program; Hospital Outpatient Prospective Payment and Ambulatory Surgical Center Payment Systems and Quality Reporting Programs; Price Transparency of Hospital Standard Charges; Radiation Oncology Model	3
	Regulatory Clean Up Initiative	2
	Securities and Exchange Commission	Safety and Effectiveness of Consumer Antiseptic Rubs; Topical Antimicrobial Drug Products for Over-the-Counter Human Use
Safety and Effectiveness of Consumer Antiseptics; Topical Antimicrobial Drug Products for Over-the-Counter Human Use		1
Safety and Effectiveness of Health Care Antiseptics; Topical Antimicrobial Drug Products for Over-the-Counter Human Use		1
Transportation Department	Securing Updated and Necessary Statutory Evaluations Timely	2
	Cybersecurity Talent Management System	2
Treasury Department	Modification of Registration Requirement for Petitioners Seeking To File Cap-Subject H-1B Petitions	1
	Privacy Act of 1974: Implementation of Exemptions; U.S. Department of Homeland Security/U.S. Immigration and Custom Enforcement-018 Analytical Records System of Records	2
Homeland Security Department	Retention of EB-1, EB-2, and EB-3 Immigrant Workers and Program Improvements Affecting High-Skilled Nonimmigrant Workers	1
	Modernization of Regulation S-K Items 101, 103, and 105	1
Securities and Exchange Commission	Management's Discussion and Analysis; Selected Financial Data; and Supplementary Financial Information	1
	Revisions to Digital Flight Data Recorder Rules	1
Transportation Department	Streamlined Launch and Reentry License Requirements	1
	Exemptions to Suspicious Activity Report Requirements	3
Treasury Department	Provisions Pertaining to Certain Investments in the United States by Foreign Persons	1

Table 5-15: Rule Titles for Top Agencies (2010-2022)

Beyond the Federal register, a useful report published by the Administrative Conference of the United States, titled *Government by Algorithm: Artificial Intelligence in the Federal Administrative Agencies* in 2020 and provides seemingly the most comprehensive inventory and analyses of current Federal Government AI use cases (Engstrom et al., 2020). The Administrative Conference of the United States is an independent agency established in 1964 to promote “efficiency, adequacy, and fairness” in federal agency procedures to administer federal programs (*The Administrative Conference of the United States (ACUS)*, 2019). This report documented 157 use cases of AI across 64 federal agencies, however, it noted that most use cases are concentrated in a small number of agencies, and many were nascent, exploratory, or in development. Only 53 use cases were “fully deployed,” while the remainder were either in the planning stages or in the development/piloting stages. Interestingly, more than 80 use cases were being developed within the agency, while the remainder were developed by private entities or through a partnership with private organizations.

TABLE 2. TOP TEN AGENCIES AND SUBAGENCIES BY NUMBER OF USE CASES

Agency Name	Number of Use Cases
Office of Justice Programs	12
Securities and Exchange Commission	10
National Aeronautics and Space Administration	9
Food and Drug Administration	8
United States Geological Survey	8
United States Postal Service	8
Social Security Administration	7
United States Patent and Trademark Office	6
Bureau of Labor Statistics	5
Customs and Border Protection	4

Table 5-16: ACUS Table of Top AI Use Cases by Agency (Engstrom et al., 2020)

FIGURE 1. AI USE CASES BY POLICY AREA

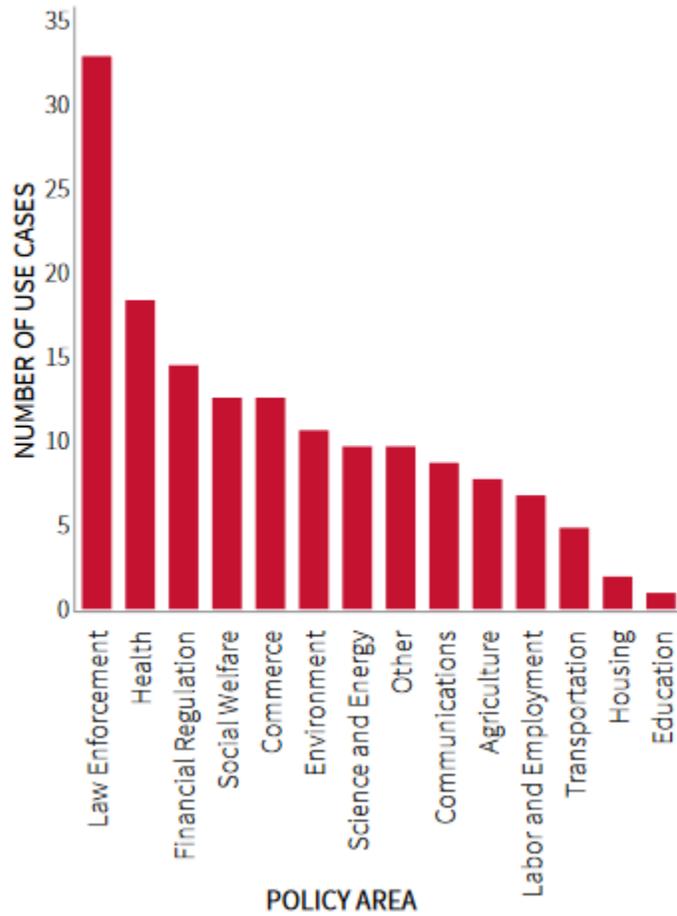


Figure 5-4: ACUS Graph of AI Use Case by Policy Areas

FIGURE 2. AI USE CASES BY GOVERNANCE TASK

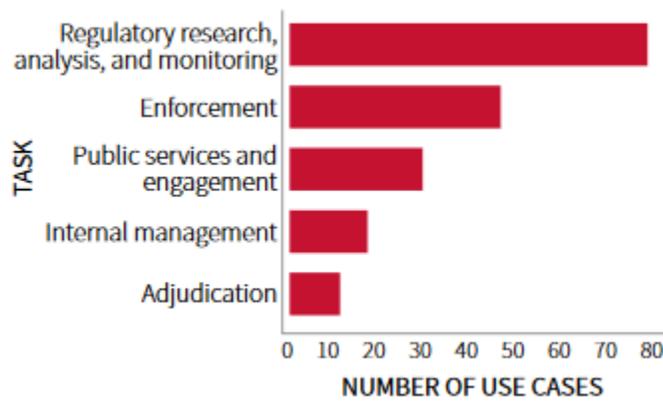


Figure 5-5: ACUS Graph of AI Use Case by Task

FIGURE 3. AI USE CASES BY IMPLEMENTATION STAGE

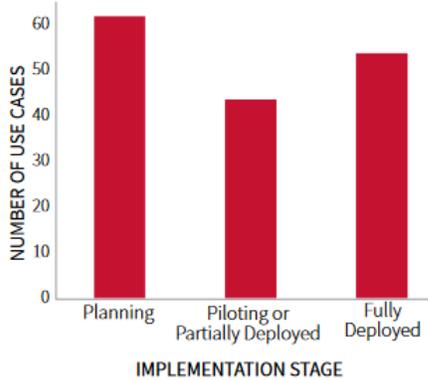


FIGURE 4. AI USE CASES BY DEVELOPER TYPE

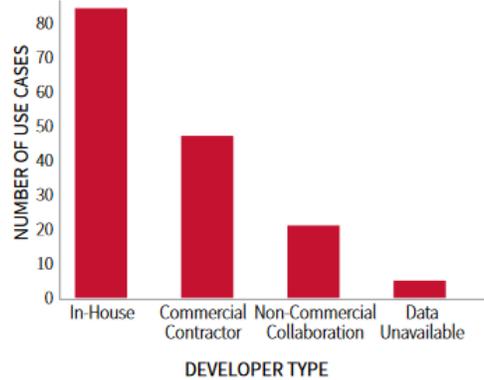


Figure 5-6: ACUS Graphs of AI Use Cases by Development Stages and Development Type

Additionally, the report noted the majority of the use cases were in law enforcement or regulatory instances, and only 12 use cases were in “adjudicatory” tasks, which might impact the programs in a manner that would affect the level of administrative burdens costs. Of note, within the 153 use cases, the majority of the AI/ML approaches were classification or regression or structured data, which is in line with my proposed ML solutions to the identified administrative burden cost solutions.

FIGURE 5. AI USE CASES BY MACHINE LEARNING METHOD

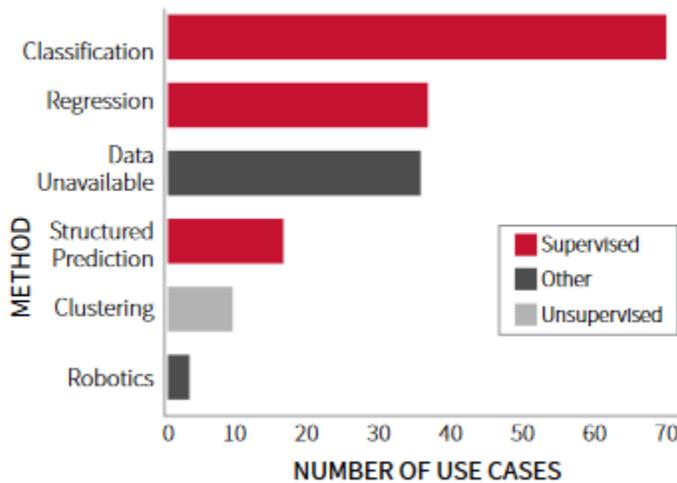


Figure 5-7: ACUS Graph of AI Use Cases by ML Method

FIGURE 6. AI USE CASES BY DATA TYPE

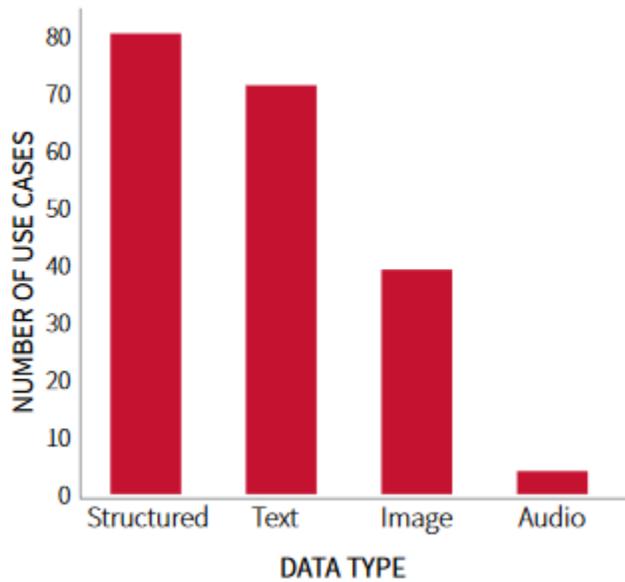


Figure 5-8: ACUS Graph of AI Use Cases by Data Type

Despite these findings and the additional use cases analyses in the report, significant conclusions were that there was insufficient publicly available information to make a detailed examination of more than 60% of the use cases. Additionally, the report cited the lack of linkage between AI uses in these agencies and administrative procedures such as public notice and regulatory guidance in the Federal Register (Engstrom et al., 2020). Next, I coded the report’s use cases based on the four types of government and individual interactions as defined in Chapter 1 to identify those that may impact the levels of administrative burdens in a federal program, as shown in Table 5-17: AI Use Cases Coded by Interaction Types. Of the 19 use cases detailed in the report, only two were associated with potential changes to administrative burden levels. These were ML systems to help the SSA adjudicate SSDI claims.

Interaction Type	Count of Interaction Type
Type 1 - Red Tape/Green Tape	8
Type 2 - Administrative Burdens	2
Type 3 - Regulatory/Enforcement	8
Type 4 - Individuals and Private Entity	1
Total	19

Table 5-17: AI Use Cases Coded by Interaction Types

Agency	AI Use Case Name	AI Use Case Explanation	Interaction Type
--------	------------------	-------------------------	------------------

SSA	Clustering for Micro-Specialization	Case Clustering for Adjudication Expertise to increase accuracy and efficiency	Type 2 - Administrative Burdens
SSA	Appeal Acceleration	Predicting the likelihood of successful adjudication/appeal outcomes to facilitate processing.	Type 2 - Administrative Burdens

Table 5-18: ACUS AI Use Cases for Administrative Burdens

This is obviously a minority of the use cases focused on, none of which were mentioned in the federal register either in specific notices or in public notifications of regulation procedures, which enforces the earlier finding of the lack of federal government transparency of ML use in administrative procedures. There is also insufficient information about the two use cases in the federal government based on the sources for the report and the current publicly available information for the agencies to make an assessment of their compliance with E.O. 13960, MLOps best practices, or methodological approaches. Additionally, I am not actually certain, based on the current publicly available information, if these programs would result in lowered administrative burdens for applicants or whether their intended outcomes were more focused on internal SSA processing frictions (which would be reduced “red tape” rather than administrative burdens) but I am assuming these efficiency and accuracy gains would at least indirectly lower administrative burdens on applicants.

Framework 2 Summary

Framework 2 shows a clear path toward using ML in the federal government to reduce different types of administrative burdens and costs in government programs. I have shown how to map the types of ML solutions and approaches based on the costs and the measurements of those costs to form the basis of the ML solutions design and development. I have also shown how to leverage the MLOps design, development, and deployment steps as adapted to the public sector to ensure compliance with legal, regulatory, ethical, and scientific requirements for ML use in the public sector. In the next section, I will explore the using the requirements for evidence-based policymaking and evaluations to measure the impacts of the ML models on administrative burdens cost levels and the overall impacts of reducing administrative burdens costs on the outcomes and impacts of the federal government programs.

Chapter 6 - Framework 3 Results - Evidence-based Policymaking: Evaluating the Outcomes of Machine Learning to Reduce or Eliminate Administrative Burdens

Overview

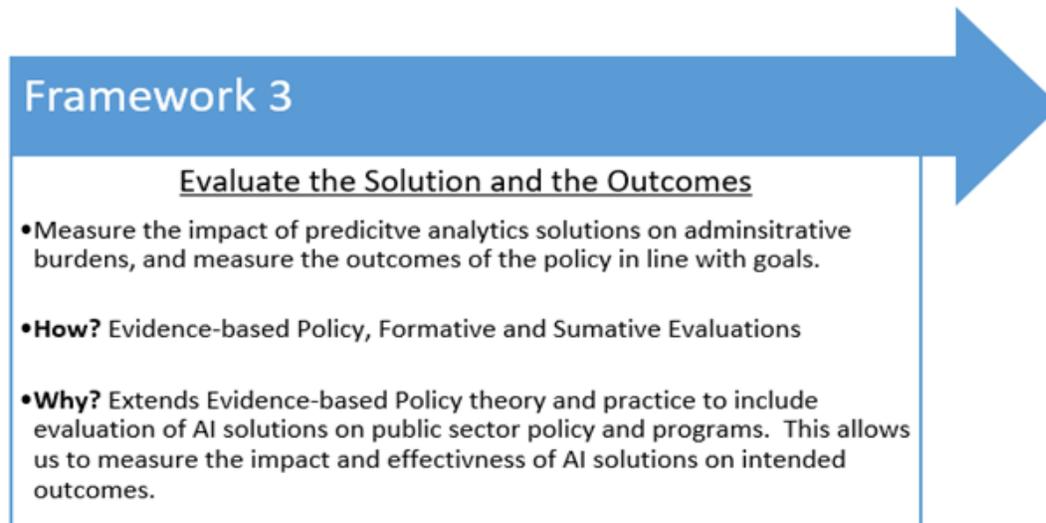


Figure 6-1: Framework 3

Similar to the increased requirements and belief in performance management in the federal government, there have been increasing requirements for the use of evaluation and scientific methods to determine which policies and programs can withstand rigorous evaluations and provide evidence they are likely to be effective. This has culminated in several requirements for the federal government, many of which were codified in the Foundations for Evidence in the Federal Government Act (Evidence Act) in 2018. Just as I approached the requirements of performance management in Framework 1, I will look to these requirements for the federal government to establish ways to evaluate the effectiveness of ML programs to lower administrative burdens and their impacts on the outcomes and goals of the programs. In Framework 2, I discussed the importance of evaluating ML models and programs to ensure they are accurate and valid, so Framework 3 will focus on the impact on the goals and outcomes of the levels of administrative burdens as well as on the overall goal of the program. This section will not seek to propose changes to the evidence-based policymaking requirements in the government or those for evaluation. Instead, it will understand the current requirements and show how they can be applied to our goals of Frameworks 1 and 2 to help us even better

understand the changes to the program and better understand what can be further implemented to reach our goals.

The benefits of this approach are several. First, I have already discussed how performance measurement and management tools and data can be useful for monitoring processes in line with meeting strategic goals, but also how performance data can be used to perform more in-depth evaluations. Secondly, I have also discussed the benefits of both performance measurement and evaluation plans for programs – providing valuable information to policymakers and policy administrators. Lastly, just as in Framework 1, but leveraging existing requirements, I am not creating new, unfunded, and unpracticed requirements for researchers and administrators. Instead, this framework will show how to use the existing requirements to help the overarching goals of reducing administrative burdens in programs through ML techniques. Further to these goals, it also helps researchers and administrators understand how to measure the impacts on the overall goals of the policy or program to see what the impacts are from the reduced administrative burdens.

Results: An evaluation framework for machine learning solutions for administrative burdens

In this section, I will walk through the requirements for evaluation and evidence-based policy as applied to ML solutions for reducing administrative burdens that were dealt with in Frameworks 1 and 2. Specifically, this means that I will develop and design an example evaluation framework to be applied to the ML solutions of high administrative burdens to show how administrators and researchers can evaluate and collect evidence on the impact and outcomes of Frameworks 1 and 2.

In Figure 6-2, I develop a theory of change model for the solutions this work proposes. This model is based on the guidance and requirements for an evidence-based policy as detailed in chapter 2, which are the current requirements for federal agencies. The following sections will outline an example evaluation framework to help administrators and researchers collect evidence about these solutions and bring additional evidence to an implementation using Frameworks 1 and 2.

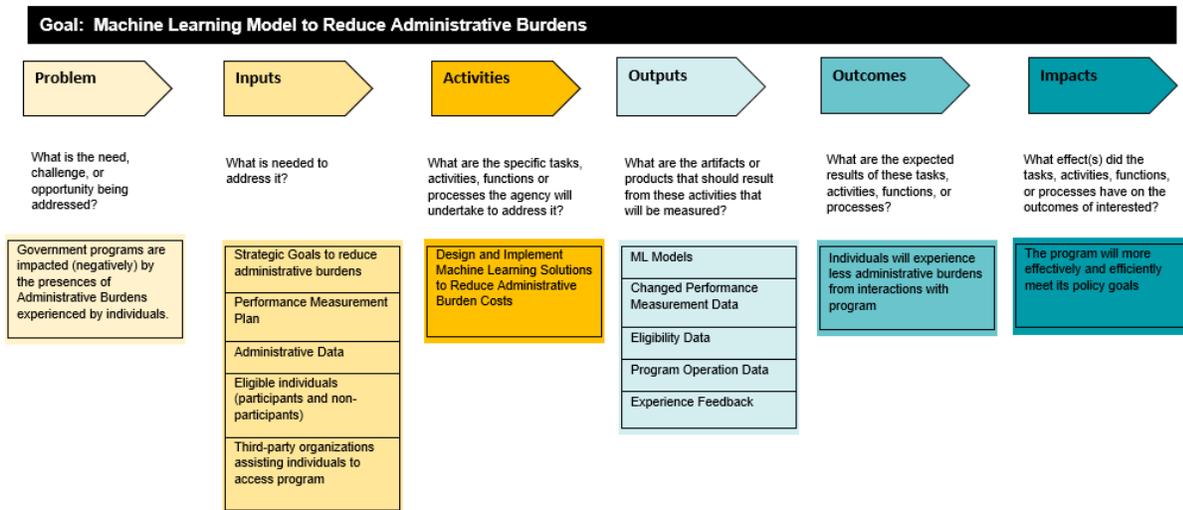


Figure 6-2: Theory of Change for ML methods to Reduce Administrative Burdens

Evaluation Questions to Be Answered

- How effectively does the agency implement the program for eligible individuals?
- What are the levels of administrative burdens, and can they be reduced with the implementation of ML solutions?
- How does the ML solution impact the levels of administrative burden costs according to the performance measurement data?
- How does the reduction of administrative burdens affect the outcomes and impacts of the program overall?

Lead Office, Points of Contact

The agency evaluation team needs to be working closely with the agency data team and the program office, as well as the team designing and implementing the ML model within the program if not contained in the program office. Key stakeholders will include the evidence-building team, the Department leadership responsible for the APG and for program oversight, as well as CDOs, evaluation Officials, Statistical Officials, and Official for Responsible AI, as this will have a broad requirement within the agency organizational structure.

The Rationale for the Evaluation

Addresses key questions in the learning agenda associated with the Agency Priority Goal in more effectively implementing this program.

Purpose of the Evaluation

Provides insights into the effects of ML solutions on levels of administrative burdens in the program and the changes to program outcomes and goals based on reduced administrative burdens.

Audience

The audience for this evaluation is agency leadership associated with the APG and program, program staff, and the public.

Outcome Evaluation

The outcome evaluation will focus on the design and implementation of the ML solution within the programs to determine if it is reducing administrative burdens as designed.

Methods and Design

Methods of measurement will include:

- Statistical analysis of administrative program data before and after the implementation of the ML model to measure administrative burdens over time.
- Feedback from individuals based on survey responses before and after the implementation of the ML solutions.
- A/B testing of the implementation of ML solutions to different parts of the program process to provide treatment and control groups for statistical analysis.

Data Needed for the Evaluation

Performance measurement data, outcome data.

Anticipated Challenges

Challenges will include political concerns about the impacts of ML solutions on subsets of the program participants and applicants, and the agency will be under pressure to quickly implement any solutions which appear to be effective to the population to minimize negative impacts on individuals who are not yet experiencing the reduced administrative burdens provided by the ML solutions.

Summative/Impact Evaluation

The summative or impact evaluation will compare outcomes with the reduced administrative burdens to outcomes without the program, as well as outcomes without the reduced administrative burdens and those without the program to determine the impact of the

overall program on individuals who participate and how those impacts differ with reduced administrative burdens.

Methods and Design

Pilot programs of the ML solutions will allow for the collection of data from treatment and control groups as well as comparisons to individuals otherwise eligible who do not apply for or participate in the program. This will allow the evaluation team to perform statistical analyses of the causality of the level of administrative burdens and on the outcomes experienced by individuals as compared to those with lower experienced administrative burdens and those who are eligible but do not participate in the program. Pilot participants will be randomly assigned to pilot treatments through the application of our ML solutions and a randomness generator for this pilot treatment when interacting with potential program applicants and participants.

Data Needed for the Evaluation

- Administrative data for all program applicants and participants, disaggregated by pilot participation and not.
- Administrative data for individuals who are eligible but do not apply for or participate in the program.
- Survey data from individuals about their experiences with the program.
- Outcomes of individuals participating in the program (based on the policy goals of the program).

Anticipated Challenges

Challenges for this evaluation are going to include risks of the impacts of selection bias for pilot participants as compared to non-pilot participants of the program, as well as selection biases, history impact, and other internal validity considerations because we will not be able to control variables associated with our control group of individuals who do not apply for or participate in the program.

Dissemination Plan

We will disseminate the evaluation data and the evaluation conclusions and recommendations to program staff, agency leadership, and the public through our website and publication on www.evaluation.gov.

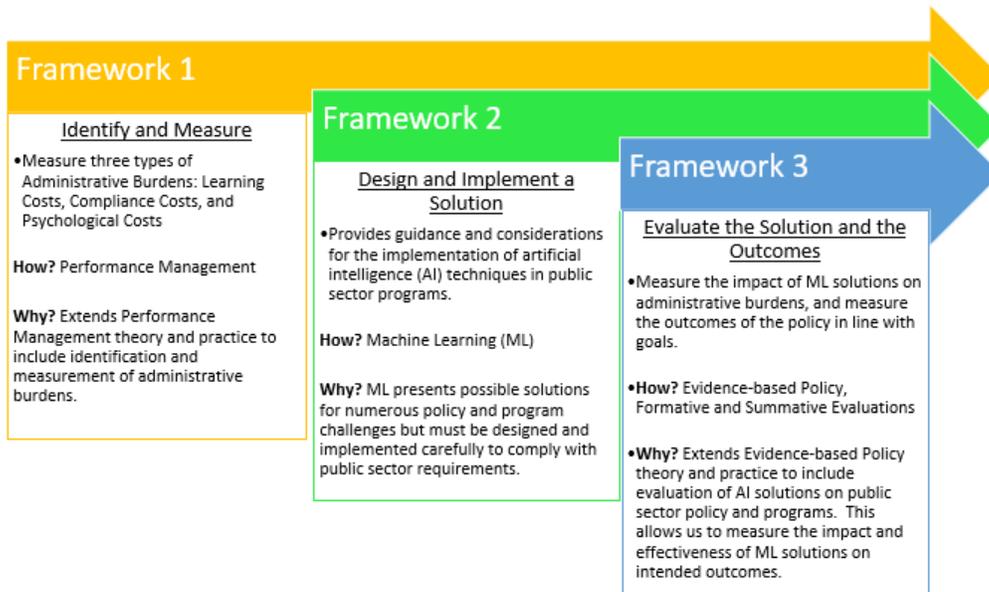
Framework 3 Summary

Framework 3 allows researchers and administrators to use the requirements of evidence-based policymaking and evaluation within the federal government that already need to be complied with to understand the impacts of their ML solutions to administrative burdens as well as the overall program impacts based on the reduction of administrative burdens. As discussed before, these are important goals because administrators and researchers need to understand how our ML solutions are performing in regard to the intended goals they were designed for. Additionally, I want to conduct valid causal analyses to understand how the reduction of administrative burdens in a program impacts the overall goals of the program. For example, with a program like Section 8 housing vouchers, does the ML solution decrease the experienced administrative burdens for applicants and participants? And if so, does the overall reduction of administrative burden cause an increased impact on the program as far as fewer individuals who are unhomed or cannot afford housing in our treatment group?

Additionally, Framework 3 can easily be changed to evaluate the design and implementation of ML solutions for purposes other than reduced administrative burdens, which is useful because the government is going to see the proliferation of ML solutions for many types of goals and program outcomes beyond just administrative burdens. Additionally, there as I have discussed in this paper, there are other potential solutions to reduce administrative burdens in the program beyond ML. This framework can also be modified to evaluate the impact on non-ML solutions to reduce administrative burdens in programs. What is important is that government agencies are not simply making changes to programs without evaluating the outcomes of those changes (formative and outcome evaluations) and the impacts on the overall program based on changes (Summative or impact evaluations).

Chapter 7 - Conclusions and Discussion

Overview of the Resulting Frameworks



There is a lot to these three frameworks. However, they are grounded in current requirements and guidance for the federal government. Therefore, these are tasks that government administrators are already complying with for programs and policies. Additionally, researchers are familiar with them and doing work to help us understand what is useful for successful performance management and evidence-based policy in the federal government. The frameworks also help tie together performance management and evidence-based policy in a concrete way, as espoused in the Evidence Act and resulting guidance.

As a reminder, Framework 1 guides the identification and measurement of administrative burdens in line with strategic goals to reduce them. Reduction of administrative burdens is important because of the potential positive impact on program take-up rates, the effectiveness of policy goals, and more equitable program administration, in addition to the other potential benefits outlined in chapter 2. Framework 2 provides administrators with a clear understanding of ML solutions to reduce administrative burdens as well as how to implement those solutions using MLOps to help address public sector ML considerations and requirements. Finally, Framework 3 builds on requirements from the Evidence Act to provide administrators with a method to evaluate the outcomes and impacts of the ML solutions and the reduced administrative burdens in programs.

The frameworks add additional benefits and tools to these important research areas as well. Framework 1 provides administrators and researchers with methods to identify and categorize administrative burden measurement processes and categories of measurement. This will further the academic work on administrative burdens because it will help build up the corpus of case examples and add comparable definitions and measurements that facilitate larger empirical and quantitative studies of administrative burdens. Framework 1 also provides additional guidance and examples about how to develop and structure performance management and measurement processes which can be applied even beyond the study of administrative burdens.

Framework 2 contributes to the administrative burdens research because it systematizes the analysis of burden causes and the design and implementation of potential solutions. It also provides useful guardrails to the continued implementation focus of ML in the public sector by grounding the current requirements and considerations in a realistic use case. Framework 2 can be extended to develop ML solutions for other program and policy considerations beyond administrative burdens. As discussed, this is becoming more and more common as the government makes headway on overcoming obstacles to ML adoption in order to successfully realize the potential benefits.

Framework 3 is not novel in terms of evidence-based policy or evaluation, but it does provide new insights and direction for administrators and researchers on the evaluation of ML in the public sector, as well as the impacts or reduced administrative burdens on program outcomes and goals. This helps bring important areas of research and focus together and again provides very practical methods for federal agencies to leverage existing requirements they have resourced to focus on these challenging and important areas.

Implementation of the Frameworks

Practically speaking, I believe that most agencies will need to take the lead in implementing these frameworks since I've already shown that the needed data isn't available to the public for researchers to be able to implement a measurement scheme for administrative burden costs. However, with some additional work, potentially seeking a public/private partnership with a federal agency or program office, request for additional data (potentially through a Freedom of Information Act request if the data does exist), or creating mechanisms to gather data directly from participants or other means available to the public could establish performance measurement processes absent the federal government taking the lead. More

realistically, though, government agencies should take the lead. Perhaps as a response to E.O. 14058 on Customer Experience, or perhaps the current administration or Congress will eventually focus on the importance of understanding administrative burdens for programs.

Whatever the path forward, I believe these frameworks show how existing requirements, which have already been resourced, can be used to focus on identifying and measuring administrative burdens, implementing ML solutions to reduce them, and evaluating the impacts of the ML solutions as well as the impacts to the program based on lower levels of burdens. As I have discussed, there are several organizations that are also working on wrap-around applications or services that might also be able to help provide paths forward for the government, such as Code for America and their many state and local chapters, as well as DataKind, and other nonprofits organizations in the “CivTech” or GovTech” space. Additionally, there are many organizations that are part of the federal government service-providing process as grantees who could implement much of this with their service delivery model or process.

Potential Limitations of the Frameworks

These frameworks are obviously limited by the fact that I was unable to apply actual performance measurement data to Framework 1, I could not find many publicly available ML solutions to compare Framework 2, and again there aren't any current applications of evaluations of ML programs or on the reduction of administrative burdens on program outcomes. Additionally, because I grounded my framework in existing federal government requirements, these frameworks will become out of date and need edits when any of these requirements change. Specifically, since nearly every administration changed some details of the performance management process within the existing legal requirements, it is likely that Framework 1 will need to be updated soon. While there has been a noticeable absence of guidance, clarifications, and rules from OMB regarding the requirements of E.O. 13960, any new changes (including rescinding or superseding the E.O.) would require substantial updates as the federal government continues to adopt ML guidance and requirements. Framework 2 also needs to stay compliant and up to date with any new court rulings and legislation regarding federal government use of ML techniques, as well as any international laws if the federal services would be affected. . However, I believe this is easily achieved through the framework provided.

Framework 3 is also limited by the fact that there are no existing evaluations on the impact of administrative burdens on program outcomes and goals specifically. I do believe this is an area that the federal government and academic researchers can improve upon in short order.

As discussed throughout this work, there are academic studies linking the impacts of programs to the antecedent causes of administrative burdens, but a specific, focused evaluation or research would help validate the existing work and theory of administrative burdens as well as provide additional fuel for focus within the academic community and the federal government.

What I Didn't Find

In addition to the above limitations, it is important to call out the importance of the null findings in this research, as they point to needed improvements in data and understanding for administrators, researchers, and citizens. Namely, the lack of publicly available data from Federal agencies allows us to understand the presence and magnitude of administrative burdens. As the United States is moving closer to meaningful implementation of the Open Data Act, we want to see that publicly available data is used in meaningful ways to research, understand, and interact with our government. Therefore, even data that isn't specifically created and reported for one task, like measuring administrative burdens, should be published in ways that allow us to interrogate other phenomena. This includes understanding the provenance of data and its limitations, being able to search through metadata to combine performance measurement data from across program areas and different processes to create a larger understanding of time series and complete programs, and creating new value from existing data without significantly increasing the level of effort of agencies.

Secondly, the lack of public information about government implementations of ML is both troubling because it doesn't allow us to understand and study the public sector's use of ML as applied to government programs and policy, as well as the potential impact on individuals. Additionally, we are blocked from understanding the impact of ML on overall program goals and outcomes. Significantly, this may be because there are very few ML implementations at this time, but more likely, it is because there are no enforced clear requirements and mechanisms for public reporting of federal government ML. Ideally, E.O. 13960 will eventually result in a much better understanding and ability to interact with public sector ML design and implementation. But additionally, in line with the requirements of Responsible ML and Administrative Procedures, I hope we will begin to see meaningful public notice of how ML programs are designed and deployed as part of federal register notices, rules, and regulations which situate these explanations as part of the overall system and process to help the public understand their full implementation and changed to process and adjudication.

And finally, evidence-based policy requirements should ideally result in a similar release of the underlying evaluation and scientific trial data. Just as we are concerned with reproducibility in academic research, we should seek out and create the mechanisms to have and facilitate the reproducibility of government evidence-building and scientific research that informed policymaking. These were some of the principles of the final report from the Commission on Evidence-based Policymaking, which didn't completely translate into the Evidence Act as signed into law. And as there are increased techniques and technologies which allow us to interact with data without necessarily having access to the sensitive elements of the data, hopefully, this will allow the government to open up access to the trove of data available so that the public and researchers can add to the body of knowledge and understanding as well as the fuel the evidence-building that the government is currently mostly undertaking on their own.

Next Steps and Future Research Agenda

There are a lot of places to go from here to extend this research. One area that needs attention is to create new technology to collect publicly available data to measure administrative burdens in programs. This should be done to help increase transparency and understanding of administrative burdens. As my research found a few data sets which might be leveraged, I believe that the government is headed in the direction of additional publication of usable data. This means that concerted work to collect, interpret, and analyze this data can help drive additional research and insight on the presence of and impact of administrative burdens. This technology can easily leverage data and information that the federal government makes available regarding processing time, steps, take-up, and information about program recipients that helps the agency and researchers measure and understand the impacts of administrative burdens. This can be done in ways that still protect and adhere to rules about privacy and data sharing.

Another approach is to build tools that can be used by the government and organizations to collect the needed data themselves. Even aggregating data or samples of data can provide significant insights into administrative burdens. For example, we could train ML models to predict the presence of administrative burdens based on labeled data derived from my definitions to help automate the collection of examples for further research. More simply, the academic community can also begin to discuss administrative burdens and the measurement of them with a shared lexicon and definitions to help build existing research data and corpora, which can fuel this important topic area. This is an area that researchers should support and work on because it will bring closer the use of research and government data to understand and inform public policy,

which is a cornerstone of evidence-based policymaking. This can be done by increasing research focuses on the adaptation of government data sets as well as increased development and release of useful data from the federal government. Hopefully, this will increase based on further adoption of Open Data Act requirements but can also be done through further government-academic partnerships or through advocacy with administrators and lawmakers.

As agencies gain more experience in identifying and measuring administrative burdens, especially in regard to program evaluations, this will become more feasible, and agencies should have more information to allow them to more easily keep track of the majority of administrative burden causes and measurements within their programs. However, for the initial APG milestones, I recommend that agencies review their programs and processes to identify key components of each administrative burden costs that exist in their programs and develop baselines and milestones to reduce based on these priority components. Additionally, by identifying these priority areas, agency leadership will also identify the areas that are likely most ripe for reduction within the twenty-four-month period.

While working on Framework 2, I became convinced that if they do not exist already, there will soon be more examples of applied ML solutions that can impact administrative burdens. A simple tool would allow researchers to continually probe federal register documents using NLP to identify new use cases of applied ML, as well as rule and process changes likely to impact the presence of administrative burdens in programs. This tool should be developed by researchers and nonprofit partnerships through legal and open source means to help provide more accurate and timely information. Additionally, the federal government should be held accountable to comply with the requirements of E.O. 13960 to create and publish inventories of existing AI use cases. Further, OMB should promulgate additional guidance on how these inventories should be published to facilitate aggregation and research methods. The guidance should also ensure that these inventories are kept up to date, include internally developed and procured AI use cases, and provide sufficient details to help the public and researchers understand how the ML solutions are designed, developed, and comply with the principles of Responsible AI.

Once designed and developed, this tool would provide a low costs way to ensure research in this area is continued with little resources or overhead. Additionally, future research can expand on Framework 2 to include more detailed and granular information to inform the design, development, and deployment of ML models in compliance with federal government requirements and considerations. These additional details can include specific research focused

on critical model metrics for public sector implementations based on the potential impacts of the model uses. It can also focus on the legalistic interpretation of current law and policy in this emerging area to ensure the federal government has the guidance and academic results needed to ensure the appropriate use of ML applications in the government.

Overall, the cross-section of administrative burdens and ML in the federal government programs have many opportunities for continued research that will have positive impacts on how government services are designed and delivered. Additionally, the opportunity is ripe to make the phenomena of administrative burdens a transparent area of focus to ensure that policy is not made without the knowledge and consent of the public and so that the field can better understand the impacts of these costs and the decisions which lead to the experiences of the costs. There are many hypotheses of the impacts of administrative burdens on policy, programs, and individuals, which the research has already begun to explore through particular use cases. However, the government should expand the publicly available information about programs and administrative burdens to help researchers understand the correlation and causations of administrative burdens throughout the public sector. These priorities align with public sector goals which are lauded by practically every political ideology and values-based goals, such as efficiency, equity, evidence-based, fiscal transparency, individualism and oversight, and transparency. Therefore, I do not believe the focus on administrative burdens can be argued to favor one political ideology over another. Instead, I believe it is more likely to be a tool used by serious individuals and organizations interested in how the United States governs.

Overall, the path forward for myself and other researchers involves a continued focus on administrative burdens in federal programs as well as solutions to them. There are tools that can be built to assist these efforts as detailed above, but the consistency in research focus and discussions will help bring this topic into clearer focus for administrators and legislators. Similar to the trajectories of both performance management and evidence-based policymaking, administrative burdens can be seen as an area that already links to important policy priorities of government programs and government processes, as can be seen by the extensions I have drawn in this work. Therefore, it is not unlikely to see legislative interest in a focus on administrative burdens, especially in bringing them into existing processes and as a means to achieve more efficient, more effective, and increased oversight of current tax expenditures in the government.

Appendix A – Raw Data

Type of Cost	Diagnostic Question	Potential Cause and Measurement	Refined Measurement	Explanation	Simplified Measurement	Source	Potential ML Solution	Simplified ML Solution	Type of ML	Subtype of ML
Compliance Costs	How many forms must applicants complete?	All forms and supporting documentation requirements.	Amount of application and compliance forms	A number of forms, length of forms.	Form Measurement	Herd & Moynihan, 2018	Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.	Autocompletion	Supervised Learning	Classification
Compliance Costs	How many questions are on the forms?	PRA estimates for forms.	Time spent on forms	What are the average amounts of time that individuals spend completing the forms?	Form Measurement	Herd & Moynihan, 2018	Reduction of data needed based on a minimum amount of data to generate threshold accurate adjudication predictions/decisions.	Autoadjudication	Unsupervised Learning	Dimensionality Reduction
Compliance Costs	Do individuals have to input data multiple times?	Review of data collected more than once.	Duplicity in information collections	Is information collected more than once? If so, how many times?	Form Measurement	Herd & Moynihan, 2018	Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.	Autocompletion	Supervised Learning	Classification
Compliance Costs	How much documentation is required?	Supporting documentation and validation requirements.	Amount of forms, time spent on forms	What are the amounts of supporting documentation or verification required? How much time do these take?	Form Measurement	Herd & Moynihan, 2018	Reduction of data needed based on a minimum amount of data to generate threshold accurate adjudication predictions/decisions.	Autoadjudication	Unsupervised Learning	Dimensionality Reduction
Compliance Costs	How do individuals submit the forms (e.g., in-person, by mail, online)?	Scores for less onerous processes.	Variety of application processes and measurement of ease of use	Are there different application processes for individuals to choose based on their levels of comfort?	Application Measurement	Herd & Moynihan, 2018	Predicting optimal placement of needed application support centers based on historical data, predictions of how individuals will wish to apply (online, by mail, via phone, in-person)	Location Optimization	Supervised Learning	Classification
Compliance Costs	Is information collected from individuals available through government administrative data?	A review of existing information collection requirements against what is available through government data sharing.	The amount of information collected from applicants was available through administrative data sources	Is information collected from individuals available elsewhere as government administrative data?	Application Measurement	Herd & Moynihan, 2018	Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.	Autocompletion	Supervised Learning	Classification
Compliance Costs	Do individuals require an interview, consultation, etc.? If so, is that available in person, by phone, or online?	Scores for validation, consultations, or interview processes are based on the results of those processes in terms of information collection.	Amount of forms, time spent on forms	In addition to forms, what is the amount of time spent on other parts of the application, verification, or certification process (if applicable)? This could	Form Measurement	Herd & Moynihan, 2018	Prediction for adjudication of eligibility determinations based on submitted and administrative data to minimize those who require any additional verification or	Autoadjudication	Supervised Learning	Classification

				include interviews, picking up vouchers, check-ins, and other compliance requirements (drug tests, employment searches, etc.)			information submission.			
Compliance Costs	Do applicants have help in completing the application processes?	Scores for resources to assist individuals, including varying levels of assistance provided based on individual needs.	Measurement of the level of assistance available, disaggregated by individual needs	Third-party helps with applications.	Assistance Measurement	Herd & Moynihan, 2018	Prediction of individuals who will require additional assistance based on submitted information for targeted, proactive outreach. Success is measured by the proportion of the additional non-predicted population that requests additional assistance.	Assistance Optimization	Supervised Learning	Classification
Compliance Costs	How frequent is re-enrollment?	The multiplier effect of costs is based on the frequency of re-enrollment or verification requirements.	How often do individuals need to re-apply, and what proportion of initial requirements must be resubmitted	How often must individuals re-apply, certify, etc.?	Application Measurement	Herd & Moynihan, 2018	Automatic re-enrollment based on predicted eligibility determination.	Autoenrollment	Unsupervised Learning	Clustering
Compliance Costs	How much time do individuals commit to the process?	PRA estimates for forms but also other procedural requirements.	Time spent on forms	PRA estimates + all other application and requirement time estimates.	Form Measurement	Herd & Moynihan, 2018	Autocompletion of forms based on the identification of individuals' data from administrative data, prior forms, accounts, and open sources of information which is then presented for correction or validation before completion.	Autocompletion	Supervised Learning	Classification
Compliance Costs	How much does an individual's inquiry for the application process cost them?	Distributive costs are based on requirements, allowable formats, and missed opportunity costs associated with missed employment time.	Amount of time spent on the application process	Based on total application time, what are the opportunity costs? Time spent on benefits x average value of their time per government numbers.	Application Measurement	Herd & Moynihan, 2018	Automatic enrollment and proactive request for validation of information/eligibility status based on administrative data.	Autoenrollment	Supervised Learning	Classification
Compliance Costs	Costs of Private healthcare when government services are not trusted	Preferred private options are too expensive to participate	Comparison of government and non-government options	Comparison of costs between government service and private services?	Process Measurement	Ali & Altaf, 2021	Increased trust in the effectiveness of government services (auxiliary effect).	N/A	N/A	N/A
Compliance Costs	Take-up rate of a program based on compliance costs	Compliance costs can impact the take-up rate of a program	Program take-up rate	Take-up rate based on the amount of compliance costs experience	Program Measurement	Bhargava and Manoli, 2015	Automatic enrollment of predicted eligible individuals.	Autoenrollment	Supervised Learning	Classification
Compliance Costs	What is the level of "administrative exclusion"?	Does the level of caseworker/adjudicator discretion result in more or less program uptake?	Program take-up rate compared to the evaluation of program staff	Formal and informal rules within public sector organizations must be navigated, complied with, and overcome to receive services or benefits.	Program Measurement	Brodin and Majumdar, 2010	Automation of eligibility determinations at multiple steps in the program to increase the accuracy of eligibility determinations and reduce administrative discretion in a majority of cases.	Autoadjudication	Supervised Learning	Classification

Compliance Costs	What is the ratio of burden or costs applied to the government agency versus the individual?	Shifting the burden or cost to the agency will reduce them from the individuals.	Program requirements for individuals versus organization	Perception of current burden and ideal burden placement based on survey responses;	Process Measurement	Burden et al., 2012	Increased automated processes developed by the government reduce necessary inputs and participation from individuals.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	What are the causal mechanisms of burdens within a specific program?	"Causal Process tracing" of a program allows a within-program case study association of causes of burdens.	Analysis of Survey responses	Statistical analyses of responses to identify causality	Feedback Measurement	Camillo, 2020	The automated root-cause analysis process increases the accuracy of causal process tracing.	Causal Inference	Unsupervised Learning	Clustering
Compliance Costs	How does the required task or process differ based on the executive functioning of individuals?	First-time applicant vs. repeat applicants (more or less administrative experience); health and executive functioning; scarcity will decrease human capital	Program compliance data disaggregated by participant executive function variables	Executive functioning (as controlled by an individual's health, level of scarcity, and cognitive decline) can impact the perceived experience with a requirement or action and, therefore, will impact the level of the cost of administrative burdens.	Program Measurement	Christensen et al., 2020	Proactive, differentiated outreach based on predictions of executive functioning levels.	Assistance Optimization	Supervised Learning	Classification
Compliance Costs	Is non-take-up due to being aware by choosing not to apply?	(1) Non-knowledge. Do welfare clients and ex-clients know that the program exists? Do potential claimants understand the program's eligibility rules? What learning costs do they face? Do service providers actively propose and "sell" the program to potential claimants?	Survey of program participants and eligible non-participants	Structure interviews aimed at assessing root causes on non-take-up	Feedback Measurement	Daigneaule and Mace, 2020	Proactive program outreach of likely eligible individuals	Assistance Optimization	Supervised Learning	Classification
Compliance Costs	Are eligible people applying but not receiving the benefits due to organizational error or decisions?	(2) Non-demand. Does the program answer the needs of long-term welfare clients and clients (program relevance)? What are the psychological and compliance costs experienced by potential and actual participants? Is the claiming process sufficiently simple to encourage individuals to apply?	Rates of false negatives in program adjudication	Structure interviews aimed at assessing root causes on non-take-up	Process Measurement	Daigneaule and Mace, 2020	Automated program application and eligibility determinations.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	How long does the individual wait to speak to someone?	Wait-time for applicants (either in person at interviews or on the phone)	Wait times for phone and in-person conversations, email response times	Longer wait times are associated with higher costs and therefore reduce program participation	Assistance Measurement	Deshpande and Li, 2019	Chatbots are accessible via web, phone, or email that can accurately answer questions and maximize personal phone operators for more advanced issues.	Virtual Assistant	Supervised Learning	Classification
Compliance Costs	How far long do individuals need to travel to locations?	Where in-person events are required or used, and how do the locations associate with individuals who are applying (or are eligible)?	Measurement of distance from individuals to application/program center	Field office closings are associated with lower take-up unless there are other changes to offset.	Assistance Measurement	Deshpande and Li, 2019	More efficient application center location for individuals.	Location Optimization	Unsupervised Learning	Clustering
Compliance Costs	Digital Access	Increased digital access to application/compliance increases	Are digital application and compliance	rules that ease the cognitive burden associated with	Process Measurement	Fox et al. 2020	Automatic application completion through the use of	Autocompletion	Supervised Learning	Classification

		<i>program up-take and reduces administrative burdens</i>	<i>options available, variety, and use</i>	<i>enrollment, including receiving real-time eligibility decisions in less than 24 hours and a variety of changes in enrollment rules (including presumptive eligibility, express lane eligibility based on other cumulatively had a significant and substantive effect on enrollments. program determinations, and reduced wait times)</i>			<i>administrative data via entity resolution minimized data collection from individuals.</i>			
<i>Compliance Costs</i>	<i>Enrollment Ease</i>	<i>Factors that make enrollment easier decrease administrative burdens</i>	<i>Measure of application "accelerators" - which are processes that speed up or make burdensome the application process for individuals</i>	<i>Measuring program changes or options which decrease the burden on the application</i>	<i>Application Measurement</i>	<i>Fox et al. 2020</i>	<i>Increased application automation</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
<i>Compliance Costs</i>	<i>How easy is it to renew program eligibility?</i>	<i>Enrollment ease is associated with administrative burdens</i>	<i>Measures of the application process for renewals/extensions of the program</i>	<i>Measuring program changes or options that decrease burdens during re-enrollment</i>	<i>Application Measurement</i>	<i>Fox et al. 2020</i>	<i>Automatic re-enrollment</i>	<i>Autoenrollment</i>	<i>Supervised Learning</i>	<i>Classification</i>
<i>Compliance Costs</i>	<i>Why do eligible people exit the program?</i>	<i>Some may no longer be eligible, or they can not comply with compliance sots, or they are deemed too onerous to continue to comply</i>	<i>Measurement of program exits was still eligible</i>	<i>Random selection of program participants and those who exited the program to determine the cause of program exits</i>	<i>Program Measurement</i>	<i>Heinrich, 2016</i>	<i>Automated eligibility determination through administrative data and eligibility determinations, automated re-enrollment to decrease eligible non-take up rates.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
<i>Compliance Costs</i>	<i>What level of administrative discretion is there in program adjudication?</i>	<i>The scale of discretion of program adjudicators</i>	<i>Measurement of program administrator discretion</i>	<i>Increased discretion is associated with increased burdens</i>	<i>Process Measurement</i>	<i>Heinrich, 2018</i>	<i>Automated information collection and eligibility determination for most individuals, reduced discretionary decisions by program staff.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
<i>Compliance Costs</i>	<i>Time Series of compliance requirements can identify increased or decreased levels of compliance costs</i>	<i>Increased or decreased compliance costs caused changed impact levels</i>	<i>Measurements of application processes over time</i>	<i>How have the measures of compliance costs changed over time within the program? What are the causes of those changes?</i>	<i>Application Measurement</i>	<i>Heinrich and Brill, 2015</i>	<i>Automated application completion through administrative data and entity resolution.</i>	<i>Autocompletion</i>	<i>Supervised Learning</i>	<i>Classification</i>
<i>Compliance Costs</i>	<i>Aged individuals have lower administrative capital and therefore experience higher administrative burdens with program applications and renewal</i>	<i>Measuring the age and health factors of individuals in the program can determine levels of administrative burdens</i>	<i>Program application process data disaggregated by participant age/health factor variables</i>	<i>Disaggregating specific factors related to decreased competency may indicate different experiences of burdens by different groups of individuals for the same procedures</i>	<i>Process Measurement</i>	<i>Herd, 2015</i>	<i>Targeted, Practice assistance to individuals predicted to need it based on age and health factors.</i>	<i>Assistance Optimization</i>	<i>Supervised Learning</i>	<i>Classification</i>

Compliance Costs	Shifted burdens to state vs. individual	Does the state take responsibility for determining eligibility?	Measures of application requirements based on responsibility (government versus individuals)	Identification of burdens in processes, and then whether they are the responsibility of individuals or the government. Switching responsibility to the government will reduce individuals' burden.	Application Measurement	Herd et al., 2013	Automated application processes and eligibility determination shift the burden to the government from individuals.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Benefits Bundling	Does requested information or adjudication determine eligibility for multiple benefits?	Number of program eligibility determinations per application	Using application processes for multiple benefits can cut down on experienced burdens by factors.	Process Measurement	Herd et al., 2013	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoenrollment	Unsupervised Learning	Clustering
Compliance Costs	Rate of program application start as compared to those that complete program application	Higher rates of program application completion can be representative of lower compliance costs when controlling for other variables	Application completion rates	Measuring application completion rates of all eligible individuals, all individuals who begin the process, or all individuals who are inquiring about the program. Need to account for other factors of why people may not apply or complete their applications.	Application Measurement	Masood and Nisar, 2021	Automated application process based on entity resolution and use of administrative data.	Autoenrollment	Supervised Learning	Classification
Compliance Costs	Are program take-up rates different based on protected categories?	Potential indicators of biases or administrative exclusion based on protected attributes (Sex, race, gender, age, etc.)	Program take-up rates disaggregated by protected category factors	Identifying the correlation between protected categories and program participation or take-up rates can identify biases in burden experiences.	Program Measurement	Moreno and Mullins, 2017	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoenrollment	Unsupervised Learning	Clustering
Compliance Costs	Does the program account easily for non-traditional applicant situations?	Individuals who do not neatly fit into program rules or processes increase administrative burdens	Program participant edge cases, take-up rate	Third-gender individuals are applying for identification, etc. Requirements for parental consent without parents	Program Measurement	Nisar, 2018	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoenrollment	Supervised Learning	Classification
Compliance Costs	How do individual agencies modify experienced burdens over time?	Individuals can react and respond to burdens overtime to modify their effects	Compliance cost measures disaggregated by participant administrative experience (first-time applicant versus returning applicant)	Increased agency and bureaucratic competence can offset administrative burdens over time	Application Measurement	Peeters and Campos, 2021	Proactive, targeted outreach and program assistance based on a prediction of the level of individual agency of program experience.	Assistance Optimization	Supervised Learning	Classification
Compliance Costs	Time spent dealing with application	Polling information about time spent seeking to comply	Amount of application and compliance forms	Translating application compliance costs times into	Application Measurement	Pfeffer et al., 2020	Automated program application and	Autoadjudication	Supervised Learning	Classification

	and benefits procedures	with program requirements; turning time into economic costs		monetary values based on economic estimates.			eligibility determinations.			
Compliance Costs	Increased red tape is correlated with increased compliance costs, especially when this is shifted to the individuals rather than the government	Internal bureaucratic processes can correlate with increased compliance costs to individuals (unless all red tape costs are born by the government with no noticeable delay in adjudication or administration)	Measures of compliance costs compared to red tape measurements	Voter restoration laws leave much up to applicants to navigate, and internal confusion amongst agencies increases the burdens on individuals to comply and the feelings of their "worthiness" of the benefits	Process Measurement	Selin, 2019	Automated, more accurate processes based on entity resolution and eligibility predictions.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Increased access of groups or organizations to navigate the application process, which benefits individuals	Reliance on groups and organization to comply with program benefits are associated with decreased compliance costs	Compliance steps available and taken by for third-party on individuals' behalf	NGOs, nonprofits, condo associations, etc., can apply for benefits on behalf of individuals, and a group of individuals allow the reliance on "experts" or distribute costs/frictions among groups.	Assistance Measurement	Shybalkina, 2020	Proactive assignment to a third-party group to assist with application based on predicted need/benefit.	Assistance Optimization	Supervised Learning	Classification
Compliance Costs	Measures of burden based on PRA estimates	Higher PRA burdens are associated with higher compliance costs for program applications	PRA Measures	PRA measures as currently applied in the federal government	Application Measurement	Sunstein, 2018	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Measures of costs to individuals multiplied by all individuals compared to savings of the outcomes or processes	Sludge Audits should be conducted on administrative burdens associated with impacts on individuals	PRA Measures and benefits-costs analysis	Sludge audits are a combination of PRA estimates, benefit-cost analysis, and qualitative evaluations of burdens	Application Measurement	Sunstein, 2020	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	The proportion of eligible individuals not accessing the program or not applying for the program	Burdens are harmful or too onerous if they are screening out eligible individuals from the program	Program take-up rates	Take-up rates that indicate eligible individuals are not applying for, found eligible, or using the program indicate burdens are too high.	Program Measurement	Sunstein and Gosset, 2020	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Do third-party organizations absorb frictions or costs rather than passing them onto individuals?	How much of the administrative burdens are experienced by individuals rather than the government or third parties?	Compliance measures borne by third parties versus individuals (when third parties are present)	If third-party groups are available, what proportion of individuals use them?	Process Measurement	Wiley and Berry, 2018	Proactive assignment to a third-party group to assist with application based on predicted need/benefit.	Assistance Optimization	Supervised Learning	Classification
Compliance Costs	Call abandonment rate, call answered rates	Answer times increase compliance costs; measurements of time spent attempting to access	Call abandonment rate, call answered rates	Measuring specific customer interactions rates and time periods	Assistance Measurement	Code for America, 2021	Chatbot to increase call assistance and first-time resolution rates.	Virtual Assistant	Supervised Learning	Classification
Compliance Costs	First call resolution rates	Answer times increase compliance costs; measurements of time spent attempting to access	First call resolution rates	Resolving inquiries and programs with fewer interaction means fewer compliance	Assistance Measurement	Code for America, 2021	Chatbot to increase call assistance and first-time resolution rates.	Virtual Assistant	Supervised Learning	Classification

				costs experienced.						
Compliance Costs	Application Completion Rates	Ability to begin and complete the applications	Application Completion Rates	Higher application rates can indicate that compliance costs are not too high	Application Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Application times	Time spent accessing an application	Application times	A measure of time (can be a PRA estimate or direct measurements)	Application Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Procedural denials (missed interview, document, requirements)	The amount of compliance that determines intelligibility	Procedural denials (missed interview, document, requirements)	Procedural denials can indicate compliance costs are too high.	Process Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Renewal Rates	ability to reapply and recertify after already being eligible	Renewal Rates	The ability and willingness to renew program eligibility is an indicator of compliance cost levels.	Application Measurement	Code for America, 2021	Automated re-enrollment based on administrative data and adjudication predictions.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Accuracy Rates of Benefits	Administrative accuracy and inaccuracy can increase compliance costs	Accuracy Rates of Benefits	Increased accuracy in program administration lowers compliance costs	Process Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	Appeals resulting in decision reversal	False negatives of adjudication	Appeals resulting in decision reversal	Successful appeals may indicate compliance costs that are too high	Process Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	False positives and negatives for benefits decisions	False negatives of adjudication	False positives and negatives for benefits decisions	Solicit feedback about experienced stress of application process.	Process Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Compliance Costs	The churn rate of participants, especially eligible participants.	Ability to remain eligible/compliant	The churn rate of participants, especially eligible participants.	Program churn of participants may be related to compliance costs being too high	Application Measurement	Code for America, 2021	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Learning Costs	How do participants find out about the program?	Amount of outreach about the program, as measured by efforts, budget, and reach.	Outreach Measurement	The amount of proactive information about the program.	Outreach Measurement	Herd & Moynihan, 2018	Automatic proactive eligibility determination and targeted outreach about the program.	Autoenrollment	Supervised Learning	Classification
Learning Costs	How do individuals establish eligibility?	Amount of information and tools that help individuals understand eligibility requirements.	Availability and use of eligibility tools	Do tools or consultations exist for individuals to understand the eligibility as applied to their circumstances?	Tool measurement	Herd & Moynihan, 2018	ML categorization model that predicts eligibility and benefits levels based on minimum hypothetical data entered.	Autoadjudication	Supervised Learning	Classification

Learning Costs	How do individuals understand the benefits?	Information and tools to provide information about benefits, including resources for specific and individual situational understanding.	Availability and use of program and benefit tools	Do tools or consultations exist for individuals to understand the benefits as applied to their circumstances?	Tool measurement	Herd & Moynihan, 2018	ML categorization model that predicts eligibility and benefits levels based on minimum hypothetical data entered.	Autoadjudication	Supervised Learning	Classification
Learning Costs	How do individuals learn about the application process?	Resources and services focused on explaining and assisting with the application process.	Availability and use of application tools and services	Application checklists, consultation services, assistance completing applications?	Tool measurement	Herd & Moynihan, 2018	Chat Bot is based on natural language processing that can answer questions about the application process via web or phone.	Virtual Assistant	Supervised Learning	Classification
Learning Costs	Does distrust of government or organization make it more difficult to learn about the program?	Distrust of the State can lead to more information gathering about the program because of distrust of government-sourced information.	Customer feedback about a government entity	Gathering information from non-government sources	Feedback Measurement	Ali & Altaf, 2021	Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.	Assistance Optimization	Supervised Learning	classification
Learning Costs	Do individuals consider the government a reliable source of information?	Unreliable state services require greater diligence and alternative sources of information	Customer feedback about program use	Survey of individuals' beliefs about government reliability about the program.	Feedback Measurement	Ali & Altaf, 2021	Increased opinions and experiences of government reliability based on increased accuracy to benefit adjudications and program administration (auxiliary effect).	N/A	N/A	N/A
Learning Costs	Do individuals have to locate other sources of information?	Prevalence of government information-seeking vs. outside information seeking	Customer feedback about government vs. non-government sources of information	Looking to personal contacts or NGOs for information rather than government sources	Feedback Measurement	Ali & Altaf, 2021	Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.	Assistance Optimization	Supervised Learning	classification
Learning Costs	Are there more or fewer rules and requirements about how program benefits are redeemed?	Redemption of benefits with particular rules and requirements can increase compliance costs	Amount of rules regarding program use	SNAP benefits rules about how to redeem benefits, what they can be spent on, and where they can be redeemed in relation to areas where products or services would be purchased without benefits.	Process Measurement	Barnes, 2020	The burden for benefits redemption is automated and removed from the participant's responsibility.	Autoadjudication	Supervised Learning	Classification
Learning Costs	Take-up rate based on learning costs	Learning costs can also impact the take-up rate	Program take-up rate	Take-up can be calculated for EITC because the administrative data is available. Why not automatically enroll and adjudicate eligibility?	Program Measurement	Bhargava and Manoli, 2015	Targeted, proactive outreach to educate likely eligible individuals about the program.	Assistance Optimization	Supervised Learning	Classification
Learning Costs	What are the causal mechanisms of burdens within a specific program?	"Causal Process tracing" of a program allows a within-program case study association of causes of burdens.	Analysis of Survey responses	Statistical analyses of responses to identify causality	Feedback Measurement	Camillo, 2020	The automated root-cause analysis process increases the accuracy of causal process tracing.	Causal Inference	Unsupervised Learning	Clustering
Learning Costs	How does the required task or process differ based on the executive functioning of individuals?	First-time applicant vs. repeat applicants (more or less administrative experience); health and executive functioning; scarcity will decrease human capital	Program application process data disaggregated by participant executive function variables	Executive functioning (as controlled by an individual's health, level of scarcity, and cognitive decline) can impact the perceived experience with a requirement or	Application Measurement	Christensen et al., 2020	Proactive, differentiated outreach based on predictions of executive functioning levels.	Assistance Optimization	Supervised Learning	Classification

				action and, therefore, will impact the level of the cost of administrative burdens.						
Learning Costs	Is Non-take up due to lack of knowledge?	(1) Non-knowledge. Do welfare clients and ex-clients know that the program exists? Do potential claimants understand the program's eligibility rules? What learning costs do they face? Do service providers actively propose and "sell" the program to potential claimants?	Survey of program participants and eligible non-participants	Structure interviews aimed at assessing root causes on non-take-up	Feedback Measurement	Daigneaullt and Mace, 2020	Practice targeted outreach based on likely eligible individuals with high non-take-up rates.	Assistance Optimization	Supervised Learning	Classification
Learning Costs	Shifted burdens to state vs. individual	Does the state take responsibility for finding potentially eligible individuals?	Measures of program learning based on responsibility (government versus individuals)	Identification of burdens in processes, and then whether they are the responsibility of individuals or the government. Switching responsibility to the government will reduce individuals' burden.	Outreach Measurement	Herd et al., 2013	Automated application processes and eligibility determination shift the burden to the government from individuals.	Autoadjudication	Supervised Learning	Classification
Learning Costs	Benefits Bundling	Does one application review eligibility for multiple benefits?	Number of program eligibility determinations per application	Using application processes for multiple benefits can cut down on experienced burdens by factors.	Process Measurement	Herd et al., 2013	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoadjudication	Unsupervised Learning	Clustering
Learning Costs	Early Nudging about program information/eligibility	Providing behavioral nudges early in program learning can reduce learning costs	Measure of program up-take based on nudges	Despite increased overall administrative steps, learning costs can be lessened by behavioral nudges aimed at education	Program Measurement	Linos et al., 2020	Targeted, proactive outreach based on predicted lower take-up rates of individuals.	Assistance Optimization	Supervised Learning	Classification
Learning Costs	Are program take-up rates different based on protected categories?	Potential indicators of biases or administrative exclusion based on protected attributes (Sex, race, gender, age, etc.)	Application completion rates, disaggregated by protected category factors	Identifying the correlation between protected categories and program participation or take-up rates can identify biases in burden experiences.	Application Measurement	Moreno and Mullins, 2017	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoenrollment	Unsupervised Learning	Clustering
Learning Costs	Does the program account easily for non-traditional applicant situations?	Individuals who do not neatly fit into program rules or processes increase administrative burdens	Program participant edge cases, take-up rate	Third-gender individuals are applying for identification, etc. Requirements for parental consent without parents	Program Measurement	Nisar, 2018	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and	Autoenrollment	Supervised Learning	Classification

							programs needed and eligible for.			
Learning Costs	Increased access of groups or organizations to navigate the application process, which benefits individuals	Reliance on groups and organization to comply with program benefits are associated with decreased compliance costs	The proportion of third-party entities performing compliance processes	If third-party groups are available, what proportion of individuals use them?	Assistance Measurement	Shybalkina, 2020	Proactive assignment to a third-party group to assist with application based on predicted need/benefit.	Assistance Optimization	Supervised Learning	Classification
Learning Costs	Measures of burden based on PRA estimates	Higher estimates of PRA burdens are associated with more difficult requirements to learn and understand, as well as more difficult program eligibility requirements	PRA Measures	PRA measures as currently applied in the federal government	Application Measurement	Sunstein, 2018	Automated application completion and eligibility determination based on the use of administrative data and adjudication.	Autoadjudication	Supervised Learning	Classification
Psychological Costs	Are interactions with the application process stressful?	Customer feedback survey	Customer feedback about the application process	Solicit feedback about experienced stress of application process.	Feedback Measurement	Herd & Moynihan, 2018	Targeted proactive outreach to individuals based on predicted psychological costs experienced.	Assistance Optimization	Supervised Learning	Classification
Psychological Costs	Do people receive respectful treatment?	Customer feedback and employee evaluation.	Customer feedback about the application process	Solicit feedback about experiences with program staff.	Feedback Measurement	Herd & Moynihan, 2018	Targeted proactive outreach to individuals based on predicted psychological costs experienced.	Assistance Optimization	Supervised Learning	Classification
Psychological Costs	Do people enjoy some autonomy in the interaction?	Qualitative assessment of the application and benefit redemption process.	Customer feedback about program use	Solicit feedback about experiences and self-worth based on the program.	Feedback Measurement	Herd & Moynihan, 2018	Targeted proactive outreach to individuals based on predicted psychological costs experienced.	Assistance Optimization	Supervised Learning	classification
Psychological Costs	Degradation, disempowerment, and frustration at intrusive, directive, or judgmental bureaucratic encounters	Negative opinions or prior experience with the state produce negative associations or feelings about the program or about worthiness for the program.	Customer feedback about prior government interactions	Degradation, disempowerment, and frustration at intrusive, directive, or judgmental bureaucratic encounters	Feedback Measurement	Ali & Altaf, 2021	Automatic categorization for third-party outreach based on a prediction of individuals most likely to have low government trust.	Assistance Optimization	Supervised Learning	classification
Psychological Costs	Loss of Autonomy through interaction with the program	Resentment and fear of the state and its representatives as a repressive, controlling, or extractive entity caused by Waiting times and spaces communicating the state's (dis)regard of its citizens	Measurement of application step wait times, customer feedback about experiences	Survey about individuals' experience during application process wait times, interactions with program staff, etc.	Application Measurement	Ali & Altaf, 2021	Automatic data completion, enrollment, re-enrollment from administrative data, and proactive benefits eligibility determinations and requests for verification to individuals.	Autoenrollment	Supervised Learning	classification
Psychological Costs	The stigma associated with the program or having to interact with the government for assistance	The stigma of associating with the state because of negative associations with government services	Customer feedback about program use	A survey about individuals who experienced feelings based on associating with the program	Feedback Measurement	Ali & Altaf, 2021	Targeted proactive outreach to individuals based on predicted psychological costs experienced.	Assistance Optimization	Supervised Learning	Classification
Psychological Costs	Stress is associated with distrust of the state and having to rely on it	The stress of greater diligence required to determine the reliability and safety of services (vaccines)	Customer feedback about program use	Survey of feelings of stress, do they have the means to use non-government services?	Feedback Measurement	Ali & Altaf, 2021	Targeted proactive outreach to individuals based on predicted psychological costs experienced.	Assistance Optimization	Supervised Learning	Classification
Psychological Costs	Does the program offer benefits based on qualifications that form a label about the recipient?	Certain means-tested programs potentially identify applicants within a category or label which has stigma within society.	Population attitude towards program and recipients	Survey of individuals' understanding of identity based on program participation	Feedback Measurement	Baekgaard et al., 2021	Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program	Autoenrollment	Supervised Learning	Classification

							participation more opaque.			
Psychological Costs	Does complying increase stress or stigma specific to the compliance process?	Increased compliance demands are correlated with increases in psychological costs of stigma, loss of autonomy, and stress.	A measure of compliance costs as compared to customer feedback about the program	Looking at the changes to program applications brought about by digital applications	Feedback Measurement	Baekgaard et al., 2021	Decrease compliance requirements through automatic form completion and automatic enrollment/re-enrollment.	Autocompletion	Supervised Learning	Classification
Psychological Costs	Survey questions about stress	Stress experienced based on program participation	Customer feedback about program use	Survey regarding experienced stress based on program application or participation.	Feedback Measurement	Baekgaard et al., 2021	Decreased experienced stress based on program changes (auxiliary effect).	N/A	N/A	N/A
Psychological Costs	Survey questions about stigma	Stress experienced based on program participation	Customer feedback about program use	Survey regarding experienced stress based on program participation and benefits use.	Feedback Measurement	Baekgaard et al., 2021	Decreased experienced stress based on program changes (auxiliary effect).	N/A	N/A	N/A
Psychological Costs	Survey questions about Autonomy Loss	Qualitative assessment of the application and benefit redemption process.	Customer feedback about program use	Decreased compliance costs result in an increased sense of autonomy when participating in the program.	Feedback Measurement	Baekgaard et al., 2021	Increased sense of autonomy based on program changes and reduced compliance costs (auxiliary effect).	N/A	N/A	N/A
Psychological Costs	What is the overall balance of clients' evaluations of their experiences (positive, negative, or neutral)? Do clients distinguish bureaucrats from bureaucracies in making these assessments? What causal explanations (attributions of control) do clients provide for the basis of their evaluations?	Attributions of control over burdens (individual vs. government actor) can impact the associated magnitude of the burden.	Customer feedback about program staff	Feedback about experiences and understanding of program staff	Feedback Measurement	Barnes and Henly, 2018	Potential auxiliary impact of other program changes.	N/A	N/A	N/A
Psychological Costs	What is the rate of take-up of program participation as compared to the eligible or potentially eligible populations?	Psychological costs can attribute to lower program take-up because of the associated frictions.	Program take-up rate	Take-up can be calculated for EITC because the administrative data is available. Why not automatically enroll/adjudicate eligibility?	Program Measurement	Bhargava and Manoli, 2015	Targeted proactive outreach changes for individuals less likely to successfully access the program.	Assistance Optimization	Supervised Learning	Classification
Psychological Costs	What are the causal mechanisms of burdens within a specific program?	"Causal Process tracing" of a program allows a within-program case study association of causes of burdens.	Analysis of Survey responses	Statistical analyses of responses to identify causality	Feedback Measurement	Camillo, 2020	The automated root-cause analysis process increases the accuracy of causal process tracing.	Causal Inference	Unsupervised Learning	Clustering
Psychological Costs	How does the required task or process differ based on the executive functioning of individuals?	First-time applicant vs. repeat applicants (more or less administrative experience); health and executive functioning; scarcity will decrease human capital	Program survey feedback disaggregated by participant executive function variables	Executive functioning (as controlled by an individual's health, level of scarcity, and cognitive decline) can impact the perceived experience with a requirement or action and, therefore, will impact the level of the cost of	Feedback Measurement	Christensen et al., 2020	Proactive, differentiated outreach based on predictions of executive functioning levels.	Assistance Optimization	Supervised Learning	Classification

				administrative burdens.						
Psychological Costs	Perceptions of programs and participants?	An attitude of individuals participating in the program?	Customer feedback about program use	Measuring attitudes and opinions of citizens about the programs and participants to determine psychological costs on potential participants.	Feedback Measurement	Haeder et al., 2021	Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in an understanding of individuals' eligibility, or making program participation more opaque (auxiliary effect).	N/A	N/A	N/A
Psychological Costs	Measuring emotions through physiological measurements (facial coding, electrodermal activity, heart rate) to determine psychological costs	A physical manifestation of stress and physiological feels which can be measured	Direct measurements of program participant's feelings	Physical measurements of stress	Feedback Measurement	Hattke et al., 2020	Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).	N/A	N/A	N/A
Psychological Costs	What are the levels of compliance costs in the programs?	Reducing compliance costs lowers psychological costs	Participant feedback on the program as compared to individuals compliance costs	Psychological costs increase as compliance costs increase	Feedback Measurement	Baekgaard et al., 2021	Automated application completion and program adjudication.	Autocompletion	Supervised Learning	Classification
Psychological Costs	Are program take-up rates different based on protected categories?	Potential indicators of biases or administrative exclusion based on protected attributes (Sex, race, gender, age, etc.)	Program take-up rates disaggregated by protected category factors and compared to participant feedback on the program	Identifying the correlation between protected categories and program participation or take-up rates can identify biases in burden experiences.	Program Measurement	Moreno and Mullins, 2017	Automated application process and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoenrollment	Unsupervised Learning	Clustering
Psychological Costs	Associations between public support for administrative burdens can increase psychological costs on potential program participants	High learning and compliance costs can increase psychological costs	Measures of learning and compliance costs compared to applicant feedback about the program	Identifying the correlation between protected categories and program participation or take-up rates can identify biases in burden experiences.	Feedback Measurement	Nicholson-Crotty et al., 2021	Automatic enrollments and re-enrollment potentially reduce stigma because of changed attitudes about the program, changed attitudes in understanding or individuals' eligibility, or making program participation more opaque (auxiliary effect).	Autoenrollment	N/A	N/A
Psychological Costs	Does the program account easily for non-traditional applicant situations?	Individuals who do not neatly fit into program rules or processes increase administrative burdens	Program participant edge cases, take-up rate	Third-gender individuals are applying for identification, etc. Requirements for parental consent without parents	Program Measurement	Nisar, 2018	Automated application processes and eligibility determination shift the burden to the government from individuals for multiple programs based on identifying groups of individuals and programs needed and eligible for.	Autoenrollment	Supervised Learning	Classification
Psychological Costs	Increased red tape is correlated with increased compliance costs, especially when this is shifted to	Increased compliance demands are correlated with increases in psychological costs	Measures of compliance costs borne by participants versus government	Voter restoration laws leave much up to applicants to navigate, and	Process Measurement	Selin, 2019	Automated, more accurate processes based on entity resolution and eligibility predictions.	Autoadjudication	Supervised Learning	Classification

	<i>the individuals rather than the government</i>	<i>of stigma, loss of autonomy, and stress.</i>		<i>internal confusion amongst agencies increases the burdens on individuals to comply and the feelings of their "worthiness" of the benefits</i>						
<i>Psychological Costs</i>	<i>Measures of burden based on PRA estimates</i>	<i>Higher PRA estimates are associated with increased feelings of privacy impositions and increased judgment of "worthiness."</i>	<i>PRA Measures</i>	<i>PRA measures as currently applied in the federal government</i>	<i>Application Measurement</i>	<i>Sunstein, 2018</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>
<i>Psychological Costs</i>	<i>Customer satisfaction rates</i>	<i>rates of individuals' experiences with the program</i>	<i>Customer satisfaction rates</i>	<i>High satisfaction rates with the program can indicate lower stress and stigma associated with program interactions</i>	<i>Feedback Measurement</i>	<i>Code for America, 2021</i>	<i>Auxiliary effect based on increased accuracy of program administration and decreased experienced burdens.</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
<i>Psychological Costs</i>	<i>Appeals resulting in decision reversal</i>	<i>False negatives of adjudication</i>	<i>Appeals resulting in decision reversal</i>	<i>Higher appeal rates likely indicate higher levels of experiencing stress and stigma</i>	<i>Process Measurement</i>	<i>Code for America, 2021</i>	<i>Automated application completion and eligibility determination based on the use of administrative data and adjudication.</i>	<i>Autoadjudication</i>	<i>Supervised Learning</i>	<i>Classification</i>

Appendix B – Frameworks Guide

In this appendix, I pull together the tools from the three frameworks in one consistent location to provide an easy-to-follow guide for practitioners who wish to use them as designed.

Performance Measurement – Strategic Goal Setting

1. Set the specific strategic goal for the target program:

Improve the outcomes of [the program] by reducing administrative burdens. By [time period], administrative burdens will be reduced by [X] percent overall, with at least a [X] percent reduction falling into each component of learning cost, compliance costs, and physiological costs.

2. Identify existing administrative burdens using example diagnostic questions and causes in Appendix A.

3. Identify measurement processes for each administrative burden cause that you will focus on using the below definitions.

Simplified Measurement	Explanation
Application Measurement	Measures of applications, ways to apply, methods of application, the burden of application (government vs. individuals), validation and certification processes, where in-person steps are required.
Assistance Measurement	Measures of the amount and variety of methods of assistance to applicants
Feedback Measurement	Measures of direct experiences based on feedback in the form of direct observations, surveys, interviews, focus groups, etc.
Form Measurement	Measures of information collections and forms such as numbers, time to complete, and amount of data elements.
Outreach Measurement	Measurement of the types, methods, resources, and presence of outreach to individuals.
Process Measurement	Measures of variety, types, responsible entity, error rates, number of instances of processes including adjudication decisions, appeals, re-enrolment, benefit redemption, etc.
Program Measurement	Measures of the overall program such as proportions of covered compared to eligible individuals (take-up rate), measurement of attrition, etc.

Tool measurement	Measurement of the availability, variety, and usefulness of tools designed to facilitate program exploration, application, explanation, or application processes.
-------------------------	---

4. Use the identified administrative burdens and measurement approaches to create milestones for your program in line with the overall strategic goal:

<p><u>Reduce Learning costs by [X] percent within the [time] period.</u></p> <ul style="list-style-type: none">• Component/Milestone 1.1: [identify program area of focus and target goals for reduction]• Component/Milestone 1.N: <p><u>Reduce Compliance costs by at least [X] percent with the [time] period.</u></p> <ul style="list-style-type: none">• Component/Milestone 2.1:• Component/Milestone 2.N: <p><u>Reduce Psychological costs by at least [X] percent during the [time] period.</u></p> <p>Component/Milestone 3.N:</p>

5. Create performance measurement indicators for each milestone:

APG Milestone 1.1: [identify program area of focus and target goals for reduction]

Indicator 1.1.1:

Value Type	FY1 Q1	FY1 Q2	FY1 Q3	FY1 Q4	FY2 Q1	FY2 Q2	FY2 Q3	FY2 Q4
Target	N/A	N/A	2	3	4	5	5	5
Available tools								
Actual Available Tools								

Indicator 1.1.2:

Value Type	FY1 Q1	FY1 Q2	FY1 Q3	FY1 Q4	FY2 Q1	FY2 Q2	FY2 Q3	FY2 Q4
Target Proportion of Tool Use	N/A	N/A	10%	20%	30%	40%	50%	50%
Actual								

Machine Learning Solutions

6. Based on component/milestone goals and areas for administrative burden reduction, identify appropriate machine learning solutions, if applicable, for those areas using ML definitions and examples solutions in Appendix A:

Type of ML	Sub-type	Definition
Supervised Learning	Classification	ML program draws conclusions from observed values to determine a category for new observations.
Supervised Learning	Regression	The ML program must estimate the relationships between variables and make predictions on the independent variable.
Supervised Learning	Forecasting	ML program learns trends from historical data and variables and applies this knowledge to predict future trends in data.

<i>Unsupervised Learning</i>	Clustering	ML program clusters by grouping sets of similar data (based on defined criteria) from a larger set of data.
<i>Unsupervised Learning</i>	Dimension Reduction	The ML program reduces the number of variables being considered based on variables that have no impact on the target variable or removing variables that covary with each other.
<i>Reinforcement Learning</i>	Positive Reinforcement	An event occurs because of specific behavior that is desirable, and therefore, the algorithm reinforces this behavior.
<i>Reinforcement Learning</i>	Negative Reinforcement	Strengthens behavior that occurs because of a negative condition or the absence of something which should be stopped or avoided.

Simplified ML Solution	Explanation
Assistance Optimization	ML solutions that predict and recommend methods, types, groups, or individuals to receive assistance, usually proactively before they request it.
Autoadjudication	ML solutions adjudicate decisions about eligibility, benefit types, amounts, and benefit redemption decisions.
Autocompletion	ML models autocomplete forms and other information collections from administrative data, user-submitted data, and open data sources through a variety of entity resolution techniques.
Autoenrollment	ML solutions automatically enroll individuals or groups into government services or benefits programs based on entity resolution and autoadjudication across single or multiple programs.
Causal Inference	ML models use administrative data, program data, and theoretical and logic models to predict causality between input variables and outcomes.
Location Optimization	ML solutions that predict the optimal types and placements of application assistance, benefit redemption, or other in-person offices involved in the program.

Virtual Assistant	ML solution that uses natural language processing to interact with individuals to provide assistance, information, and answer questions.
--------------------------	--

<i>Type of Cost</i>	<i>Simplified ML Solution</i>	<i>Count</i>
Compliance Costs	Assistance Optimization	7
	Autoadjudication	22
	Autocompletion	7
	Autoenrollment	8
	Causal Inference	1
	Location Optimization	2
	N/A	1
	Virtual Assistant	3
Compliance Costs Total		51
Learning Costs	Assistance Optimization	7
	Autoadjudication	6
	Autoenrollment	3
	Causal Inference	1
	N/A	1
	Virtual Assistant	1
Learning Costs Total		19
Psychological Costs	Assistance Optimization	8
	Autoadjudication	3
	Autocompletion	2
	Autoenrollment	7
	Causal Inference	1
	N/A	5
Psychological Costs Total		26
Grand Total		96

7. Based on identified ML solutions, use the MLOps framework to design, develop, and implement:

*Design**Problem Development*

- **What is the problem trying to be solved?** Determining effectively and efficiently if an individual is eligible or not eligible for program benefits.
- **What is the definition of success? What are the acceptance tests?** If the ML solution can determine eligibility at least as accurately as manual adjudication rates and more efficiently based on time and resources, then it should be used.
- **What are the performance indicators that would be used to measure success?** Accuracy rates of eligibility predictions across all individual types. The efficiency of predictions as compared to manual adjudications.

Data

- **Do you have access to the data needed?** The agency has access to more than ten years of application and program data; through data, sharing has access to individual data from multiple records systems for identity resolution, including socioeconomic indicators.
- **Can the data be used for this purpose?** The agency has updated our regulations and information collections to notify the public about this use of their data, how these adjudications will be made, and other pertinent information about the ML process for oversight.
- **Is the data up-to-date and accurate?** Data sharing agreements and data pipelines provide real-time access to all data sources.
- **Is the data representative of the population of interest?** New groups of individuals who never applied or were found eligible before are under-represented in the data. As program changes are made which create this gap, data transformation will be used to provide sufficient representation in training data.
- **Could using this data lead to biased results?** Because prior adjudication results will be used, and there is a potential for biased adjudications previously made by human adjudicators, the data and the model results will be tested for biases. If found, data will be transformed to create unbiased training and testing data to reduce or eliminate training bias.

Baselining

- **What is the baseline to measure the model against?** The baseline for comparison of the ML model is the results and indicators of performance for manual adjudications being replaced by the ML autoadjudication model.

*Development**Preprocessing*

- **What are the results from the exploratory data analysis?** EDA results indicate no biases in the underlying training data, and the data is sufficiently clean for ML model development.
- **Are the assumptions about the model documented and translated into automated checks?** Assumptions about how the model will be documented and automated checks are coded into the procedure, which will notify of any actions which fail these checks.
- **Are the data cleaning, preprocessing, and feature tuning steps documented for replication?** All preprocessing, training, and tuning steps taken are documented to allow for reproducibility.
- **How is missing data handled, and why?** No missing data is imputed into any of the data sets because we have multiple data sets to perform entity resolution, and the model has been tested and remains accurate without any missing data.

Model Evaluation

- **Does the ML solution work based on predetermined criteria?** The model is accurate based on predetermined criteria (accuracy, precision, sensitivity, F1 scores, specificity, AUC, etc.)
- **Do you have documented model performance metrics and acceptance ranges?** Model performance parameters with associated tolerance ranges are documented and coded to trigger model stoppage if thresholds are exceeded.
- **Check for overfitting, biases, and data leaks?** Model and data are evaluated for overfitting and bias.
- **Manually check misclassified examples. When does the model make mistakes?** There are no detectable patterns or causes for misclassifications.
- **Are the code, dependencies, and technical requirements documented?** All code and technical environment requirements and dependencies are documented and reproducible.

*Deployment/Operations**Production Technical Requirements*

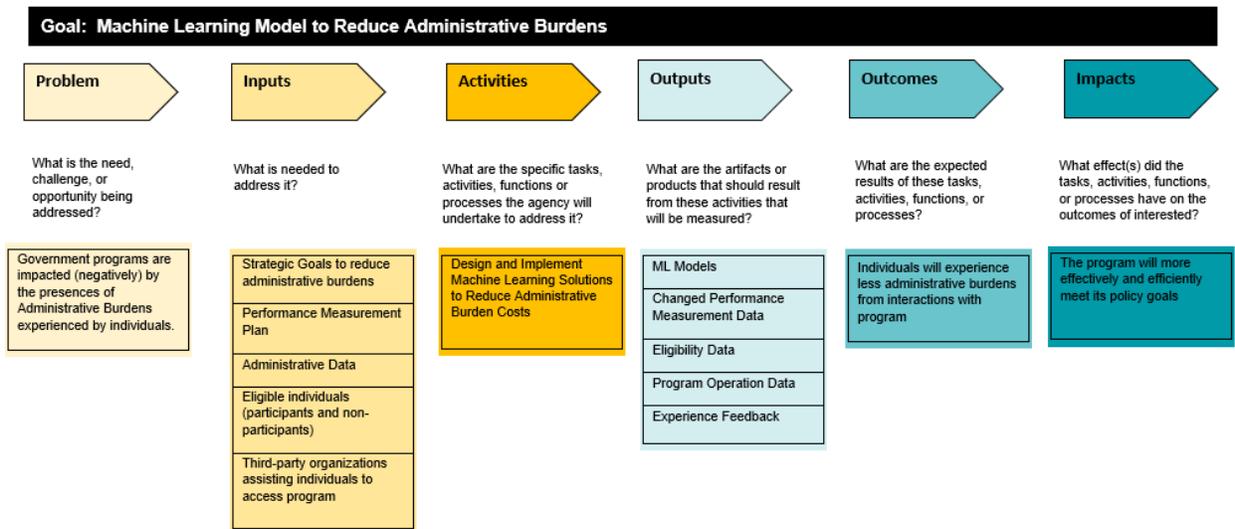
- **Will the ML model run in real-time or in batches?** Autoadjudication will run in batch once daily for all applications meeting requisite data requirements and will generate results and notifications the following business day.
- **Will there be additional processes or reviews for anomalous predictions?** Any anomalous results, as defined by the model and data thresholds and evaluation metrics, will be held in abeyance and reviewed the following business day.
- **What metadata, artifacts, and audit logs will be collected? When will they be reviewed?** All model metrics, adjudication and prediction metrics, and results will be logged. Representative samples of logs will be reviewed for accuracy and compliance once per week.
- **How often does the model need to be retrained?** The model must be retrained when policy or program changes are made regarding eligibility.

Monitoring

- **What data quality metrics need to be monitored in production?** Model and prediction accuracy, computational resources, and runtimes will be monitored.
- **What model criteria need to be monitored?** Accuracy, specificity, and F1 scores appropriate for the classification model will be monitored as well as data and model bias.
- **How will input drift and model degradation be monitored?** ML model predictions will be compared to eligibility criteria and determinations based on random sampling to determine model drift (becoming less accurate).
- **What feedback loops need attention?** Predictions of “eligible” where individuals decline benefits; predictions on “not eligible” where appeals are filed to determine if eligibility was wrongly predicted.

Evaluate ML Solutions and Impacts

8. Use and adapt the theory of change model to your specific program to help develop evaluations:



9. Based on current guidance, develop an evidence-building and evaluation plan for your program:

Evaluation Questions to Be Answered

- How effectively does the agency implement the program for eligible individuals?
- What are the levels of administrative burdens, and can they be reduced with the implementation of ML solutions?
- How does the ML solution impact the levels of administrative burden costs according to the performance measurement data?
- How does the reduction of administrative burdens affect the outcomes and impacts of the program overall?

Lead Office, Points of Contact

[The agency evaluation team/program office/etc.]

The rationale for the Evaluation

Addresses key questions in the learning agenda associated with the Agency Priority Goal in more effectively implementing this program.

Purpose of the Evaluation

Provides insights into the effects of ML solutions on levels of administrative burdens in the program and the changes to program outcomes and goals based on reduced administrative burdens.

Audience

The audience for this evaluation is [*agency leadership associated with the APG and program, program staff, and the public*].

Outcome Evaluation

The outcome evaluation will focus on the design and implementation of the ML solution within the programs to determine if it is reducing administrative burdens as designed.

Methods and Design

Methods of measurement will include:

-

Data Needed for the Evaluation

Performance measurement data, outcome data.

Anticipated Challenges

[*Challenges will include...*]

Summative/Impact Evaluation

The summative or impact evaluation will compare outcomes with the reduced administrative burdens to outcomes without the program, as well as outcomes without the reduced administrative burdens and those without the program to determine the impact of the overall program on individuals who participate and how those impacts differ with reduced administrative burdens.

Methods and Design

[X]

Data Needed for the Evaluation

- [Based on performance measurement plan and ML solutions]

Anticipated Challenges

Challenges for this evaluation are going to include...

Dissemination Plan

We will disseminate the evaluation data and the evaluation conclusions and recommendations to program staff, agency leadership, and the public through our website and publication on www.evaluation.gov.

Bibliography

- 40 USC 11301: Responsibility of Director. (n.d.). Retrieved February 27, 2022, from [https://uscode.house.gov/view.xhtml?req=\(title:40%20section:11301%20edition:prelim\)](https://uscode.house.gov/view.xhtml?req=(title:40%20section:11301%20edition:prelim))
- Aarøe, L., Baekgaard, M., Christensen, J., & Moynihan, D. P. (2021). Personality and Public Administration: Policymaker Tolerance of Administrative Burdens in Welfare Services. *Public Administration Review*, puar.13381. <https://doi.org/10.1111/puar.13381>
- Abelkop, A. D. K. (2010). Shifting Administrative Burdens: An Examination of Iowa's Medicaid Payment Rules on Specialized Wheelchairs for Nursing Facility Residents. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2164852>
- Adam, C., Steinebach, Y., & Knill, C. (2018). Neglected challenges to evidence-based policy-making: The problem of policy accumulation. *Policy Sciences*, 51(3), 269–290. <https://doi.org/10.1007/s11077-018-9318-4>
- Administrative Procedure Act (5 U.S.C. Subchapter II)*. (2016, August 15). National Archives. <https://www.archives.gov/federal-register/laws/administrative-procedure>
- Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.
- Alhadad, S. (2018). Visualizing Data to Support Judgement, Inference, and Decision Making in Learning Analytics: Insights from Cognitive Psychology and Visualization Science. *Journal of Learning Analytics*, 5(2), 60–85. <https://doi.org/10.18608/jla.2018.52.5>
- Ali, S. A. M., & Altaf, S. W. (2021). Citizen trust, administrative capacity and administrative burden in Pakistan's immunization program. *Journal of Behavioral Public Administration*, 4(1). <https://doi.org/10.30636/jbpa.41.184>
- Alla, S., & Adari, S. K. (2021). What Is MLOps? In S. Alla & S. K. Adari (Eds.), *Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure* (pp. 79–124). Apress. https://doi.org/10.1007/978-1-4842-6549-9_3
- Ames, F. L. (2015). *The Drive to Improve Performance in the Federal Government: A Longitudinal Case Study of Managing for Results* [Ph.D., George Mason University]. <http://www.proquest.com/docview/1695284737/abstract/9B8AFEC0DADA445DPQ/1>
- Andersen, L. B., Boesen, A., & Pedersen, L. H. (2016). Performance in Public Organizations: Clarifying the Conceptual Space. *Public Administration Review*, 76(6), 852–862. <https://doi.org/10.1111/puar.12578>

- Androusoyopoulou, A., & Charalabidis, Y. (2018). A framework for evidence based policy making combining big data, dynamic modelling and machine intelligence. *ICEGOV*. <https://doi.org/10.1145/3209415.3209427>
- Anna-Katharina Dhungel, Wessel, D., Zoubir, M., Mourad Zoubir, & Heine, M. (2021). Too Bureaucratic to Flexibly Learn About AI? The Human-Centered Development of a MOOC on Artificial Intelligence in and for Public Administration. *Mensch Und Computer*, 563–567. <https://doi.org/10.1145/3473856.3473998>
- Aridor, G., Che, Y.-K., Nelson, W., & Salz, T. (2020). Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR. *Social Science Research Network (SSRN)*, 67.
- Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions* (Rev. and expanded ed., 3. [print]). Harper Collins Publ.
- Arinder, M. K. (2016). Bridging the Divide between Evidence and Policy in Public Sector Decision Making: A Practitioner’s Perspective. *Public Administration Review*, 76(3), 394–398.
- Artificial Intelligence | GSA - IT Modernization Centers of Excellence*. (n.d.). Retrieved March 5, 2022, from <https://coe.gsa.gov/coe/artificial-intelligence.html>
- Ayers, R. S., Malgeri, J. R., & Press, J. E. (2014). ASSESING SENIOR PERFORMANCE COUNCILS Structures, Processes, and Promising Practices for Implementing the GPRA Modernization Act of 2010. *PUBLIC PERFORMANCE & MANAGEMENT REVIEW*, 38(1), 152–186. <https://doi.org/10.2753/PMR1530-9576380107>
- Barnes, C. Y. (2020). “It Takes a While to Get Used to”: The Costs of Redeeming Public Benefits. *Journal of Public Administration Research and Theory*, muaa042. <https://doi.org/10.1093/jopart/muaa042>
- Baron, J. (2018). A Brief History of Evidence-Based Policy. *The ANNALS of the American Academy of Political and Social Science*, 678(1), 40–50. <https://doi.org/10.1177/0002716218763128>
- Battaglio, R. P., Belardinelli, P., Bellé, N., & Cantarelli, P. (2019). Behavioral Public Administration ad fontes: A Synthesis of Research on Bounded Rationality, Cognitive Biases, and Nudging in Public Organizations. *Public Administration Review*, 79(3), 304–320. <https://doi.org/10.1111/puar.12994>
- Baylor, D., Koc, L., Koo, C. Y., Lew, L., Mewald, C., Modi, A. N., Polyzotis, N., Ramesh, S., Roy, S., Whang, S. E., Wicke, M., Breck, E., Wilkiewicz, J., Zhang, X., Zinkevich, M.,

- Cheng, H.-T., Fiedel, N., Foo, C. Y., Haque, Z., ... Jain, V. (2017). TFX: A TensorFlow-Based Production-Scale Machine Learning Platform. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 1387–1395. <https://doi.org/10.1145/3097983.3098021>
- Bekhet, A. K., & Zauszniewski, J. A. (2012). Methodological triangulation: An approach to understanding data. *Nurse Researcher*, 20(2), 40–43. <https://doi.org/10.7748/nr2012.11.20.2.40.c9442>
- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *ArXiv:1810.01943 [Cs]*. <http://arxiv.org/abs/1810.01943>
- Berendt, B., & Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. *Artificial Intelligence and Law*, 22(2), 175–209. <https://doi.org/10.1007/s10506-013-9152-0>
- Berman, E. (2018). A Government of Laws and Not of Machines. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3098995>
- Bernick, E. D. (2021). ENVISIONING ADMINISTRATIVE PROCEDURE ACT ORIGINALISM. *SSRN Electronic Journal*, 49.
- Bhanot, S. P., & Linos, E. (2020). Behavioral Public Administration: Past, Present, and Future. *Public Administration Review*, 80(1), 168–171. <https://doi.org/10.1111/puar.13129>
- Bhargava, S., & Manoli, D. (2015). Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review*, 105(11), 3489–3529. <https://doi.org/10.1257/aer.20121493>
- Bogenschneider, K., & Corbett, T. (2010). *Evidence-based policymaking: Insights from policy-minded researchers and research-minded policymakers*. Routledge Academic.
- Bogenschneider, K., & Corbett, T. J. (2021a). *When Researchers Successfully Engaged Policymakers*. 235–253. <https://doi.org/10.4324/9781003057666-12>
- Bogenschneider, K., & Corbett, T. J. (2021b). *Why There Is a Disconnect Between Research and Policy, and What We Can Do*. 3–23. <https://doi.org/10.4324/9781003057666-2>
- Bogenschneider, K., & Corbett, T. J. (2021c). *Understanding Policymakers: Insights From Science*. 34–59. <https://doi.org/10.4324/9781003057666-4>

- Bourdeaux, C. (2008). Integrating Performance Information into Legislative Budget Processes. *Public Performance & Management Review*, 31(4), 547–569.
<https://doi.org/10.2753/PMR1530-9576310403>
- Bozeman, B. (1993). A Theory Of Government “Red Tape.” *Journal of Public Administration Research and Theory*. <https://doi.org/10.1093/oxfordjournals.jpart.a037171>
- Breuel, C. (2020, January 3). *ML Ops: Machine Learning as an Engineering Discipline*. Medium. <https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f>
- Burden, B. C., Canon, D. T., Mayer, K. R., & Moynihan, D. P. (2012). The Effect of Administrative Burden on Bureaucratic Perception of Policies: Evidence from Election Administration. *Public Administration Review*, 72(5), 741–751.
<https://doi.org/10.1111/j.1540-6210.2012.02600.x>
- Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 205395171562251.
<https://doi.org/10.1177/2053951715622512>
- Busuioc, M. (2020). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, puar.13293. <https://doi.org/10.1111/puar.13293>
- Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2), 277–292.
<https://doi.org/10.1007/s10618-010-0190-x>
- Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*, 103, 513–564.
- Calo, R. (2017). Calo, Ryan, Artificial Intelligence Policy: A Primer and Roadmap. *SSRN*, 28.
<http://dx.doi.org/10.2139/ssrn.3015350>
- Carrigan, C., Pandey, S. K., & Van Ryzin, G. G. (2020). Pursuing Consilience: Using Behavioral Public Administration to Connect Research on Bureaucratic Red Tape, Administrative Burden, and Regulation. *Public Administration Review*, 80(1), 46–52.
<https://doi.org/10.1111/puar.13143>
- Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The Use of Triangulation in Qualitative Research. *Oncology Nursing Forum*, 41(5), 545–547.
<https://doi.org/10.1188/14.ONF.545-547>
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

- Cartwright, N., & Stegenga, J. (2011). A theory of evidence for evidence-based policy. In P. Dawid, W. Twining, & M. Vasilaki (Eds.), *Evidence, Inference and Enquiry* (p. 291). Oup/British Academy.
- Cavalluzzo, K. S., & Ittner, C. D. (2003). *IMPLEMENTING PERFORMANCE MEASUREMENT INNOVATIONS: EVIDENCE FROM GOVERNMENT*. 54.
- Chassang, G. (2017). The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*, *11*. <https://doi.org/10.3332/ecancer.2017.709>
- Choi, I., & Moynihan, D. P. (2019). How to foster collaborative performance management? Key factors in the US federal agencies. *Public Management Review*, *21*(10), 1538–1559. <https://doi.org/10.1080/14719037.2019.1571275>
- Christensen, J., Aarøe, L., Baekgaard, M., Herd, P., & Moynihan, D. P. (2020). Human Capital and Administrative Burden: The Role of Cognitive Resources in Citizen-State Interactions. *Public Administration Review*, *80*(1), 127–136. <https://doi.org/10.1111/puar.13134>
- Christensen, J., Dahlmann, C. M., Mathiasen, A. H., Moynihan, D. P., & Petersen, N. B. G. (2018). How Do Elected Officials Evaluate Performance? Goal Preferences, Governance Preferences, and the Process of Goal Reprioritization. *Journal of Public Administration Research and Theory*, *28*(2), 197–211. <https://doi.org/10.1093/jopart/muy001>
- Chudnovsky, M., & Peeters, R. (2021). The unequal distribution of administrative burden: A framework and an illustrative case study for understanding variation in people's experience of burdens. *Social Policy & Administration*, *55*(4), 527–542. <https://doi.org/10.1111/spol.12639>
- Citron, D. K., & Pasquale, F. (2011). THE SCORED SOCIETY: DUE PROCESS FOR AUTOMATED PREDICTIONS. *WASHINGTON LAW REVIEW*, *89*, 34.
- Clark, C. (2013, April 26). Reinventing Government—Two Decades Later. *Government Executive*. <https://www.govexec.com/management/2013/04/what-reinvention-wrought/62836/>
- Coglianesi, C. (2019). AI in Adjudication and Administration. *Public Law and Legal Theory Research Paper Series*, *19*(41), 32.
- Coglianesi, C., & Lehr, D. (n.d.). Regulating by Robot: Administrative Decision Making in the Machine-Learning Era. *THE GEORGETOWN LAW JOURNAL*, *105*, 77.
- Coglianesi, C., & Lehr, D. (2017). Regulating by Robot: Administrative Decision Making in the Machine-Learning Era. *Georgetown Law Journal*, *105* *Geo. L.J.*(1147).

- Coglianesse, C., & Lehr, D. (2019). TRANSPARENCY AND ALGORITHMIC GOVERNANCE. *ADMINISTRATIVE LAW REVIEW*, 56.
- Cohen, A. B., Colby, D. C., Wailoo, K., & Zelizer, J. E. (Eds.). (2015). *Medicare and Medicaid at 50: America's entitlement programs in the age of affordable care*. Oxford University Press.
- Commission on Evidence-Based Policymaking. (2017). *The Promise of Evidence-based Policymaking*. <https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf>
- Congdon, W. J., & Shankar, M. (2018). The Role of Behavioral Economics in Evidence-Based Policymaking. *The ANNALS of the American Academy of Political and Social Science*, 678(1), 81–92. <https://doi.org/10.1177/0002716218766268>
- Connolly, J. M., Klofstad, C., & Uscinski, J. (2021). Administrative Burdens and Citizen Likelihood to Seek Local Public Services: The Case of Hurricane Shelters. *Public Performance & Management Review*, 44(3), 560–579. <https://doi.org/10.1080/15309576.2020.1818588>
- Crowley, D. M., & Scott, J. T. (2017). Bringing Rigor to the Use of Evidence in Policy Making: Translating Early Evidence. *Public Administration Review*, 77(5), 650–655. <https://doi.org/10.1111/puar.12830>
- Daigneault, P.-M., & Macé, C. (2020). Program Awareness, Administrative Burden, and Non-Take-Up of Québec's Supplement to the Work Premium. *International Journal of Public Administration*, 43(6), 527–539. <https://doi.org/10.1080/01900692.2019.1636397>
- Data.gov. (n.d.). Data.Gov. Retrieved January 10, 2022, from <https://www.data.gov/>
- Desmidt, S., & Meyfrootd, K. (2020). How does public disclosure of performance information affect politicians' attitudes towards effort allocation? Evidence from a survey experiment. *Journal of Public Administration Research and Theory*, muaa054. <https://doi.org/10.1093/jopart/muaa054>
- Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*, 49(7), 1043–1064. <https://doi.org/10.1177/0095399714555751>
- Destler, K. N. (2016). Creating a Performance Culture: Incentives, Climate, and Organizational Change. *The American Review of Public Administration*, 46(2), 201–225. <https://doi.org/10.1177/0275074014545381>

- Dey, S. S., Thommana, J., & Dock, S. (2015). Public Agency Performance Management for Improved Service Delivery in the Digital Age: Case Study. *Journal of Management in Engineering*, 31(5), 05014022. [https://doi.org/10.1061/\(ASCE\)ME.1943-5479.0000321](https://doi.org/10.1061/(ASCE)ME.1943-5479.0000321)
- Dhar, V. (2020). *Data Science and Prediction*. <https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext>
- Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Dooren, W. van, Bouckaert, G., & Halligan, J. (2015). *Performance management in the public sector* (Second edition). Routledge.
- Dull, M. (2008). Results-Model Reform Leadership: Questions of Credible Commitment. *Journal of Public Administration Research and Theory*, 19(2), 255–284. <https://doi.org/10.1093/jopart/mum043>
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Crick, T., Duan, Y., Dwivedi, R., Edwards, J. S., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., ... Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>
- Elias, R. A. (2016). The legislative history of the administrative procedure act. *Fordham Environmental Law Review*, 27(2), 19.
- Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3551505>
- Estimating Burden | A Guide to the Paperwork Reduction Act*. (2022a). <https://pra.digital.gov/burden/>
- Estimating Burden | A Guide to the Paperwork Reduction Act*. (2022b). [Federal Government]. *Estimating Burden | A Guide to the Paperwork Reduction Act*. <https://pra.digital.gov/burden/>
- Evaluation.gov*. (2022). <https://www.evaluation.gov/about/>
- Evelyn Z. Brodtkin. (1997). Inside the Welfare Contract: Discretion and Accountability in State Welfare Administration. *Social Service Review*. <https://doi.org/10.1086/604228>
- Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government – The White House*. (2020).

<https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/>

Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government. (2021, December 13). The White House.

<https://www.whitehouse.gov/briefing-room/presidential-actions/2021/12/13/executive-order-on-transforming-federal-customer-experience-and-service-delivery-to-rebuild-trust-in-government/>

FACT SHEET: Putting the Public First: Improving Customer Experience and Service Delivery for the American People. (2021, December 13). The White House.

<https://www.whitehouse.gov/briefing-room/statements-releases/2021/12/13/fact-sheet-putting-the-public-first-improving-customer-experience-and-service-delivery-for-the-american-people/>

Federal Data Strategy 2021 Action Plan. (2021). 24.

Fejes, E., & Futó, I. (2021). Artificial Intelligence in Public Administration – Supporting Administrative Decisions. *Pénzügyi Szemle = Public Finance Quarterly*, 66(Special edition 2021/1), 23–51. https://doi.org/10.35551/PFQ_2021_s_1_2

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 259–268. <https://doi.org/10.1145/2783258.2783311>

Fobia, A. C., Holzberg, J., Eggleston, C., Childs, J. H., Marlar, J., & Morales, G. (2019). Attitudes towards Data Linkage for Evidence-Based Policymaking. *Public Opinion Quarterly*, 83(S1), 264–279. <https://doi.org/10.1093/poq/nfz008>

Franzke, A. S., Muis, I., & Schäfer, M. T. (2021). Data Ethics Decision Aid (DEDA): A dialogical framework for ethical inquiry of AI and data projects in the Netherlands. *Ethics and Information Technology*. <https://doi.org/10.1007/s10676-020-09577-5>

Fuentes, A. (2018). *Hands-on predictive analytics with Python: Master the complete predictive analytics process, from problem definition to model deployment*. Packt Publishing. <http://proquestcombo.safaribooksonline.com/9781789138719>

Funk, W. F. (1987). The paperwork reduction act: Paperwork reduction meets administrative law. *Harvard Journal on Legislation*, 24(1), 117.

- Gamage, P. (2016). New development: Leveraging ‘big data’ analytics in the public sector. *Public Money & Management*, 36(5), 385–390.
<https://doi.org/10.1080/09540962.2016.1194087>
- Gao, J. (2015). Performance Measurement and Management in the Public Sector: Some Lessons from Research Evidence: PERFORMANCE MEASUREMENT AND MANAGEMENT IN THE PUBLIC SECTOR. *Public Administration and Development*, 35(2), 86–96.
<https://doi.org/10.1002/pad.1704>
- General Services Administration (GSA). (2020). *CoE Guide to AI Ethics*.
<https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf>
- Gerrits, L. (2020). Soul of a new machine: Self-learning algorithms in public administration. *Information Polity*, 26(3), 237–250. <https://doi.org/10.3233/ip-200224>
- GetYourRefund. (2021). *Code for America*. <https://codeforamerica.org/programs/tax-benefits/getyourrefund/>
- Giest, S. (2017). Big data for policymaking: Fad or fasttrack? *Policy Sciences*, 50(3), 367–382.
<https://doi.org/10.1007/s11077-017-9293-1>
- Gohwong, S. (2015). The Investigation of Artificial Intelligence Application in the Public Administration’s Literature. *RSU International Journal of College of Government (RSUIJCG)*, 2(1), 7.
- Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation.” *AI Magazine*, 38(3), 50.
<https://doi.org/10.1609/aimag.v38i3.2741>
- Gore, A. (1997). *Access America: Reengineering Through Information Technology*. DIANE Publishing.
- Granville, V. (2017, July 27). *Types of Machine Learning Algorithms in One Picture*—*DataScienceCentral.com*. Data Science Central.
<https://www.datasciencecentral.com/types-of-machine-learning-algorithms-in-one-picture/>
- Greco, C. (2021, October 25). *ML and MLOps at a Reasonable Scale*. Medium.
<https://towardsdatascience.com/ml-and-mlops-at-a-reasonable-scale-31d2c0782d9c>
- Grimmelikhuijsen, S., Jilke, S., Olsen, A. L., & Tummers, L. (2017). Behavioral Public Administration: Combining Insights from Public Administration and Psychology: PUBLIC ADMINISTRATION AND THE DISCIPLINES. *Public Administration Review*, 77(1), 45–56.
<https://doi.org/10.1111/puar.12609>

- Hälterlein, J. (2021). Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. *Big Data & Society*, 8(1), 205395172110031. <https://doi.org/10.1177/20539517211003118>
- Hamburger, J. (2014). The “Access Report”: Balancing the Privacy Benefits with the Administrative Burdens. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2381112>
- Haskins, R. (2018). Evidence-Based Policy: The Movement, the Goals, the Issues, the Promise. *The ANNALS of the American Academy of Political and Social Science*, 678(1), 8–37. <https://doi.org/10.1177/0002716218770642>
- Hassan, S., & Hatmaker, D. M. (2015). Leadership and Performance of Public Employees: Effects of the Quality and Characteristics of Manager-Employee Relationships. *Journal of Public Administration Research and Theory*, 25(4), 1127–1155. <https://doi.org/10.1093/jopart/muu002>
- Hatry, H. P. (2002). Performance Measurement: Fashions and Fallacies. *Public Performance & Management Review*, 25(4), 352–358. <https://doi.org/10.1080/15309576.2002.11643671>
- Hatry, H. P. (2010). Looking into the Crystal Ball: Performance Management over the Next Decade. *Public Administration Review*, 70, s208–s211. <https://doi.org/10.1111/j.1540-6210.2010.02274.x>
- Hatry, H. P. (2013). Sorting the Relationships Among Performance Measurement, Program Evaluation, and Performance Management. *New Directions for Evaluation*, 2013(137), 19–32. <https://doi.org/10.1002/ev.20043>
- Head, B., Ferguson, M., Cherney, A., & Boreham, P. (2014). Are policy-makers interested in social research? Exploring the sources and uses of valued information among public servants in Australia. *Policy and Society*, 33(2), 89–101. <https://doi.org/10.1016/j.polsoc.2014.04.004>
- Head, B. W. (2010). Reconsidering evidence-based policy: Key issues and challenges. *Policy & Society*, 29(2), 77.
- Head, B. W. (2016). Toward More “Evidence-Informed” Policy Making? *Public Administration Review*, 76(3), 472–484. <https://doi.org/10.1111/puar.12475>
- Heinrich, C. J. (2007). Evidence-Based Policy and Performance Management: Challenges and Prospects in Two Parallel Movements. *The American Review of Public Administration*, 37(3), 255–277. <https://doi.org/10.1177/0275074007301957>

- Heinrich, C. J. (2016). The Bite of Administrative Burden: A Theoretical and Empirical Investigation. *Journal of Public Administration Research and Theory*, 26(3), 403–420. <https://doi.org/10.1093/jopart/muv034>
- Heinrich, C. J., Camacho, S., Henderson, S. C., Hernández, M., & Joshi, E. (2021). Consequences of Administrative Burden for Social Safety Nets that Support the Healthy Development of Children. *Journal of Policy Analysis and Management*. <https://doi.org/10.1002/pam.22324>
- Herd, P., DeLeire, T., Harvey, H., & Moynihan, D. P. (2013). Shifting Administrative Burden to the State: The Case of Medicaid Take-Up. *Public Administration Review*, 73(s1), S69–S81. <https://doi.org/10.1111/puar.12114>
- Herd, P., & Moynihan, D. P. (2018). *Administrative burden: Policymaking by other means*. Russell Sage Foundation.
- Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170355. <https://doi.org/10.1098/rsta.2017.0355>
- Höchtel, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26(1–2), 147–169. <https://doi.org/10.1080/10919392.2015.1125187>
- House, P. W., & Shull, R. D. (1988). *Rush to policy: Using analytic techniques in public sector decision making*. Transaction Books.
- IRS. (2021). *Statistics for Tax Returns with the Earned Income Tax Credit (EITC) | Earned Income Tax Credit*. <https://www.eitc.irs.gov/eitc-central/statistics-for-tax-returns-with-eitc/statistics-for-tax-returns-with-the-earned-income>
- Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, 70, 338–345. <https://doi.org/10.1016/j.jbusres.2016.08.007>
- Jennings, E. T., & Hall, J. L. (2012). Evidence-Based Practice and the Use of Information in State Agency Decision Making. *Journal of Public Administration Research and Theory*, 22(2), 245–266. <https://doi.org/10.1093/jopart/mur040>
- Jensen, L. (2003). *Patriots, settlers, and the origins of American social policy*. Cambridge University Press.
- Jilke, S., Van Dooren, W., & Rys, S. (2018). Discrimination and Administrative Burden in Public Service Markets: Does a Public-Private Difference Exist? *Journal of Public*

- Administration Research & Theory*, 28(3), 423–439.
<https://doi.org/10.1093/jopart/muy009>
- Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Johansen, M., Kim, T., & Zhu, L. (2018). Managing for Results Differently: Examining Managers' Purposeful Performance Information Use in Public, Nonprofit, and Private Organizations. *The American Review of Public Administration*, 48(2), 133–147.
<https://doi.org/10.1177/0275074016676574>
- Katz, D. (1975). *Bureaucratic Encounters: A Pilot Study in the Evaluation of Government Services*. Survey Research Center, Institute for Social Research, University of Michigan.
- Katzenbach, C. (2021). “AI will fix this” – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society*, 8(2), 20539517211046184.
<https://doi.org/10.1177/20539517211046182>
- Keiser, L. R., & Miller, S. M. (2020). Does Administrative Burden Influence Public Support for Government Programs? Evidence from a Survey Experiment. *Public Administration Review*, 80(1), 137–150. <https://doi.org/10.1111/puar.13133>
- Keller, D. (2018). *The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation*. <https://doi.org/10.15779/z38639k53j>
- Khagram, S., & Thomas, C. W. (2010). Toward a Platinum Standard for Evidence-Based Assessment by 2020. *Public Administration Review*, 70, S100–S106.
<https://doi.org/10.1111/j.1540-6210.2010.02251.x>
- Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*, 57(3), 78–85. <https://doi.org/10.1145/2500873>
- Kissinger, H., Schmidt, E., Huttenlocher, D. P., & Schouten, S. (2021). *The age of AI: And our human future* (First edition). Little Brown and Company.
- Kroll, A. (2015a). Explaining the Use of Performance Information by Public Managers: A Planned-Behavior Approach. *The American Review of Public Administration*, 45(2), 201–215. <https://doi.org/10.1177/0275074013486180>
- Kroll, A. (2015b). Exploring the Link Between Performance Information Use and Organizational Performance: A Contingency Approach. *Public Performance & Management Review*, 39(1), 7–32. <https://doi.org/10.1080/15309576.2016.1071159>

- Kroll, A., & Moynihan, D. (2015). Does Training Matter? Evidence from Performance Management Reforms. *Public Administration Review*, 75(3), 411–420.
<https://doi.org/10.1111/puar.12331>
- Kroll, A., & Moynihan, D. (2017). The Design and Practice of Integrating Evidence: Connecting Performance Management with Program Evaluation: The Design and Practice of Integrating Evidence: Connecting Performance Management with Program Evaluation. *Public Administration Review*. <https://doi.org/10.1111/puar.12865>
- Kroll, A., & Moynihan, D. P. (2020). Tools of Control? Comparing Congressional and Presidential Performance Management Reforms. *Public Administration Review*. <https://doi.org/10.1111/puar.13312>
- Kroll, J. A., Huey, J., Barocas, S., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). ACCOUNTABLE ALGORITHMS. *University of Pennsylvania Law Review*, 165, 74.
- Landry, R., Lamari, M., & Amara, N. (2003). The Extent and Determinants of the Utilization of University Research in Government Agencies. *Public Administration Review*, 63(2), 192–205. <https://doi.org/10.1111/1540-6210.00279>
- Lane, J. (2018). Building an Infrastructure to Support the Use of Government Administrative Data for Program Performance and Social Science Research. *The ANNALS of the American Academy of Political and Social Science*, 675(1), 240–252.
<https://doi.org/10.1177/0002716217746652>
- Lavertu, S. (2016). We All Need Help: “Big Data” and the Mismeasure of Public Administration. *Public Administration Review*, 76(6), 864–872.
<https://doi.org/10.1111/puar.12436>
- LeRoux, K., & Wright, N. S. (2010). Does Performance Measurement Improve Strategic Decision Making? Findings From a National Survey of Nonprofit Social Service Agencies. *Nonprofit and Voluntary Sector Quarterly*, 39(4), 571–587.
<https://doi.org/10.1177/0899764009359942>
- Liu, H.-W., Lin, C.-F., & Chen, Y.-J. (2019). Beyond State v Loomis: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, 27(2), 122–141. <https://doi.org/10.1093/ijlit/eaz001>
- Lubbers, J. S. (1997). PAPERWORK REDUX: THE (STRONGER) PAPERWORK REDUCTION ACT OF 1995. *ADMINISTRATIVE LAW REVIEW*, 49, 13.
- MacCarthy, M. (2018). Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3154788>

- Maciejewski, M. (2017). To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences*, 83(1_suppl), 120–135. <https://doi.org/10.1177/0020852316640058>
- Madsen, J. K., Mikkelsen, K. S., & Moynihan, D. P. (2020). Burdens, Sludge, Ordeals, Red tape, Oh My!: A User's Guide to the Study of Frictions. *Public Administration*, n/a(n/a). <https://doi.org/10.1111/padm.12717>
- Mäkinen, S., Skogström, H., Laaksonen, E., & Mikkonen, T. (2021). Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, 109–112. <https://doi.org/10.1109/WAIN52551.2021.00024>
- March, J. G., & Olsen, J. P. (1984). The New Institutionalism: Organizational Factors in Political Life. *The American Political Science Review*, 78(3), 734–749. <https://doi.org/10.2307/1961840>
- March, J. G., & Olsen, J. P. (2008). *Elaborating the "New Institutionalism."* Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199548460.003.0001>
- Masood, A., & Nisar, M. (2020). Administrative Capital and Citizens' Responses to Administrative Burden. *Journal of Public Administration Research and Theory*. <https://doi.org/10.1093/jopart/muaa031>
- Matsumi, H. (2017). PREDICTIONS AND PRIVACY: SHOULD THERE BE RULES ABOUT USING PERSONAL DATA TO FORECAST THE FUTURE? *CUMBERLAND LAW REVIEW*, 48, 63.
- Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big Data in Public Affairs. *Public Administration Review*, 76(6), 928–937. <https://doi.org/10.1111/puar.12625>
- Mettler, S. (2011). *The submerged state: How invisible government policies undermine American democracy*. University of Chicago Press.
- Mihm, J. C. (2017). OMB and Agencies Should More Fully Implement the Process to Streamline Reporting Requirements. *GAO Reports*, i–ii.
- Moynihan, D. (2008). *The dynamics of performance management: Constructing information and reform*. Georgetown University Press.
- Moynihan, D., & Herd, P. (2010). Red Tape and Democracy: How Rules Affect Citizenship Rights: *The American Review of Public Administration*. <https://doi.org/10.1177/0275074010366732>

- Moynihan, D., Herd, P., & Harvey, H. (2015). Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions. *Journal of Public Administration Research and Theory*, 25(1), 43–69. <https://doi.org/10.1093/jopart/muu009>
- Moynihan, D., & Kroll, A. (2016). Performance Management Routines That Work? An Early Assessment of the GPRA Modernization Act. *Public Administration Review*, 76(2), 314–323. <https://doi.org/10.1111/puar.12434>
- Moynihan, D., & Lavertu, S. (2012). Does Involvement in Performance Management Routines Encourage Performance Information Use? Evaluating GPRA and PART. *Public Administration Review*, 72(4), 592–602. <https://doi.org/10.1111/j.1540-6210.2011.02539.x>
- Moynihan, D., Neilsen, P., & Krull, A. (2017). Managerial Use of Performance Data by Bureaucrats and Politicians. In *Experiments in Public Management Research: Challenges and Contributions* (pp. 244–269). Cambridge University Press.
- Moynihan, D., & Pandey, S. K. (2010). The Big Question for Performance Management: Why Do Managers Use Performance Information? *Journal of Public Administration Research and Theory*, 20(4), 849–866. <https://doi.org/10.1093/jopart/muq004>
- Office of Management and Budget, Exec. (2021). *OMB Circular A-11, Part 6, The Federal Performance Framework for Improving Program and Service Delivery*. Executive Office of the President. <https://www.whitehouse.gov/wp-content/uploads/2018/06/a11.pdf>
- O’Leary, P., Walker, E., & Roessel, E. (2015). *Social Security Disability Insurance at Age 60: Does It Still Reflect Congress’ Original Intent?* (No. 2015-01). Office of Retirement and Disability Policy. <https://www.ssa.gov/policy/docs/issuepapers/ip2015-01.html>
- Olsen, A., & James, O. (2017). Citizens and Public Performance Measures: Making Sense of Performance Information. In *Experiments in Public Management Research: Challenges and Contributions* (pp. 270–290). Cambridge University Press.
- OMB. (2022). *Ai.gov*. National Artificial Intelligence Initiative. <https://www.ai.gov/about/>
- O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.
- OPM. (2019). *Data Scientist Titling Guidance | CHCOC*. <https://www.chcoc.gov/content/data-scientist-titling-guidance>
- Orr, L. L. (2018). The Role of Evaluation in Building Evidence-Based Policy. *The ANNALS of the American Academy of Political and Social Science*, 678(1), 51–59. <https://doi.org/10.1177/0002716218764299>

- Orr, L. L., Olsen, R. B., Bell, S. H., Schmid, I., Shivji, A., & Stuart, E. A. (2019). Using the Results from Rigorous Multisite Evaluations to Inform Local Policy Decisions. *Journal of Policy Analysis and Management*, 38(4), 978–1003.
<https://doi.org/10.1002/pam.22154>
- Osborne, D., Osborne, D. E., Silverberg, R., & Gaebler, T. A. (1992). *Reinventing Government: How The Entrepreneurial Spirit Is Transforming The Public Sector*. Basic Books.
- Paperwork Reduction Act Guide*. (2017). OPM.GOV. <https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf>
- Patton, M. Q. (2012). *Essentials of utilization-focused evaluation*. SAGE.
- Pawson, R. (2002). Evidence-based policy: The promise of `realist synthesis`. *Evaluation: The International Journal of Theory, Research and Practice*, 8(3), 340–358.
<https://doi.org/10.1177/135638902401462448>
- Performance Improvement Council (Ed.). (2019). *P3 Playbook | PIC.gov*.
<https://www.pic.gov/playbook>
- Performance Improvement Council. (2022). *PIC Resources | PIC.gov*. PIC.Gov.
<https://www.pic.gov/goalplaybook/>
- Radin, B. A. (2003). Caught Between Agendas: GPRA, Devolution, and Politics. *International Journal of Public Administration*, 26(10–11), 1245–1255. <https://doi.org/10.1081/PAD-120019930>
- Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2* (Third edition). Packt.
- Riccucci, N. (2010). *Public administration: Traditions of inquiry and philosophies of knowledge*. Georgetown University Press.
- Robinson, D. G. (2017). *The Challenges of Prediction: Lessons from Criminal Justice* (SSRN Scholarly Paper ID 3054115). Social Science Research Network.
<https://papers.ssrn.com/abstract=3054115>
- Rogge, N., Agasisti, T., & De Witte, K. (2017). Big data and the measurement of public organizations' performance and efficiency: The state-of-the-art. *Public Policy and Administration*, 32(4), 263–281. <https://doi.org/10.1177/0952076716687355>
- Ryan, P. D. (2019, January 14). *H.R.4174 - 115th Congress (2017-2018): Foundations for Evidence-Based Policymaking Act of 2018 (2017/2018)* [Legislation].
<https://www.congress.gov/bill/115th-congress/house-bill/4174>

- Sanderson, I. (2002). Making Sense of ‘What Works’: Evidence Based Policy Making as Instrumental Rationality? *Public Policy & Administration*, 17(3), 61.
- Shybalkina, I. (2020). The role of organized groups in administrative burdens of property taxation. *Journal of Behavioral Public Administration*, 4(1).
<https://doi.org/10.30636/jbpa.41.179>
- Silver, N. (2012). *The signal and the noise: Why so many predictions fail--but some don't*. Penguin Press.
- Stack, K. (2018). The Office of Management and Budget: The Quarterback of Evidence-Based Policy in the Federal Government. *The ANNALS of the American Academy of Political and Social Science*, 678(1), 112–123. <https://doi.org/10.1177/0002716218768440>
- Stanley, J. W., & Lutz, R. (2021). Implementation of the Federal Performance Framework under Presidents Obama and Trump: A Comparative Analysis of Agency Strategic Plans. *Public Performance & Management Review*, 44(3), 682–705.
<https://doi.org/10.1080/15309576.2021.1884577>
- Sunstein, C. R. (2018). Sludge and Ordeals. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.3288192>
- Sunstein, C. R. (2020). Sludge Audits. *Behavioural Public Policy*, 1–20.
<https://doi.org/10.1017/bpp.2019.32>
- Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. *The Administrative Conference of the United States (ACUS)*. (2019).
<https://www.acus.gov/administrative-conference-united-states-acus>
- Three-in-One: The New Evidence Act*. (2020). | IBM Center for The Business of Government.
<http://businessofgovernment.org/blog/three-one-new-evidence-act>
- Treveil, M., Omont, N., Stenac, C., Lefevre, K., Phan, D., Zentici, J., Lavoillotte, A., Miyazaki, M., & Heidmann, L. (2020). *Introducing MLOps*. O'Reilly Media, Inc.
- United States Congress. (2021). *WILLIAM M. (MAC) THORNBERRY NATIONAL DEFENSE AUTHORIZATION ACT FOR FISCAL YEAR 2021*.
- U.S. Department of Health and Human Services. (2021). *HHS AI Strategy*.
- U.S. Government Accountability Office. (2021a, March). *Program Evaluation: Key Terms and Concepts*. <https://www.gao.gov/products/gao-21-404sp>
- U.S. Government Accountability Office. (2021b). *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*. GAO.
<https://www.gao.gov/products/gao-21-519sp>

- Vought, R. T. (2019). *Memorandum for Heads of Executive Departments and Agencies: Phase I Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance*. OMB.
<https://www.whitehouse.gov/wp-content/uploads/2019/07/M-19-23.pdf>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.3063289>
- Wagner, R. E. (2013). *Charging for Government (Routledge Revivals): User Charges and Earmarked Taxes in Principle and Practice*. Routledge.
- Weimer, D. L., & Vining, A. R. (2017). *Policy analysis: Concepts and practice* (Sixth edition). Routledge, Taylor & Francis Group.
- White House. (2014). *Big Data: Seizing Opportunities, Preserving Values*.
https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
- White, J., & Anderson, B. (2012). Playing the Wrong PART: The Program Assessment Rating Tool and the Functions of the President's Budget [with commentary]. *Public Administration Review*, 72(1), 112–122.
- Wichowsky, A., & Moynihan, D. P. (2008). Measuring How Administration Shapes Citizenship: A Policy Feedback Perspective on Performance Management. *Public Administration Review*. <https://doi.org/10.1111/j.1540-6210.2008.00931.x>
- Wojciech Kopczuk & Cristian Pop-Eleches. (2005). *Electronic Filing, Tax Preparers, and Participation in the Earned Income Tax Credit*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv:1910.03771 [Cs]*.
<http://arxiv.org/abs/1910.03771>
- Yang, K., & Yi Hsieh, J. (2007). Managerial Effectiveness of Government Performance Measurement: Testing a Middle-Range Model. *Public Administration Review*.
- Zamora, P., & McNeil, P. (2012). *Government Efficiency and the GPRA Modernization Act*. Nova Science Publishers, Inc. <http://proxy-ub.researchport.umd.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsebk&AN=1441496&site=eds-live>

