Artificial Intelligence and Administrative Evil

Matthew M Young
Syracuse University

Johannes Himmelreich
Syracuse University

Justin B Bullock
Texas A&M University

Kyoung-Cheol Kim
University of Georgia

Abstract

Artificial intelligence (AI) offers challenges and benefits to the public sector. We present an ethical framework to analyze the effects of AI in public organizations, guide empirical and theoretical research in public administration, and inform practitioner deliberation and decision-making on AI adoption. We put forward six propositions on how the use of AI by public organizations may facilitate or prevent unnecessary harm. The framework builds on the theory of administrative evil and contributes to it in two ways. First, we integrate the theory of administrative evil with agency theory. We examine how the mechanisms stipulated by the former relate to the underlying mechanisms of the latter. Specifically, we highlight how mechanisms of administrative evil can be analyzed as information problems in the form of adverse selection and moral hazard. Second, we locate the causal pathways of the theory of administrative evil on multiple levels of analysis, including the individual (micro), organizational (meso), and cultural (macro) levels. We then develop both descriptive and normative propositions on AI's potential to increase or decrease the risk of administrative evil. The article hence contributes an institutional and public administration lens to the growing literature on AI safety and value alignment.

Keywords: Administrative Evil, Principal-Agent Theory, Artificial Intelligence, Decision Making, Public Sector Innovation.

Note: This is the penultimate draft of a paper published in *Perspectives on Public Management and Governance*, Volume 4, Issue 3, September 2021, Pages 244–258. The version of record can be found here: https://doi.org/10.1093/ppmgov/gvab006.

Introduction

Artificial intelligence (AI) comes with significant risks. In a famous thought experiment, an AI system that is tasked with making paperclips eventually converts all available matter, including all of humanity, either into paperclips or into machines that make paperclips (Bostrom 2014, 123). This thought experiment illustrates what is known as the value alignment problem:[1] When an AI is sufficiently powerful and built without clarity of purpose, then great harms may result. Although this thought experiment is absurd, it surfaces important technical and philosophical challenges (Russell 2019; Gabriel 2020). The pursuit of seemingly benign goals can produce bad outcomes. Such bad outcomes by public organizations are already occurring, for example in the form of algorithmic bias and discrimination in bail decisions, criminal investigations, hiring decisions, or the allocation of child welfare services (O'Neil 2016; Angwin et al. 2016; Eubanks 2018). As scholars of public administration, we are interested in the institutional aspect of the value alignment problem. How can insights from management studies and institutional design help to conceptualize this problem and prevent such bad outcomes?

To address these questions, we draw on well-established theories in public administration and in particular on the theory of administrative evil (Balfour, Adams, and Nickels 2020). The theory of administrative evil is inspired by Arendt's (1963) idea of the banality of evil: it is not the malice of the powerful few but instead the thoughtlessness of the ordinary many that leads to horrendous wrongs through the execution of mundane tasks. The theory of administrative evil builds on this idea by contending that the causes of evil are often structural. Norms, conventions,

[1] A similar story is the myth of King Midas and the golden touch (Russell 2019, 136).

and a culture of "technical rationality" explain why evil occurs. By taking such a structural approach, the theory of administrative evil resembles current theories of sexism, misogyny, and racism. We combine the theory of administrative evil with agency theory and apply both to the case of AI. From this perspective, we analyze AI as masking or un-masking administrative evil or as increasing or decreasing agency costs associated with the adverse selection and moral hazard problems endemic to principal–agent relationships.[2]

We proceed in three steps. First, we review the theories of administrative evil, organizational decision making, and agency theory. Second, we develop six propositions of how AI may increase or decrease the chances of administrative evil through factors on the micro, meso and macro level of an organization. Third, we provide directions for further empirical and theoretical research and give recommendations for public administration practitioners. As our main practical implication, we caution against the cavalier or spurious use of AI in the public sector. AI implementation carries significant risk for increasing administrative evil.

Administrative Evil

"Evil" can be defined as unjust or unnecessary suffering that humans or human organizations, intentionally or unintentionally, inflict on other humans.[3] Evil can occur on a small scale: a wrongfully denied application for temporary assistance for needy families may count as evil using

[2] To be clear, we do not impute moral agency to AI. For our purposes here, we can assume that AI has no values "of its own."

[3] The theory of administrative evil resists any explicit definition of the term "evil" (Balfour, Adams, Nickels 2020, 3). Our own definition is compatible with how the theory of administrative evil implicitly understands the term. First, similar to the theory of administrative evil, our definition casts "evil" more narrowly compared to the theological problem of evil, which understands "evil" as meaning any bad state of affairs – including earthquakes and pandemics. Second, our definition casts "evil" more broadly compared to definitions which reserve the term exclusively for particularly egregious harms, heinous deeds, or deep injustice (Calder 2018).

this definition. Administrative evil, as investigated by Balfour et al. (2020), is a special form of evil in three ways.

First, administrative evil is typically the product of organization-level behavior. Organizations, according to the theory of administrative evil, can be understood broadly as including national governments (e.g., Nazi Germany), or narrowly as individual agencies or their branches (e.g., NASA). On the level of individuals, administrative evil is furthered by the individual tendency to comply, carry out routine tasks, and yield to authority (March and Simon 1993).

Second, administrative evil is facilitated by technology in that technology worsens individual tendencies to comply and yield to authority (Balfour, Adams, and Nickels 2020, xvi). The theory of administrative evil refers to this as a culture of *technical rationality*, that is, "a way of thinking and living that elevates the scientific-analytical mindset and belief in technological progress over all other forms of rationality"; technical rationality reduces ethical and normative concerns in favor of efficiency and expertise (Balfour, Adams, and Nickels 2020, 28). Similar phenomena have recently been studied elsewhere (Glikson and Woolley forthcoming), often as a form of "automation bias" (Cummings 2004). In the literature of public management and on AI specifically, related concerns have been highlighted under the heading of "artificial discretion" (Young, Bullock, and Lecy 2019).

Third, administrative evil invites self-reflection. The theory contends that "the pathways to administrative evil … most often emanate from within, ready to coax and nudge any professional down a surprisingly familiar route: first toward moral inversion, then to complicity in crimes against humanity" (Balfour, Adams, and Nickels 2020, 8). The central observation that administrative evil "emanates from within" invites those who are within – scholars and practitioners – to reflect critically on possible causes of evil in their organizations or how they themselves might be

involved in evil. Such deliberate self-reflection is required because administrative evil may not be salient insofar as it results from a gradual, continuous process and not a discrete choice.

The theory of administrative evil postulates two key explanatory concepts: masking and moral inversion. *Masking* refers to mechanisms that make administrative evil difficult to recognize as such. Evil can be masked either by the use of euphemisms or obfuscatory jargon but also through psychological or physical distancing. A current example of masking is the US immigration system where the "DHS takes pains to say that [the sites where migrants are confined] are 'detention centers,' 'servicing processing centers,' or 'residential centers' – anything but jails or prisons" (Hernández 2019). Evil is unmasked when a critical inflection point is reached and evil unambiguously needs to be recognized as such.

Moral inversion refers to the misrepresentation or misperception that what is evil is actually good. Moral inversion is typically facilitated through dehumanization, or more generally, by "portray[ing] the victims as deserving of their treatment" (Balfour, Adams, and Nickels 2020, 18). A glaring example of moral inversion can again be found in Nazi Germany, which promoted an ideology that dehumanized members of nameable groups, stigmatized individuals as deserving of their death, and that framed a genocide as an exercise in service of racial purity and hygiene. A more recent example of moral inversion is the mistreatment of refugees and asylum seekers. In many countries throughout the world, including the US, refugees are criminalized to promote national security or cultural unity. Because their immigration is made a crime, they are portrayed as deserving of their treatment in pursuit of moral goods such as security and unity. A more subtle example is the NASA Challenger disaster, where the obdurate rigidity of deadlines led decisionmakers to systematically discount warnings by safety engineers in order to salvage a waning reputation, at the expense of astronauts' lives.

Despite concentrating on often extreme examples, the theory of administrative evil, as reflected in our definition of "evil," is applicable to mundane and daily tasks – recall its ancestry in the banality of evil. We contribute to the development of the theory of administrative evil and its two key explanatory elements of masking and moral inversion by analyzing their mechanisms through the lens of agency theory.

Agency Theory and Administrative Evil

Agency theory models the relationships between actors – individuals, organizations, etc. – that stand in a vertical division of labor: One actor (the principal) requires the other (the agent) to perform one or more tasks.[4] This approach is particularly useful for analyzing both intra-organizational hierarchies and inter-organizational contracts. Agency theory is often used in public administration to understand the relationship between public managers and third parties contracted to provide public services (Lambright 2009; Eisenhardt 1989; Brown, Potoski, and Van Slyke 2015). It is also employed to model the relationship between elected politicians and the administrative bureaucracy in general (Epstein and O'Halloran 1994; Gailmard 2002; Bertelli and Lynn 2006).

Of central concern for our purposes are agency theory's twin problems of hidden information and hidden action (Arrow 1984). Hidden information, or information asymmetry, arises because agents have more information about their capabilities than principals do. Hidden action, or moral hazard, arises because agents' incentives do not perfectly align with their principals', and so the agent's level of effort is lower than what the principal desires or expects. In both cases, the

[4] This only illustrates a dyadic, vertical division of labor. Extensions to agency theory include both multiple agent and multiple principal scenarios.

fundamental problem is either a lack of information on the part of one party, or a misalignment of goals and values between parties. Solutions to these problems include both *a priori* protection via more complete contracts between principals and agents, and *in situ* measures to reduce information asymmetries in the form of agent monitoring (e.g., performance measurement and management) (Bertelli 2012).

Classic approaches to agency theory are unidirectional in their understanding of these problems: Agents are uniformly motivated via rational self-interest and preferences that do not align with their principal's, and agents always possess more information about their capacity and actions.

We instead adopt an alternative understanding of agency theory that relaxes these assumptions. This understanding allows for information asymmetries in which principals have more information about their desired objectives than their agents, and goal and value misalignments can arise from "honest incompetence" or other factors besides agent opportunism (Hay 2004; Hendry 2002; Kauppi and van Raaij 2014). Allowing for agency problems to flow in both directions makes them useful for understanding the concepts of masking and moral inversion in administrative evil.

In agency theoretical terms, masking can be understood as a problem of hidden information: either principal or agent (or both) is incapable of identifying administrative evil because they do not possess sufficient information.[5] Moral inversions can occur both through information asymmetries – agents do not have the information necessary to recognize the inversion – and through moral hazard in the form of "honest incompetence," where agents (or principals) simply

[5] Masking can concern different forms of information. Specifically, masking can concern information about an expression's referent as well as its sense. Euphemisms and obfuscating jargon convey less sense-information even if the reference-information is the same. "Processing facility" and "prison where children are separated from their parents" may refer to the same place but they convey very different senses.

do not possess the rational capacity to recognize the inversion (Hendry 2002). This recasting administrative evil's causal pathways in agency theoretic terms of information acquisition and processing allows us to map the potential use of AI by public organizations to these organizations' risk of administrative evil.

Organizational Decision Making and Technology

Decision making in organizations presents two long-standing challenges: delegation and discretion.[6] The challenge of *delegation* consists in principal–agent problems as just described. On the one hand, the agent is expected to carry out the political principal's will. On the other hand, the agent is driven by her or his own individual values and interests, including a need for status, recognition, or compensation (Selznick 1948). The agent has private information that the principal lacks. Delegation hence leads to moral hazard and adverse selection.

A second challenge of decision making in organizations is that of *discretion*. Administrative decision makers may have significant leeway in applying statutes and policies that are incomplete (Huber and Shipan 2002). Street-level bureaucrats often need to choose between allocations. For example, a NSF program officer needs to decide whether to recommend a research proposal for funding, and an employee in a local department of social services needs to decide whether a client is eligible for SNAP and about the length of their certification period. Such decisions are rooted in an assessment of rich facts and complex values (Simon 1997). That decision makers need to make normative and factual determinations in order to arrive at a decision, given that the factual and normative basis is incomplete, constitutes the challenge of administrative discretion.

---

[6] This glosses over many aspects of existing scholarship for the sake of parsimony. Decision making in public organization has been examined from various perspectives, and the debates have generated different traditions of public administration scholarship. One important aspect that we leave out concerns the difficulty of using and dispersing the rich information that agents acquire on the front-line of service.

Technology, specifically information and communication technologies (ICTs), have long been identified as a solution for minimizing information-based harms in bureaucracies, including those arising from delegation and discretion (Berry, Berry, and Foster 1998; Bozeman and Bretschneider 1986; Moon and Bretschneider 2002). On delegation, ICTs reduce information asymmetries insofar as they increase data generation, bandwidth, and capacities for data analysis. This decreases agency costs; in this case, the costs of information both for the agent to advance organizational objectives, and for the principal to observe and monitor agent behavior. On the issue of discretion, ICT tools may enhance or restrict discretion (Busch and Henriksen 2018; Buffat 2015; Bullock 2019). The impact of an ICT tool on discretion and thus organizational outcomes depends upon both the characteristics of the ICT tool itself and on characteristics of the task the ICT tool is given to complete, such as the task's complexity or risk (Buffat 2015; Young, Bullock, and Lecy 2019). Generally, ICT tools reduce information asymmetries by lowering search costs, insofar as they allow bureaucrats to obtain more information and give managers new ways to monitor agent behavior and performance.

Although technology might help with some aspects of the problems of delegation and discretion, it may introduce new problems of its own. For example, service provision may degrade as front-line bureaucrats interact with people remotely. Street-level bureaucracy might be turned into "screen-" or "systems-level" bureaucracy (Fountain 2001; Bovens and Zouridis 2002; Busch and Henriksen 2018). The transcription of phenomena into the machine-readable, quantitative data required for most ICTs also creates a tradeoff between nuance and simplicity. This risks the problem of "missing the forest for the trees" on the part of both the agent performing their duties and principals reviewing their work. Directly flowing from this problem is the example of goal displacement, where the pursuit of intermediate goals – that are often easily quantified and monitored via

ICT – leads to behavior that undermines the attainment of more fundamental goals (Lavertu 2016; Moynihan 2008). Goal displacement, in turn, increases the risk that organizational goals and values become confused and misunderstood by individuals, which necessarily increases the scope and magnitude of value considerations required for decision making.

Artificial Intelligence Across the Levels of Administrative Evil

In this section, we identify factors that might positively or negatively contribute to administrative evil at the individual (micro), organizational (meso), and institutional or cultural (macro) levels of analysis. We bring together three hitherto disparate literatures – administrative evil, agency theory, and on the use of ICTs in administrative decision making, as just reviewed – to elaborate six propositions on how AI may influence the occurrence of administrative evil (see Table 1 for an overview). For analytical simplicity and parsimony, our analysis assumes that public organizations have complete property rights, including source code access, to all AI systems they use.

Table 1: Propositions of AI's propensity to increase or decrease the risk of administrative evil by unit of analysis.

Proposition	Unit of Analysis			Risk of Administrative Evil
	Micro	Meso	Macro	Increase (↑) or Decrease (↓)
Descriptive: Amount and quality of available information				
Technical Inscrutability		✓		↑
Harm Discovery	✓	✓		↓
Quantification Bias		✓	✓	↑
Normative: Attitudes and values				
Control Centralization	✓	✓		↑↓
Organizational Value Misalignment		✓		↑
AI Exuberance	✓	✓	✓	↑

Before setting out this framework, we should define key terms. We understand AI as rational decision-support or decision-making systems that employ machine learning techniques to nondeterministic and domain-specific tasks (Young, Bullock, and Lecy 2019). We understand decisions as "rational" in the sense of bounded rationality and tasks as "nondeterministic" in the sense that outcomes are uncertain (Simon 1997). These definitions imply that AI can be deployed to perform administrative tasks (Drexler 2019; Bullock 2019) and augment or even replace human decision making, discretion and judgment (Young, Bullock, and Lecy 2019).

In distinguishing the micro, meso, and macro levels, we largely follow the definitions proposed by Jilke et al (2019). Public administration scholars have long grappled with the question of what unit of analysis is most appropriate or salient for the field (Wilson 1887; Simon 1946; Dahl 1947). More recent discussions adopt a more ecumenical approach, acknowledging the value of

different units of analysis, while also noting concern over the risk of possible epistemic fragmentation (March and Simon 1993; Moynihan 2018; Roberts 2020). We believe that both administrative evil and the effects of AI adoption occur across all three commonly understood levels of analysis, and structure our analysis accordingly.

The micro level has as its unit of analysis individual-level attributes such as the psychological mechanisms and dispositions of an individual bureaucrat or the amount of discretion that a bureaucrat enjoys in making decisions. Importantly, we do not count those effects that AI decisions have on individuals who are outside of the organization as micro-level phenomena. We instead focus explicitly on within-organization micro-level effects for the purpose of this article.

At the meso level, the unit of analysis are organizations. Public organizations often consist of several semi-autonomous sub-organizations (e.g., the US Army and US Navy nested within the Department of Defense, or public works and police departments nested within a municipal government) and AI affects these nested organizations analogously. Examples of factors located on the meso level are the depth of an organizational hierarchy, corporate processes to detect and address misconduct, as well as the collective decision making concerning the adoption of AI. Most importantly, we also count the operation of AI systems themselves as located on the meso level. We locate AI on the meso level for three reasons. First, decisions to use AI are made for entire organizations or sub-organizations. Second, AI systems are developed and maintained collectively. Finally, responsibility for misconduct due to the AI tends to be held jointly by a team. In this sense, an AI system is a joint product.

On the macro level, we concentrate on the social and political-administrative environment and include as units of analysis social norms, predominant cultures, or ideologies. These elements are particularly important for the theory of administrative evil. Balfour et al. emphasize, among

other things, that substantive values have been replaced with procedural ones in the theory and practice of public administration. Specifically, they argue that "procedural correctness and efficiency can mask both the context in which they are applied and the human consequences of administrative action" (Balfour, Adams, and Nickels 2020).

When evaluating whether AI will increase or decrease the chance for administrative evil to occur our baseline for this comparison is the status quo and not ideal types (Young, Bullock, and Lecy 2019). On the micro level this would be the median human bureaucrat, on the meso level a representative organizational structure and on the macro level the political or administrative culture in a region.

Micro-level Factors: Individual Agents and their Attributes

At the micro level, how likely administrative evil arises from the use of AI depends on individuals and their attributes. Specifically, the probability of administrative evil depends on the degree of discretion that an individual decision maker possesses, the individual's relative propensity for committing or resisting evil acts, and the individual's relative propensity for engaging in satisficing or defaulting to heuristics.

We concentrate on a class of decisions that are typical for public organizations, namely decisions in which individuals bring both evidence and values to bear and need to exercise discretion. Such decisions are taken with limited information, under time pressure, and subject to the constraints of bounded rationality (Simon 1997). Even within public organizations, decisions that allow for discretion are often made on the basis of decision makers' personal beliefs, their perception of the broader public's beliefs, or both (O'Leary 2013; Zacka 2017; Lipsky 1980).

We identify three ways in which AI may change the probability of administrative evil. AI will decrease administrative evil through harm discovery, it may increase or decrease administrative evil though control centralization, and AI will increase administrative evil because of AI exuberance.

First, AI's well-known analytic capacities could help unmasking administrative evil. AI can recognize patterns of correlation that are otherwise unrecognizable to humans. To the extent that administrative evil is due to a masking effect of complexity on individual decision makers, then AI's analytic capacity makes it a potential antidote when used as a decision support tool.

To illustrate this, suppose a social services organization uses an AI system to evaluate performance (however defined). The tool estimates changes to individual and social welfare by drawing on a vast array of disparate data and taking into account both positive and negative externalities. The AI system allows an individual agent to uncover previously undetected negative externalities associated with, for example, means-testing eligibility thresholds that in aggregate are so large that they constitute an evil despite the cause being rooted in the desire for programmatic efficiency and effectiveness. Thus,

> **Harm Discovery:** AI can decrease the chance of administrative evil by improving the information available to agents about potential harmful consequences associated with the decision, for example by revealing new correlations between decisions and outcomes.

At the same time, the interplay between AI and an individual's capacity for collecting and weighing evidence may also *increase* the risk of administrative evil. Individuals are only boundedly rational, and often resort to the use of heuristics and other satisficing – that is "good enough" – choice processes, which are subject to various biases (Tversky and Kahneman 1974; Simon 1997). Specifically, when individuals interact with technology, they suffer from what is known in

computer science and industrial psychology as *automation bias*: when individuals share control with an automated decision support tool, they become overly-reliant on the tool and do not critically review its recommendations or behaviors before taking action or making a decision (Manzey, Reichenbach, and Onnasch 2012; Mosier et al. 1996).

For example, consider again our hypothetical social services organization but now with a different AI implementation. Suppose AI is used to help individual street-level bureaucrats determine whether applicants are eligible for benefits, and if so at what level. Over time, the individual can become lulled into a false sense of security that the AI will make the correct recommendation, particularly if it is highly accurate under most circumstances. But when an applicant arrives whose circumstances constitute an edge case that the AI is not calibrated to handle properly, the tool-using bureaucrat may fall victim to automation bias and deny the applicant's claim when they are in fact eligible. In agency theoretic terms, in this case the bureaucrat is the principal, the AI-enabled decision tool is the agent, and the difference between the AI's true error rate and its presumed or reported error rate constitutes an information asymmetry between the parties. This suggests in particular the first point of

> **AI Exuberance:** AI can increase the chance of administrative evil because it (1) introduces the risk of automation bias; (2) thrives within the cultural ideology of technological rationality; and therefore (3) may be enthusiastically deployed without appropriate testing or to address an issue for which AI is not the optimal available solution.

A third effect of AI on the micro level is that it typically reduces the discretion that is afforded to individuals. Technology in general, and ICTs and AI in particular, shift the locus of control away from front-line staff and street-level bureaucrats (agents) towards management (organizational principals) by curtailing the former's ability to exercise discretion (Garson 1989;

Busch and Henriksen 2018). This is part of a more general tendency of automation. AI can perform increasingly complex tasks and it thereby threatens an increasing share of white-collar professional jobs (Frey and Osborne 2017; Lee 2016). This shift of control away from individual agents and towards principals will likely affect the risk of administrative evil in a nuanced way, in that this risk is likely jointly determined by the attributes of agents and principals. If agents have a greater disposition to commit evil than their principals, then this shift will likely decrease the risk of administrative evil, and *vice versa*.

Suppose first that individual agents are more disposed to contribute to evil. This individual disposition can be attributed to prejudice, malice, or other malfeasance on the individual's part. But as the theory of administrative evil suggests, this disposition can also arise from masking, that is, that the agent has incomplete information or is unable to correctly parse the information they have. Examples of this sort of evil-through-misfeasance include decision biases introduced from organizational loyalty, aversion to interpersonal conflict, and the bounded nature of human rationality in general. AI systems do not share these risk factors. Curtailing the discretion of individuals who are disposed to contribute to evil hence reduces the risk of administrative evil.

But often the principal might instead be the source of the ethical problem. More precisely, agents may instead be less disposed to contribute to evil compared to principals. Human decision making is sensitive to values and individual discretion can be motivated ethically. In result, individual decisions need not comport with an organization's explicit tasks and goals. Such a conflict between individual values and organizational goals poses a dilemma for human agents who then can respond in several ways: they may accept the organization's goals (loyalty), raise objections (voice), resign in protest (exit), or sabotage the effort in different ways (neglect) (Hirschman 1970; O'Leary 2013; Rusbult et al. 1988). AI agents, in contrast, do not have these options. Hence, AI

may increase the risk of administrative evil when it limits the discretion of agents disposed to prevent or resist evil. Thus,

> **Control Centralization:** AI consolidates discretion at higher levels of organizational hierarchy and thereby moderates the risk of administrative evil; the direction of its moderation is a function of whether organizational leadership is more or less predisposed to commit evil than street-level staff.

A related effect, which we also subsume under the *Control Centralization* proposition, is that AI increases psychological distancing because they mediate between agents and those affected by their decisions. If an agent is otherwise disposed to prevent administrative evil, AI systems may undercut their disposition by increasing the psychological distance between the agent and those who are subject to the effects of their decision. In terms of the theory of administrative evil, this is another form of masking.

The nature of AI makes it particularly dangerous in this regard. AI decision processes fundamentally reduce to in-group/out-group classifications: a given input is evaluated for whether it meets the conditions for inclusion in an arbitrarily defined set. Even if an AI-generated risk score is not binary, it will still drive a binary decision of whether to reject or accept an applicant, grant bail, deliver a service, or even kill a human. This closely maps to the psychological phenomenon of othering, where individuals justify harms imposed on other human beings by recategorizing them as something 'other' than human. Thus, even when some discretion remains in the hands of individuals normally inclined to do good, AI can still increase the risk of administrative evil.

In sum, on the micro level, AI may increase the risk of administrative evil through *AI Exuberance*, it may decrease the risk of administrative evil through *Harm Discovery* and it may increase or decrease the risk of administrative evil through *Control Centralization* depending on the relative dispositions of current human agents and their principals to prevent or facilitate evil.

Meso-level Factors: Organizational Structural Factors and Relational Dynamics

A comprehensive theoretical framework needs to consider the role of principals and the organization as a whole and cannot restrict attention to agents and the micro level. AI shifts the locus of control away from individual agents and towards higher levels of an organizational hierarchy and the *Control Centralization* proposition hence describes effects on the micro and on the meso level (see Table 1 on page 10).

An organization's leadership is important for how the introduction of AI contributes to administrative evil, for example, because an organization's leadership decides whether or not to adopt new technologies like AI. The adoption of AI in public organizations can take on several different forms with respect to degree of automation, associated task complexity, and level of risk or uncertainty. The preferences of organizational leadership can hence have a significant effect on what form this adoption takes in practice (Young et al 2019). Analogous to the process on the micro level, the dispositions of an organization's leadership conditions whether the adoption of AI increases or decreases the chances of administrative evil.

For example, consider an agency such as law enforcement where street-level staff are entrusted with a high level of discretion and power over civilians. If senior management and street-level staff are equally susceptible (or resistant) to the conditions that foster administrative evil, there is no clear reason to believe the organization's use of AI will affect the baseline hazard rate in either direction. But because AI can crowd out their human counterparts via automation, it is possible that AI use could increase the risk of administrative evil when the primary resistance to such risk comes from the values and actions of its street-level agents. But when, inversely, street-level staff are most at risk of committing administrative evil, then the use of AI by a leadership committed to avoiding administrative evil should, on average, reduce this risk. Because the effect can be driven both by dispositions of organizational leadership as well as by dispositions of street-

level agents, the proposition of *Control Centralization* operates at the meso as well as the micro level.

Although we hypothesize that the adoption of AI centralizes control – because it centralizes decision-making power – the adoption may yet reduce control in the specific sense that AI are difficult to interpret and that their decisions can be hard to explain. One of AI's principal advantages over human agents is its ability to analyze large, high-dimensional, and complex data. Unfortunately, this advantage comes with an inherent tradeoff: As its ability to process complex data increases, an AI system becomes harder to audit.[7] The most powerful AI are often the least understandable with respect to their decision-making process *ex post* (Weld and Bansal 2019). This suggests that information-based agency problems are likely to materialize in a way that may lead to administrative evil. The organization can be seen as the principal, the AI system as the agent and the lack of interpretability as an information asymmetry.

For illustration, consider the implementation of an AI system in our hypothetical social services organization where the system monitors the behavior of disability benefit recipients to detect fraud. But insofar as the AI system is not interpretable, and that automation bias can lead many in an organization to implicitly trust machine judgement as superior to human judgement (Hoff and Bashir 2015; Parasuraman and Manzey 2010; Dzindolet et al. 2003), classification errors become increasingly difficult to rectify (O'Neil 2016; Eubanks 2018). Type I Errors, that is false positives where an innocent benefits recipient is identified as engaging in fraud, are bound to occur but might be less likely to be recognized and rectified. Analyzed through the lens of agency theory,

[7] For example, an AI system based on a decision tree algorithm generates decision outputs that are relatively easy to understand *ex post*; one can navigate the various decision branches and their associated weights to understand the underlying logic. However, decision tree-based models are poorly optimized for dealing with more complex data, especially when the potential correlates between data points are not known *a priori*.

this is a problem of moral hazard: the agent has behaved in a way that is at odds with the principal's stated objectives, but information asymmetries make it costly for the principal to notice or understand the discrepancy. Thus,

> **Technical Inscrutability:** AI can increase the chance of administrative evil because it masks decision making (due to increased complexity and decreased transparency): AI lacks explainability or requires technical expertise to understand decisions.

This form of moral hazard and its implications for administrative evil also contribute to the problem of goal displacement. Goal displacement occurs when organizations pursue intermediate and easily measurable goals instead of their originally intended but relatively unmeasurable goals. This often inadvertently leads to worse performance in terms of the originally intended goals (Lavertu 2016; Moynihan 2008). More relevant for our purposes, goal displacement may lead to administrative evil, even if only to minor forms of evil. Suppose an AI system is tasked with assigning benefit recipients with the appropriate benefit amounts. Suppose further that the system is designed to set the lowest benefit rate that meets statutory obligations.[8] For some benefit recipients, the AI agent may identify a positive correlation between lower immediate benefit rates and greater long-term cost savings, and it may thus begin to systematically assign more recipients lower average benefit amounts when the system could have used administrative discretion to assign increased benefits. But suppose that the long-term cost savings are in part attributable to the fact that these recipients died earlier than they would have if they had received increased benefits. In pursuit of the intermediate and more easily measurable goal of cost savings over time, the AI system minimized benefits levels as low as legally possible by indirectly maximizing recipient's mortality. In

[8] This is not uncommon in human-facilitated social benefit calculation, particularly in the anglosphere.

fact, AI agents demonstrate goal-displacing behavior to the extent that the field of AI research has its own associated terminology: *reward hacking* (Amodei et al. 2016). Thus,

> **Organizational Value Misalignment:** AI can increase the chance of administrative evil by increasing organizational goal displacement.

This hypothetical situation highlights how AI systems can facilitate administrative evil through goal displacement. The AI system contributed to the untimely death of benefit recipients by executing on its intermediate goal of minimizing costs. This problem is exacerbated by AI's singular capacity to identify complex correlations in high-dimensional data and its fundamentally *inhuman* reasoning.

However, this same characteristic of AI systems that may facilitate administrative evil can also be leveraged to combat it – when organizations design and implement AI systems properly. For the hypothetical social service organization, this use of AI would be tailored towards identifying the same correlates that led to goal displacement in the previous example, but this information would then be used to adapt and refine organizational decision making and processes to minimize or eliminate the risk of administrative evil in the form of contributing to the premature death of benefit recipients. As with *Control Centralization*, this leads us to propose that the proposition of *Harm Discovery* also operates at both the meso and micro levels:

> **Harm Discovery:** AI can decrease the chance of administrative evil by improving the information available to agents about potential harmful consequences associated with the decision, for example by revealing new correlations between decisions and outcomes.

Yet even when an organization's decision to adopt AI is made with the best intentions – indeed, irrespective of the intention altogether – the risk of administrative evil can increase from

AI adoption because AI can introduce or intensify decision-making biases that mask administrative evil and facilitate moral inversion. For example, a necessary precondition for the proposition of Harm Discovery to hold is that there are sufficient quantitative (and more specifically, machine-readable) data available for AI to identify the appropriate patterns correctly and reliably. Despite the exponential growth of ICTs in the late 20th and early 21st centuries, which has made AI increasingly capable and useful, many forms of data and information remain unquantifiable and, therefore, unusable by AI.

When these unquantifiable data are also the primary subject of interest, organizations set on using quantitative approaches like AI must use whatever quantifiable data they can find – often referred to as "proxy variables" – even if they are a less accurate representation of the phenomenon of interest. This tradeoff necessarily introduces the risk of both masking administrative evil and facilitating moral inversion. Proxy variables may mask administrative evil when they are related to primary outcomes in a biased way or when they make harmful algorithmic bias harder to detect (Johnson forthcoming). The extent to which using proxy variables is problematic is usually extremely difficult to measure; if it were easier there would be no need for proxy variables in the first place.

Consider the example of health. Health cannot be directly measured but needs to be operationalized. The severity of chronic conditions of a patient – one important constituent of health – are usually measured based on the medial expenditures caused by this patient (not paid by them) over a given time. Recent research has shown that this proxy variable for health masks a racial bias (Obermeyer et al. 2019). A risk-scoring AI system is used by hospitals to determine which patients with chronic conditions should be included in a special treatment program. This AI system predicted health in terms of medical expenditures accurately and fairly. But the AI system was

biased in that African American patients were much less likely to be suggested for inclusion in the special treatment program than white Americans with the same underlying chronic conditions. This bias in the prediction was due to the fact that how medical expenditures are accrued differs between racial groups. An African American patient generally accrues fewer medical expenditures compared to a white American with the same underlying chronic conditions. Unfortunately, underlying chronic conditions are difficult to measure and to aggregate (because of the highly sensitive nature of the data). Hence medical expenditure was used as a proxy variable. Thus,

> **Quantification Bias:** AI can increase the chance of administrative evil by reducing the amount or quality of data brought to bear for a decision: AI requires and reinforces a belief in the primacy of quantitative data that excludes other forms of information unless they can be readily and systematically quantified.

The underlying motivations that drive organizations to use second-best proxy variables also expose them to a more generalized risk of facilitating administrative evil. This occurs when they choose to use AI although it might not be the best available option, or when the use of AI introduces new harms or magnifies existing ones. Such an over- or mis-use of AI in pursuit of organizational objectives is more likely to take place when the organization identifies AI as a powerful and useful tool without deep understanding of how the technology works. The technology might then not fit a given task. Similarly, an organization may not properly understand the context of the task and still rely on AI despite this lack of contextual understanding. Both of these issues can be further exacerbated when the organization faces either normative or mimetic isomorphic pressure from its peer or neighbor organizations to adopt AI solutions (DiMaggio and Powell 1983; Jun and Weare 2011).

Finally, organizations might adopt AI solutions without assessing potential risks. For example, facial recognition systems are widely used despite varying significantly in accuracy depending on the shades of one's skin (Buolamwini and Gebru 2018). Nevertheless, classifications of such facial recognition systems have already been used to make false arrests (Hill 2020). Both – the tendency to adopt AI despite problems and to yield to the judgments of AI systems – are key symptoms of *AI Exuberance*.

> **AI Exuberance:** AI can increase the chance of administrative evil because it (1) introduces the risk of automation bias; (2) thrives within the cultural ideology of technological rationality; and therefore (3) may be enthusiastically deployed without appropriate testing or to address an issue for which AI is not the optimal available solution.

In sum, on the meso level, AI may increase the risk of administrative evil through *Quantification Bias*, *Organizational Value Misalignment*, *Technical Inscrutability* and *AI Exuberance*. AI may decrease the risk of administrative evil through *Harm Discovery* and it may increase or decrease the risk of administrative evil through *Control Centralization* depending on the relative dispositions of organizational leadership.

Macro level: Cultural

The theory of administrative evil identifies two cultural patterns as contributing to administrative evil: a culture of technical-rational problem solving and a "scientific-analytical mindset." We argue that AI may worsen the contribution of each of these cultural patterns to administrative evil.

Technical-rational problem solving refers to the tendency of framing and implementing policy as solution to problems. The theory of administrative evil objects to this that many social issues "transparently do not fit this image of discrete problems that can be solved once and for all

with analytic methods (Balfour, Adams, and Nickels 2020, 87)." One example is poverty. While reducing poverty may be a laudable goal, poverty is a complex, multi-faceted, enduring feature of the human condition. There has never been a moment when poverty was "solved." The realities of the condition of poverty simply do not easily reduce to an analytically-derived policy solution. This pattern of technical-rational problem solving is a form of masking. Pressing issues risk not being recognized correctly.

This enables administrative evil in two ways. First, technical-rational problem solving enables administrative evil through inaction. Technical-rational problem solving encourages a certain blindness towards any issue that cannot be framed as problem and any intervention that cannot be framed as a solution. For example, without a clear solution and elimination of the condition of poverty, poverty itself becomes a less troubling concept. Without a clear solution, inaction dominates. Second, as observed for a similar mechanism on the meso level, technical-rational problem solving enables administrative evil when a purported solution is deployed that is inappropriate for the issue at hand.

AI systems are perceived to be a powerful solution and may be easier to market and procure than non-AI alternatives. Moreover, because AI systems are perceived to be a very general solution that could help with many different kinds of problems, such systems will be considered in many different domains.[9] If AI is perceived to be more powerful and more general, then this could enable an insufficient attentiveness to the requirements of a given policy issue and thereby result in the deployment of technologies that are inappropriate or even defective given the task at hand. This is a further aspect of the *AI Exuberance* proposition.

---

[9] In this way, the perception of AI can be likened to the law of the instrument, which is often captured by the statement "if all you have is a hammer then everything you encounter looks like a nail."

In addition to the problem of masking, in terms of the theory of administrative evil, the *AI Exuberance* proposition identifies a problem of moral inversion. AI is framed as a powerful and effective new tool that symbolizes progress and expertise, when in fact it may be ineffective or even do harm. In a discussion that is immediately applicable to AI, the theory of administrative evil postulates that the invitation to do evil might come as an invitation for an expert role. AI in this case is the expert invited to do something only on the basis of perceived expertise, and their judgments are uncritically assigned credence on the same basis.

A similar critique has recently been put forth under the name of "solutionism" (Morozov 2014). Solutionism is a habit of thought in which individuals tend to seek technological fixes for social problems. Such technological solutions to social problems lack awareness of important social aspects in a way that threatens the success of the proposed solution. One prominent example of this is the initiative of one laptop per child. The one laptop per child initiative sought to remedy global poverty through education and was premised on the assumption that potential for educational attainment in the developing world can be unlocked by equipping every child with a laptop and internet access. This approach is considered to have failed (Kraemer, Dedrick, and Sharma 2009; Warschauer and Ames 2010).

The second cultural pattern that the theory of administrative evil identifies as enabling administrative evil is that of the "scientific-analytic mindset." In this diagnosis it references the work of Mannheim (1940) vis. "functional rationality," or Horkheimer (1947) vis. "instrumental reason". Framing something as scientific can be a form moral inversion to mask harm. Specifically, for the purposes of AI, we understand the "scientific-analytic mindset" as one that strives for and valorizes quantification and precise numerical measurement.

In addition to problems on the meso level that we already mentioned, quantification and measurement can lead to several problems on the macro level that are well explored in philosophy, science and technology studies (STS), and in the sciences themselves (Hausman, McPherson, and Satz 2016; Mesquita 2019). For example, categories such as health, crime and sex may be socially constructed. Measuring variables that are socially constructed is hence subject to a reflexivity and the measurement itself may interact with what is measured.

The "scientific-analytic mindset" may unjustifiably increase the belief that quantitative data – and proxy variables in particular – are a simple and unproblematic representation of an underlying feature that is, instead, complex, messy and vague. Because AI deals in quantification and numerical representation and feature engineering, it may increase the chance of administrative evil through masking and moral inversion.

Moreover, A lesson from feminist philosophy of science and the ethics of data science is that AI systems reflect subtle choices of models and measurement methods as well as empirical design (Johnson forthcoming). AI becomes an enabler of administrative evil when this lesson is disregarded and when, instead, the recommendations by AI become arguments in favor of policies only because they carry the predicate of being "scientific" or the alleged necessity of being "dictated by science."

We hypothesize that similar mechanisms of masking and moral inversion may be prominent in the case of AI, and that this is as much a cultural phenomenon as it is the product of any one decision to adopt and use AI. Whereas the overlooked lesson more generally is that it is wrong to think of something "dictated by science," in the case of AI the analogue of this idea is that it is wrong to think of something being "dictated by data." Thus,

> **Quantification Bias:** AI can increase the chance of administrative evil by reducing the amount or quality of data brought to bear a decision: AI requires and reinforces a belief in the primacy of quantitative data that excludes other forms of information unless they can be readily and systematically quantified.

In summary, on the macro level, AI may increase the risk of administrative evil in two ways – relating to how AI and how policy domains are seen. First, the *AI Exuberance* proposition rests on projecting AI as a powerful "intelligent" and general problem solver. This projection can be a form of masking and moral inversion concealing the harms that come from a careless or uncritical deployment of AI – evil is masked by how we see AI. Second, the *Quantification Bias* proposition expresses an over-valorization of numerical measurement and quantification. Quantification Bias restricts the domain of policy to tractable problems. Here, evil is masked by how we see the world. AI fits and may even entrench this quantitative outlook.

Discussion

The use of AI for administrative decision making has direct implications for the risk of administrative evil. To make this argument, we have extended the theory of administrative evil in two ways. First, we have clarified and detailed the causal pathways that the theory postulates in agency theoretic terms. Second, we located the relevant units of analysis on the individual (micro), organizational (meso), and cultural (macro) level.

We have formalized six propositions to advance our understanding of the effects that AI has in the public sector. Three of these propositions relate to the amount and quality of information available for decision making; in this sense, these three are descriptive propositions. The other three propositions relate to decision-making policies, values, and attitudes, and they are hence normative. Table 1 above identifies the units of analysis, that is, the levels at which we expect AI

to affect the risk of administrative evil in public organizations. Table 2 below lists all six propositions. This ethical framework, in addition to advancing our understanding of the effects of AI, may guide future empirical and theoretical research and it may inform deliberation and decision making.

Table 2: Summary of all propositions

Descriptive: Propositions about amount and quality of available information
Harm Discovery: AI can decrease the chance of administrative evil by improving the information available to agents about potential harmful consequences associated with the decision, for example by revealing new correlations between decisions and outcomes.
Technical Inscrutability: AI can increase the chance of administrative evil because it masks decision making (due to increased complexity and decreased transparency): AI lacks explainability or requires technical expertise to understand decisions.
Quantification Bias: AI can increase the chance of administrative evil by reducing the amount or quality of data brought to bear a decision: AI requires and reinforces a belief in the primacy of quantitative data that excludes other forms of information unless they can be readily and systematically quantified.
Normative: Propositions about attitudes and values
AI Exuberance: AI can increase the chance of administrative evil because it (1) introduces the risk of automation bias; (2) thrives within the cultural ideology of technological rationality; and therefore (3) may be enthusiastically deployed without appropriate testing or to address an issue for which AI is not the optimal available solution.
Organizational Value Misalignment: AI can increase the chance of administrative evil by increasing organizational goal displacement.
Control Centralization: AI consolidates discretion at higher levels of organizational hierarchy and thereby moderates the risk of administrative evil; the direction of its moderation is a function of whether organizational leadership is more or less predisposed to commit evil than street-level staff.

These propositions point out avenues for empirical research. Whereas some propositions, such as *AI Exuberance*, may find some support in existing studies on automation bias, econometric approaches to model innovation diffusion could test for and improve our understanding of the risks of *AI Exuberance*. Moreover, propositions such as *Control Centralization* are squarely in the ballpark of public administration and ripe for empirical analysis. Lest we fail to heed our own warning about quantification bias, we should note that qualitative techniques like ethnographies or in-depth

case studies may be best suited for testing propositions like *Control Centralization*, *Organizational Value Misalignment*, or *Quantification Bias*. Experimental techniques, both with human subjects and through computer modeling, could bring insights from cross-disciplinary collaboration to bear on matters of public administration.

Assuming that these propositions are plausible, they carry implications for public administrators and may inform deliberation and decision making. Relating to the three levels of analysis, our propositions address a practitioner in one of three roles: as a bureaucrat and individual decision maker in a public organization (micro level); as a manager of a division or team in such an organization (meso level); or as a citizen of his or her country (macro level). In what follows we highlight the key practical upshots of our analysis.

Individual decision makers are likely going to see their discretion decreased as per the *Control Centralization* proposition. As AI takes over increasingly more tasks, the practical problem comes to the fore of how the role of the individual decision maker changes in response. Our *Harm Discovery* proposition suggests that a future role for individual bureaucrats might be to audit and oversee the operation of AI systems – potentially utilizing further AI systems. This opportunity to evolve the role of bureaucrats however likely requires large, systemic changes to the ecosystem of statutory regulation, training opportunities, and individual incentives.

Our analysis has the most wide-ranging practical implications for managers in public organizations. Quantification might lead to administrative errors and administrative evil as our *Quantification Bias* and *AI Exuberance* propositions suggests. Quantification in general and AI in particular bear risks as such and contribute to the risk of goal displacement and to the value alignment problem as our proposition of *Organizational Value Misalignment* suggests. Both of these propositions – *Quantification Bias* and *Organizational Value Misalignment* – lead to one practical

upshot: AI has important in-principle limitations. In a slogan: Not everything that *could* be done by AI *should* be done by AI. Moreover, to make AI effective and safe, organizations likely require a high clarity of purpose. Managers in public organizations need to deliberate and articulate in unambiguous terms the goals, objectives, and costs of using AI. The importance of this practical upshot is heightened by what we call the *Technical Inscrutability* proposition, which highlights the risks that administrative errors and evils go undetected. Our *Harm Discovery* proposition may partially mitigate this risk as AI has the potential to identify administrative evil that would otherwise go unnoticed. Public managers need to actively plan to leverage the opportunity to use AI for this purpose. Finally, our *Control Centralization* proposition emphasizes that the impact of individual decisions is likely to increase. Since the direction of this effect can go either way, this raises the bar for self-reflection and ethical standards applied to public managers.

In the role as a citizen, our analysis offers two upshots to practitioners. First, our *Quantification Bias* proposition recommends an acknowledgement that not all outcomes can be easily measured. Albeit necessary, this acknowledgement is likely uncomfortable as it subsequently raises deep methodological challenges for the factual basis of public deliberation: How else, if not by quantification, can we even begin to address social challenges rigorously? Second, our *AI Exuberance* proposition motivates an important practical upshot across all three levels. Be it in a decision-making role, as a manager in a public organization, or as a citizen, each individual needs to check themselves against the pitfalls of the automation bias. Our hope is that the pernicious effects of the quantification bias and of AI exuberance can be mitigated with sufficient self-awareness of this phenomenon. Given the high stakes associated with AI and administrative evil, mitigating AI exuberance is of utmost importance.

Conclusion

We first extend and then employ the theory of administrative evil to consider how the public sector's use of artificial intelligence may mitigate or create individual and social harms. The theory of administrative evil contends that bureaucratic systems subtly facilitate or cause evil because such systems are difficult to understand as a whole (the problem of masking), and because individual, organizational, and cultural factors may make harmful intermediate or final outcomes appear beneficial to their producers (the problem of moral inversion). Our contribution to the theory of administrative evil consists of clarifying its causal pathways both through the use of agency theory and by explicitly mapping said pathways to different units of analysis: individual attributes on the micro level; organizational factors on the meso level; and cultural factors on the macro level. We then develop both descriptive and normative propositions on AI's potential to increase or decrease the risk of administrative evil. These propositions include harm discovery, technical inscrutability, quantification bias, AI exuberance, organizational value misalignment, and control centralization.

While harm discovery highlights an opportunity for AI to decrease the likelihood of administrative evil, each of the following propositions highlights a causal pathway for AI to increase risk of administrative evil being committed by public organizations. *Ceteris paribus*, the use of AI by governments provides for decreases in transparency in decision making (technical inscrutability), increased reliance on quantitative data to the crowding out of other sources of information (quantitative bias), overreliance on AI as a tools even when it is dangerous or inappropriate (AI exuberance), misalignment of decision making values and organizational values (organizational value misalignment), and decisions that are centralized as to takeaway decision making authority

from professionalized bureaucrats (control centralization). These propositions, taken together, constitute a clarion call of warning: Government use of AI will likely facilitate administrative evil without significant interventions and oversight before, during, and after AI's adoption and implementation.

The ethical framework that we present here may guide empirical research and inform practitioners' decision making to adopt and use AI. On the whole, the propositions suggest that if public managers and administrators desire to avoid administrative evil, they should exercise extreme caution when considering whether and how to adopt and implement AI. In particular, public managers should prioritize systems that only use AI when tasks are clearly understood; decision inputs are diverse; decision inputs and outputs are transparent as possible; and include clear statutory and cultural accountability for mistakes and unintended consequences on the part of system architects and the leadership that authorized its implementation. If these cautions are taken, then public organizations may enjoy the benefits of clever applications of AI that improve effectiveness, equity, and efficiency, rather than perpetuating administrative evils.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *ArXiv:1606.06565*, 29.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arendt, Hannah. 1963. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press.

Arrow, Kenneth J. 1984. "The Economics of Agency. In Principals and Agents: The Structure of Business, Ed. John W. Pratt and Richard J. Zeckhauser, 1–35." Boston, MA: Harvard Business School Press.

Balfour, Danny L., Guy B. Adams, and Ashley E. Nickels. 2020. *Unmasking Administrative Evil*. 5th ed. New York: Routledge.

Berry, Frances Stokes, William D Berry, and Stephen K Foster. 1998. "The Determinants of Success in Implementing an Expert System in State Government." *Public Administration Review*, 293–305.

Bertelli, Anthony M. 2012. *The Political Economy of Public Sector Governance*. Cambridge University Press.

Bertelli, Anthony M., and Laurence E. Lynn. 2006. *Madison's Managers: Public Administration and the Constitution*. Baltimore: The Johns Hopkins University Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bovens, Mark, and Stavros Zouridis. 2002. "From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control." *Public Administration Review* 62 (2): 174–84.

Bozeman, Barry, and Stuart Bretschneider. 1986. "Public Management Information Systems: Theory and Prescription." *Public Administration Review* 46 (November): 475–487.

Brown, Trevor L., Matthew Potoski, and David Van Slyke. 2015. "Managing Complex Contracts: A Theoretical Approach." *Journal of Public Administration Research and Theory* 26 (2): 294–308. https://doi.org/10.1093/jopart/muv004.

Buffat, Aurélien. 2015. "Street-Level Bureaucracy and e-Government." *Public Management Review* 17 (1): 149–61.

Bullock, Justin B. 2019. "Artificial Intelligence, Discretion, and Bureaucracy." *The American Review of Public Administration* 49 (7): 751–61. https://doi.org/10.1177/0275074019856123.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html.

Busch, Peter André, and Helle Zinner Henriksen. 2018. "Digital Discretion: A Systematic Literature Review of ICT and Street-Level Discretion." *Information Polity* 23 (1): 3–28.

Calder, Todd. 2018. "The Concept of Evil." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/concept-evil/.

Cummings, Mary. 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems." In , 6313. https://doi.org/10.2514/6.2004-6313.

DiMaggio, Paul J., and Walter W. Powell. 1983. “The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields.” *American Sociological Review* 48 (2): 147–160. https://doi.org/10.2307/2095101.

Drexler, K. Eric. 2019. “Reframing Superintelligence: Comprehensive AI Services as General Intelligence.” 2019–1. FHI Technical Report. Oxford: Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.

Dzindolet, Mary T, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. “The Role of Trust in Automation Reliance.” *International Journal of Human-Computer Studies* 58 (6): 697–718.

Eisenhardt, Kathleen M. 1989. “Agency Theory: An Assessment and Review.” *Academy of Management Review* 14 (1): 57–74.

Epstein, David, and Sharyn O’Halloran. 1994. “Administrative Procedures, Information, and Agency Discretion.” *American Journal of Political Science* 38 (3): 697–722. https://doi.org/10.2307/2111603.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press.

Fountain, Jane E. 2001. *Building The Virtual State: Information Technology and Institutional Change*. Washington, D.C.: Brookings Institution Press.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” *Technological Forecasting and Social Change* 114 (January): 254–80. https://doi.org/10.1016/j.techfore.2016.08.019.

Gabriel, Iason. 2020. “Artificial Intelligence, Values, and Alignment.” *ArXiv:2001.09768 [Cs.CY]*, January. https://arxiv.org/abs/2001.09768.

Gailmard, Sean. 2002. “Expertise, Subversion, and Bureaucratic Discretion.” *Journal of Law, Economics, & Organization* 18 (2): 536–55.

Garson, Barbara. 1989. *The Electronic Sweatshop*. New York: Simon & Schuster.

Glikson, Ella, and Anita Williams Woolley. forthcoming. “Human Trust in Artificial Intelligence: Review of Empirical Research.” *Academy of Management Annals*. https://doi.org/10.5465/annals.2018.0057.

Hausman, Daniel, Michael McPherson, and Debra Satz. 2016. *Economic Analysis, Moral Philosophy, and Public Policy*. 3 edition. New York, NY: Cambridge University Press.

Hay, Colin. 2004. “Theory, Stylized Heuristic or Self-fulfilling Prophecy? The Status of Rational Choice Theory in Public Administration.” *Public Administration* 82 (1): 39–62.

Hendry, John. 2002. “The Principal’s Other Problems: Honest Incompetence and the Specification of Objectives.” *The Academy of Management Review* 27 (1): 98–113. https://doi.org/10.2307/4134371.

Hernández, César Cuauhtémoc García. 2019. *Migrating to Prison: America’s Obsession with Locking Up Immigrants*. The New Press.

Hill, Kashmir. 2020. “Wrongfully Accused by an Algorithm.” *The New York Times*, June 24, 2020, sec. Technology. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hirschman, Albert O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Vol. 25. Cambridge Mass.: Harvard University Press.

Hoff, Kevin Anthony, and Masooda Bashir. 2015. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” *Human Factors* 57 (3): 407–34.

Horkheimer, Max. 1947. *Eclipse of Reason*. Oxford: Oxford University Press.
Huber, John D, and Charles R Shipan. 2002. *Deliberate Discretion?: The Institutional Foundations of Bureaucratic Autonomy*. Cambridge: Cambridge University Press.
Jilke, Sebastian, Asmus Leth Olsen, William Resh, and Saba Siddiki. 2019. "Microbrook, Mesobrook, Macrobrook." *Perspectives on Public Management and Governance* 2 (4): 245–53. https://doi.org/10.1093/ppmgov/gvz015.
Johnson, Gabbrielle M. forthcoming. "Algorithmic Bias: On the Implicit Biases of Social Technology." *Synthese*. https://doi.org/10.1007/s11229-020-02696-y.
Johnson, Gabbrielle M. forthcoming. "Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning." *Journal of Moral Philosophy*, 23.
Jun, Kyu-Nahm, and Christopher Weare. 2011. "Institutional Motivations in the Adoption of Innovations: The Case of E-Government." *Journal of Public Administration Research and Theory* 21 (3): 495–519.
Kauppi, Katri, and Erik M. van Raaij. 2014. "Opportunism and Honest Incompetence—Seeking Explanations for Noncompliance in Public Procurement." *Journal of Public Administration Research and Theory* 25 (3): 953–79. https://doi.org/10.1093/jopart/mut081.
Kraemer, Kenneth L., Jason Dedrick, and Prakul Sharma. 2009. "One Laptop per Child: Vision vs. Reality." *Communications of the ACM* 52 (6): 66–73.
Lambright, Kristina T. 2009. "Agency Theory and beyond: Contracted Providers' Motivations to Properly Use Service Monitoring Tools." *Journal of Public Administration Research and Theory* 19 (2): 207–27.
Lavertu, Stéphane. 2016. "We All Need Help:'Big Data' and the Mismeasure of Public Administration." *Public Administration Review* 76 (6): 864–72.
Lee, Kristin. 2016. "Artificial Intelligence, Automation, and the Economy." Whitehouse.Gov. December 20, 2016. https://obamawhitehouse.archives.gov/blog/2016/12/20/artificial-intelligence-automation-and-economy.
Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service*. New York: Russell Sage Foundation.
Mannheim, Karl. 1940. *Man and Society in an Age of Reconstruction: Studies in Modern Social Structure*. Harcourt, Brace & World.
March, J.G, and Herbert A. Simon. 1993. *Organizations*. 2nd ed. New York: Wiley.
Mesquita, Ethan Bueno de. 2019. "The Perils of Quantification." Text. Boston Review. March 9, 2019. https://bostonreview.net/forum/economics-after-neoliberalism/ethan-bueno-de-mesquita-perils-quantification.
Moon, M, and S Bretschneider. 2002. "Does the Perception of Red Tape Constrain IT Innovativeness in Organizations? Results from Simultaneous Equation Model and Implications." *Journal of Public Administration Research and Theory* 11 (3): 327–52.
Morozov, Evgeny. 2014. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.
Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Washington, D.C.: Georgetown University Press.
Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.
O'Leary, Rosemary. 2013. *The Ethics of Dissent: Managing Guerrilla Government*. Thousand Oaks, CA: CQ Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
Parasuraman, Raja, and Dietrich H Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52 (3): 381–410.
Roberts, Alasdair. 2020. "Bridging Levels of Public Administration: How Macro Shapes Meso and Micro." *Administration & Society* 52 (4): 631–56.
Rusbult, Caryl E, Dan Farrell, Glen Rogers, and Arch G Mainous III. 1988. "Impact of Exchange Variables on Exit, Voice, Loyalty, and Neglect: An Integrative Model of Responses to Declining Job Satisfaction." *Academy of Management Journal* 31 (3): 599–627.
Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Random House.
Russell, Stuart, and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach*. 3 edition. Upper Saddle River: Pearson.
Selznick, Philip. 1948. "Foundations of the Theory of Organization." *American Sociological Review* 13 (1): 25–35.
Simon, Herbert A. 1997. *Administrative Behavior*. 4th ed. New York: Simon & Schuster.
Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.
Warschauer, Mark, and Morgan Ames. 2010. "Can One Laptop per Child Save the World's Poor?" *Journal of International Affairs* 64 (1): 33–51.
Young, Matthew M, Justin B Bullock, and Jesse D Lecy. 2019. "Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration." *Perspectives on Public Management and Governance* 2 (4): 301–313. https://doi.org/10.1093/ppmgov/gvz014.
Zacka, Bernardo. 2017. *When the State Meets the Street*. Cambridge, MA: Harvard University Press.