Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions

Ali Alkhatib
Stanford University Computer Science
ali.alkhatib@cs.stanford.edu

Michael Bernstein
Stanford University Computer Science
msb@cs.stanford.edu

ABSTRACT

Errors and biases are earning algorithms increasingly malignant reputations in society. A central challenge is that algorithms must bridge the gap between high–level policy and on–the–ground decisions, making inferences in novel situations where the policy or training data do not readily apply. In this paper, we draw on the theory of *street–level bureaucracies*, how human bureaucrats such as police and judges interpret policy to make on–the–ground decisions. We present by analogy a theory of *street–level algorithms*, the algorithms that bridge the gaps between policy and decisions about people in a socio-technical system. We argue that unlike street–level bureaucrats, who reflexively refine their decision criteria as they reason through a novel situation, street–level algorithms at best refine their criteria only after the decision is made. This loop–and–a–half delay results in illogical decisions when handling new or extenuating circumstances. This theory suggests designs for street–level algorithms that draw on historical design patterns for street–level bureaucracies, including mechanisms for self–policing and recourse in the case of error.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**;

KEYWORDS

Street–level algorithms; Street–level bureaucracies; Artificial Intelligence

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300760

ACM Reference Format:
Ali Alkhatib and Michael Bernstein. 2019. Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3290605.3300760

1 INTRODUCTION

People have grown increasingly frustrated with the decisions that algorithmic systems make over their lives. These decisions can have weighty consequences: they determine whether we're excluded from social environments [**?** **?**]; they decide whether we should be paid for our work [**?**]; they influence whether we're sent to jail or released on bail [**?**]. Despite the importance of getting these decisions right, algorithmic and machine learning systems make surprising and frustrating errors: they ban good actors from social environments and leave up toxic content [**?** **?**]; they disproportionately make bail inaccessible for people of color [**?**]; they engage in wage theft for honest workers [**?**]. Across diverse applications of algorithmic systems, one aspect seems to come through: surprise and — quite often — frustration with these systems over the decisions they make.

Researchers have approached these problems in algorithmic systems from a number of perspectives: some have interrogated the hegemonic influence that these systems have over their users [**?**]; others have documented and called attention to the unfair and opaque decisions these systems can make [**?**]; others still have audited algorithms via the criteria of fairness, accountability and transparency [**?**]. But even if we could answer the questions at the hearts of these research agendas — finding clear guiding principles, incorporating the needs of myriad stakeholders, and ensuring fairness — these algorithms will always face cases that are *at the margin*: outside the situations seen in their training data. For example, even a value–sensitive [**?**], fair, and transparent content moderation algorithm will likely make errors with high confidence when classifying content with a new slur that it has never seen before. And it is exactly at these moments that the algorithm has to generalize: to "fill in the gaps" between the policy implied by training data and a new case the likes of which it has never seen before.

"Filling in the gaps" is not a problem unique to algorithms, however: it has been the chief concern for decades of what bureaucratic theorists call *street–level bureaucracies* [**?**]. A street–level bureaucracy is the layer of a bureaucracy that directly interacts with people. They are responsible for making decisions "on the street", filling in the gaps between legislated policies and the situations in front of them. Street–level bureaucrats are police officers, judges, teachers, customer service agents, and others with whom members of the public interact frequently, and who make everyday decisions that affect our lives throughout the day. In all of these roles, street–level bureaucrats make important decisions about cases both familiar and new that require them to translate an official policy into decisions about the situations they face. A police officer chooses whether to issue a warning or a traffic citation; a judge decides whether to allow a defendant to pay bail or to have them remanded to jail; a teacher determines whether to waive a course's prerequisites for a student. These decisions often involve nuance or extenuating circumstances, making it all but impossible to prescribe the right response for all situations.

Street–level bureaucracies are important because the decisions they make are the manifestation of the power of the institution, and are *effectively* policy. Regardless of what's explicitly legislated or prescribed, policy is effected by the street–level bureaucrat's execution of their duties. Police officers choose whether to issue citations to people speeding only slightly over the speed limit; judges may make every effort to make bail accessible given a defendant's circumstances; teachers may allow some flexibility in course prerequisites. In each case, these street–level decisions become expectations of the system writ large. When these effective policies are biased (e.g., [**?**]), it prompts broad critiques and reviews of the policy, organization, or system.

In this paper, we draw the analogy to pose people's interactions with algorithmic aspects of sociotechnical systems as interactions with what we term *street–level algorithms*. Street–level algorithms are algorithmic systems that directly interact with and make decisions about people in a sociotechnical system. They make on–the–ground decisions about human lives and welfare, filling in the gaps between platform policy and implementation, and represent the algorithmic layer that mediates interaction between humans and complex computational systems.

Our central claim is that street–level algorithms make frustrating decisions in many situations where street–level bureaucrats do not, because street–level bureaucrats can reflexively refine the contours of their decision boundary *before* making a decision on a novel or marginal case, but street–level algorithms at best refine these contours only *after* they make a decision. Our focus here is on cases at the margin: those representing marginal or under-represented groups, or others creating novel situations not seen often or at all in the training data. When street–level bureaucrats encounter a novel or marginal case, they use that case to refine their understanding of the policy. When street–level algorithms encounter a novel or marginal case, they execute their pre–trained classification boundary, potentially with erroneously high confidence [**?**]. For a bureaucrat, but not an algorithm, the execution of policy is itself reflexive. For an algorithm, but not for a bureaucrat, reflexivity can only occur after the system receives feedback or additional training data. The result is that street–level algorithms sometimes make nonsensical decisions, never revisiting the decision or the motivating rationale until it has prompted human review.

We will look at several case studies through this lens of street–level algorithms. First, we will discuss how YouTube's monetization algorithms targeted LGBTQ content creators. Second, we will analyze the algorithmic management of crowd workers, discussing problems that arise when algorithmic systems evaluate the quality of workers who must themselves make interpretations about underspecified tasks. Third, we will look at judicial bail–recommendation systems and interrogate their biases — for example, disproportionately recommending jail for people of color. In all of these cases, we'll illustrate the difference between policy explication and policy execution, or in some other sense, the difference between training data and test data.

We will discuss ways that designers of sociotechnical systems can mitigate the damage that street–level algorithms do in marginal cases. Drawing on analogy to street–level bureaucracies, we identify opportunities for designers of algorithmic systems to incorporate mechanisms for recourse in the case of mistakes and for self-policing and audits. Recognizing that bureaucrats often operate by precedent, we suggest that people might compare their case to similar cases seen by the algorithm, arguing specifically how their case is marginal relative to previously observed cases. We also reflect on the emerging set of cases where street–level bureaucrats utilize or contest the output of street–level algorithms, for example judges making decisions based in part on the predictions of bail-setting algorithms.

We suggest that many of today's discussions of the invisible systems in human–computer interaction might also be considered interrogations of street–level algorithms. When we talk about systems that decide which social media posts we see [**?**], when we find ourselves in conflict with Wikipedia bots [**? ?**], and when we enforce screen time with our children [**?**], we are creating, debating, and working around street–level algorithms. Our hope is that the lens we introduce in this paper can help make sense of these previously disconnected phenomena.

2 STREET–LEVEL BUREAUCRACY

Street–level bureaucracies are the layer of a bureaucratic institutions that intermediate between the public and the institution itself. They're the police, the teachers, the judges, and others who make decisions "on the street" that largely determine outcomes for stakeholders of the cities, the schools, the courts, and other institutions in which they work. These functionaries are the points of contact for people who need to interact with these organizations.

Street–level bureaucracies are important because it's here that much of the power of the institution becomes manifest. At numerous stages in interactions with the bureaucracy, officials make consequential decisions about what to do with a person. Police officers decide who to pull over, arrest, and interrogate; judges pass sentences and mediate trials; instructors decide a student's grades, ultimately determining their education outcomes. To put it another way: Universities are vested with the power to grant degrees, but that power manifests in the classroom, when an instructor gives a student a passing or failing grade. Laws govern the actions of police officer, but that power manifests in their actions when they cite someone or let them off with a warning.

Street–level bureaucracies are important for reasons beyond the outcomes of members of the public; they substantially affect the outcomes of the organizations they serve. Whether a bureaucratic institution succeeds in its goals or not is largely influenced by the street–level bureaucrats [**?** **?**]. Efforts to constrain street–level bureaucrats are also fraught with challenges [**?**]; the specialized tasks that bureaucrats perform necessitates a certain degree of autonomy, which affords them latitude to make determinations according to their best judgment.

The consequences of the responsibility to make these kinds of decisions autonomously are difficult to exaggerate. By not issuing tickets for cars speeding only marginally over the speed limit, police officers effect a policy of higher speed limits. An instructor can waive a prerequisite for their course for inquiring students, effecting a policy that overrides the prerequisite for those that ask. Like all instruments of power, the power to enact effective policy can cause harm as well: judges might dismiss cases involving white people at a higher rate than they dismiss cases involving people of color; women are less likely to ask for exceptions [**?**], which can lead to an emergent gender bias. And street–level bureaucrats can be the arbiters of prejudice, manifesting policies and regimes that bring lasting harm to groups that are consequentially regarded as lesser. In all of these cases, regardless of the details of their choices, *defined* policies transform into *effective* policies through street–level bureaucrats.

The term of "street–level bureaucracy" was introduced by **?** in his **?** working paper [**?**] and explicated more comprehensively in his book on the subject in **?** [**?**]. Although the intuition of street–level bureaucracies had existed for some time, **?** formalized the insight and the term. Prior to this work, the academic focus on politics had specifically been of the people in formally recognized positions of power: the elected officials who authored policy. **?** argued that the true locus of power lay in the people who turn the policies into actions. This point of view became a formative landmark in thinking about governmental administration and political science more broadly, shifting focus from elected officials to the everyday bureaucrat, with over 14,000 citations and counting.

Street–level algorithms

We argue in this paper that we would benefit from recognizing the agents in sociotechnical systems as analogous to **?** 's street–level bureaucrats: that we live in a world with *street–level algorithms*. Street–level algorithms are tasked with making many of the kinds of decisions that street–level bureaucrats have historically made — in some cases actually subsuming the roles of bureaucrats — with many of the same emergent tensions that **?** originally described. This framing helps us understand the nuances between expressed policy and effected policy, and how those things diverge.

Street–level algorithms are the layer of systems — especially but not exclusively sociotechnical systems — that directly interact with the people the systems act upon. These algorithmic systems are specifically responsible for making decisions that affect the lives of the users and the stakeholders who operate said systems. Street–level algorithms take the explained policy, and often data that train a model or decision–making framework, and manifest power as decisions that immediately affect stakeholders. These interactions happen every day, sometimes without us realizing what's transpiring. Facebook, Twitter, Instagram, Reddit, and myriad other sites use algorithmic systems that choose what items to show in users' news feeds, and the order in which to show them; companies like Google and Amazon use algorithms which decide which advertisements to show us; Wikipedia and other peer production websites use bots to rebuff edits that don't correspond to previously articulated standards; credit card companies, PayPal, Venmo, and others employ algorithmic systems that flag payment activity as fraudulent and deactivate financial accounts.

These systems that make and execute decisions about us represent the layer that mediates the interaction between humans and much more nebulous sets of computational systems. It's here, we argue, that we should be focusing when we discuss questions of fairness in algorithmic systems, raise

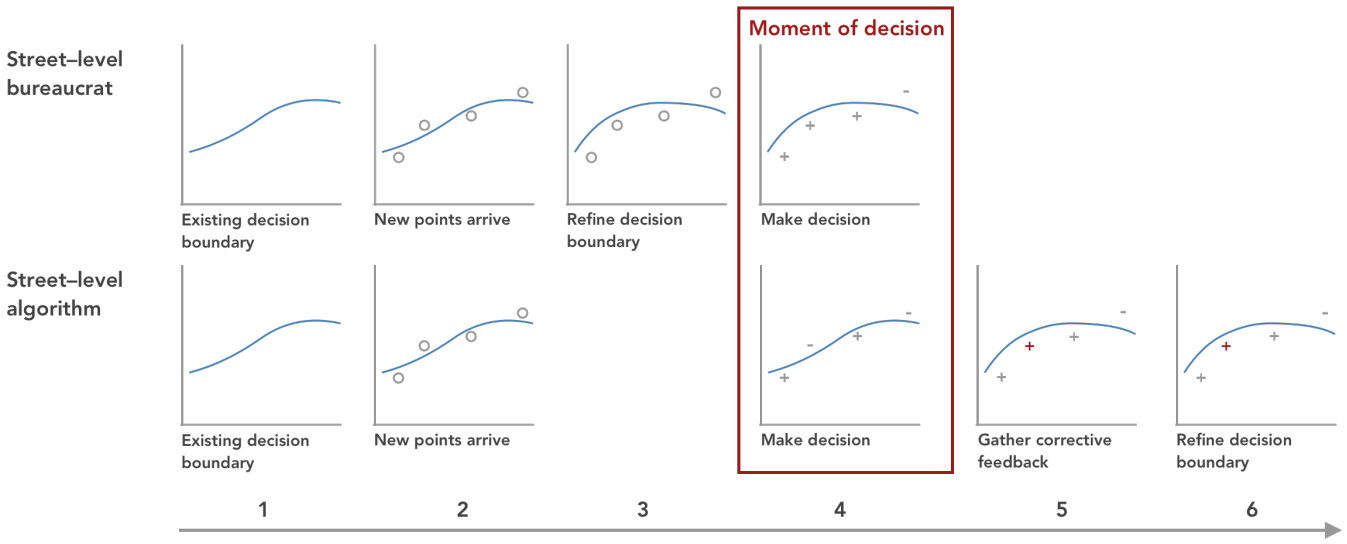

Figure 1: A timeline illustrating how reflexivity differs between street–level bureaucrats and street–level algorithms. The bureaucrat refines their decision boundary *before* making a decision about a new case, whereas the algorithm does so *afterwards*.

concerns about the lack of accountability of complex algorithms, and critique the lack of transparency in the design, training, and execution of algorithmic systems. This layer is where all of those events take place, and specifically where decisions one way or another are manifest and become real.

Moreover, we in HCI have been thinking about street–level algorithms for the better part of a decade — without naming it as such. When we talk about the systems that decide which posts we see [**?**], street–level algorithms have made decisions about the content in our news feeds, and the order in which we see it. When we fight with Wikipedia bots [**? ? ?**], street–level algorithms have erred in their enforcement of policies they've been taught either by programming or deep learning. Street–level algorithms enforce screen time with our children [**?**], manage data workers [**? ? ?**], and motivate the long–term directions of organizations [**?**].

Reflexivity in bureaucracy and algorithms

This reframing gives us traction on the issue that originally brought us here: why have street–level algorithms seen such intense criticism in the last several years? What is it about street–level algorithms that make the routinization of dealing with people so problematic that doesn't exist in, or doesn't elicit the same reactions about, bureaucracies? Street–level bureaucrats, **?** points out, practice *discretion*. Discretion is the decision of street–level bureaucrats not to enforce the policies precisely as stated, even if it means ostensibly failing their task, in favor of achieving the organization's goal. Street–level bureaucrats are capable of doing all of this only because they engage in reflexivity, thinking about their roles as observers, agents, and decision–makers in a given setting, and about the impact that their decision will have. Street–level bureaucrats reflect on their roles and on the circumstances and construct their reasoning accordingly.

Street–level bureaucrats are capable of making sense of new situations and *then* construct rationales that fill in the gaps. When police officers arrive at novel situations, they can make sense of what's happening — recognizing that this is a situation they *need* to make sense of — and they can intuit an appropriate rationale and application of rules to deal with the current case; the decision they make informs the next decision they make about incidents in a similar decision space. When instructors make decisions about whether to allow a student to take a course without a defined prerequisite, their decision contributes to a rationale that continually develops.

Street–level algorithms, by contrast, can be reflexive only *after* a decision is made, and often only when a decision has been made incorrectly. Even reinforcement learning systems, which require tight loops of feedback, receive feedback only after they take an action. Often, these algorithms only ever receive feedback after a wrong decision is made, as a corrective measure. Sometimes, algorithmic systems don't receive corrective input at all. Algorithmic systems don't make in–the–moment considerations about the decision boundary that has been formed by training data or explicit policies encoded into the program. Instead, the decision boundaries are effectively established beforehand, and street–level algorithms classify their test data without consideration of each case they encounter, and how it might influence the system to reconsider its decision boundary.

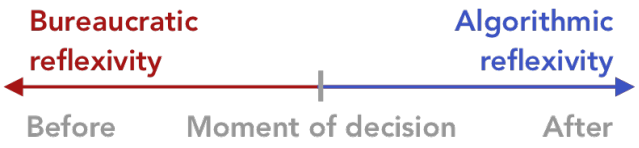

Figure 2: Bureaurats can act reflexively before making the decision; algorithms, requiring feedback, can at best retrain and act reflexively after the decision.

We can illustrate how this underlying difference between algorithmic and human agents manifests by using a single specific case where both kinds of agents intervened fundamentally differently. Facebook's policy prohibiting nudity came under scrutiny in 2008 when mothers protested action taken against pictures of them breastfeeding [**?**]. Facebook adjusted its moderation policies to allow nudity in the case of breastfeeding, but its moderation team then encountered yet more marginal and novel cases. For example, they discovered they needed to decide: is it breastfeeding if the baby isn't actually eating? What if it's not a baby, but instead an older child, an adult, or an animal that's at the breast? These interpretations had to be developed to handle these cases as they arose on Facebook. In contrast, today Facebook uses algorithms to detect nudity. For several of these classes of photos, the pretrained algorithm makes a decision guided by the data that informed it and, in accordance with its training, it removes the photos. As a street–level algorithm, the system made its decision *ex ante* and executed its decision according to the decision boundary that it generated. Facebook's moderation team, in contrast, made the decision *ex post*, debating about the implications of each option, the underlying rationale that would corroborate their intuition for whether it should be allowed, and what their boundary should be.

3 CASE STUDIES

We will focus on three cases to illustrate a methodology for thinking about the problems with street–level algorithms more broadly. We will look at moderation through YouTube's demonetization system and how it fails when it encounters underrepresented groups—in this case, members of the LGBTQ community discussing their gender identity and sexuality. We will proceed to examine the algorithmic management of workers in online labor markets and the perils of misjudgment by algorithmic quality control systems. Finally, we will discuss the emergence of algorithmic bias in judicial bail recommendation systems—a topic that has already generated substantial academic discussion in the past several years. In all of these cases, a theme will emerge of algorithmic system encountering either a novel otherwise unforeseen situation for which it was not trained—and for which, perhaps, it *could* not have been trained—and that system making a decision without reflexivity.

YouTube content moderation

YouTube enforces many of its content policies algorithmically. The decisions of how to handle user–generated content in general have enormous bearing on the culture of today's online platforms [**?**]. On YouTube, machine learning systems classify whether each uploaded video contains content that is protected by an existing copyright [**?**], and whether it violates YouTube's "advertiser–friendly content guidelines" [**?**]. The content guidelines, for example, state that to earn ad revenue, videos must not feature controversial issues, dangerous substances, harmful acts, inappropriate language, or sexually suggestive content. YouTube's advertisers do not want their ads run on videos with this content, so if a video does not meet the content deadlines, it is *demonetized*, where YouTube does not show ads on the video and the content creator receives no income for it.

But these demonetization algorithms have made highly–publicized errors. For example, YouTube began labeling videos uploaded by transgender YouTubers as "sexually explicit" [**?**], demonetizing them. When video titles included words like "transgender", they were demonetized; when the content creators removed "transgender" but left other aspects of the video as–is, the same videos were monetized normally [**?**].

YouTube's classifier for sexually explicit content misfired. Discussions of transgender issues are not necessarily about sex at all; while sex and gender have historically been conflated [**?**], they refer to different ideas [**?**]; the state of being trans does not imply any particular sexuality. The algorithm's training data or model may not have yet represented the evolution of our collective consciousness to acknowledge this distinction, or our changing desire to treat people variably along these dimensions and with more consideration than we have in the past. As people increasingly turn to and rely on YouTube to be a venue to earn a living — much like a public square — YouTube's algorithmic classification system denies people a space for this discussion about the biological and cultural distinctions of sex and gender. YouTube eventually apologized, stating "our system sometimes make mistakes in understanding context and nuances" [**?**].

These issues also emerge in the opposite direction, so to speak. The same algorithms that mischaracterized discussions about *gender* as *sexual* and consequently inappropriate for advertising failed to flag inappropriate content disguised in a large number of children's cartoons. Fraudulent videos of Peppa Pig being tortured at a dentist's office were left alone, and videos of cartoon characters assaulting and killing each other were passed over by the algorithm, in some cases being included in YouTube Kids, a subset of YouTube which offers to source content appropriate specifically for children.

This failure to identify troubling content wrapped in ostensibly child–friendly animation again led to a refinement of YouTube's policies and algorithms.

It can be useful to think about YouTube's content moderation algorithms as analogous to the class of street–level bureaucrats who monitor and interact with street performers in offline urban contexts. Street performance, or *busking*, is usually monitored by police [**?**]. Many cities have laws which restrict busking in certain contexts. The police must identify when they should enforce laws strictly and when they should take a more permissive stance: the details of enforcement of those laws is necessarily left to police officers, affording them substantial latitude. The public record is full of instances of police ranging in behavior from aggressively managing to enjoying and engaging in the performance themselves [**?**]. As performance by nature often pushes the bounds of expectations and street performance in particular is inherently experimental [**?**], police have to be flexible about the application of their powers and make reasonable decisions in new circumstances. The challenge is to characterize the novelty of the situation and reflect on the decision they're being called to make. More importantly, they must make these decisions in the moment, applying the implications of that decision to a constantly–updating intuition about the effective policy they're creating through their selective enforcement of laws.

We can think of YouTube's monetization algorithm as akin to a sort of police force that encounters hundreds of thousands of new performances every day and is expected to navigate those situations appropriately. The challenge is that this algorithmic police force trains and updates its policies in batches: it is executing on today's performances based on yesterday's data. Yesterday, transgender people weren't speaking publicly, in a venue accessible all over the world, about deeply personal aspects of their lives in the ways that they now do. The algorithm is always behind the curve: at best, it gets feedback or negative rewards only after it has executed its decision, in this case when YouTubers appeal or gather media attention. By contrast, police reflexively construct an interpretation of the situation as soon as they encounter it, rather than merely match on learned patterns.

This case highlights a shortcoming with a commonly offered solution to these kinds of problems, that more training data would eliminate errors of this nature: culture always shifts. Simply having more data will never allow us to anticipate and prevent these kinds of errors. Experimentation is often the point of performance and art, and certainly part of the nature of public performance art like YouTube [**? ?**]. Society is slowly (albeit not consistently) coming to acknowledge that transgender people are people, and in doing so recognize that their gender identities are acceptable to discuss. In statistical terms, we might say that the data in this system is a nonstationary process: the distribution (of words, topics, and meanings) changes, sometimes abruptly. YouTube was, at one time, a uniquely empowering space for members of the transgender community to be candid and to explore their shared experiences [**? ?**], but arguably culture has grown in ways that YouTube and its content classification algorithms have not. Reinforcement rewards and new training data can help an algorithm reconfigure its decision boundary, but even deep learning only gets this new training information *after* it has already made decisions, sometimes hundreds of thousands of decisions — *effecting* a policy deterring transgender YouTubers from discussing their gender identity, or risk being demonetized for discussing content the system erroneously classifies as "sexual".

The effects of bad street–level algorithms are farther–reaching than the immediate cases that are mishandled. **?** points out that people begin to work around street–level bureaucracies when they become unreliable and untrustworthy, or when the public decide that they cannot hope for bureaucrats to make favorable decisions [**?**]. We see this phenomenon unfolding on YouTube: as their demonetization algorithms mishandle more and more YouTubers [**? ?**], creators have begun to circumvent YouTube monetization entirely, encouraging audiences to support them through third–party services such as Patreon [**?**].

Quality control in crowd work

Crowd work — that is, piecework [**?**] on platforms such as Amazon Mechanical Turk — has become both a crucial source of income for many workers, and a major source of annotations for modern machine learning systems. Tasks on crowd work platforms range from the prosaic (e.g., transcribing audio) to the novel (literally, in some cases, writing novels [**?**]). Nearly every approach to crowd work requires a strategy for quality control, deciding which workers are allowed to continue working on the task, or at worst, which ones are denied payment [**? ? ? ?**]. These algorithmic approaches variously seek to test workers' abilities, measure cross-worker agreement, and motivate high–effort responses. Workers' reputations — and by extension, their prospective abilities to find work — are determined by whether these algorithms decide that workers' submissions are high quality.

Unfortunately, these quality control algorithms have — perhaps unintentionally — resulted in wage theft for many workers, because the algorithms deny payment for good-faith work. There are often many correct ways to interpret a task [**?**], and often these algorithms only recognize and pay for the most common one. Requesters' right to reject work without compensation is the second–most common discussion topic about the Mechanical Turk participation agreement [**?**], and mass rejection is a persistent, widespread fear [**?**], and the fear of getting work algorithmically rejected

looms large [**?**]. To manage the concern, workers use back–channels to share information about which requesters are reliable and don't reject unnecessarily [**? ?**].

We can think of the relationship Turkers have with the systems that manage them as analogous in some sense to the foremen who manage factory workers. Pieceworkers in railroads, car assembly, and many other industries historically interacted with foremen as their interface with the company: the foreman assigned workers to suitable work given their qualifications, responded to workers when they needed assistance, and provided feedback on the output of the work. This relationship was a crucial boundary between managers and workers that foremen had to navigate carefully — neither management nor worker, but decidedly and intentionally in the middle [**?**].

The foreman's job was important because even the most standardized work sometimes surfaces unique circumstances. As much as managers attempt to routinize work, scholarship on the subject tells us that improvisation remains a necessary aspect of even the most carefully routinized work [**? ?**].

> When performance is difficult to evaluate, imperfect input measures and a manager's subjective judgment are preferable to defective (simple, observable) output measures.
>
> — **?** [**?**], as cited in [**?**]

The challenge is that algorithmic review mechanisms are not well–equipped to understand unusual cases. A crowd worker's output is almost never evaluated by humans directly, but algorithmically scored either in comparison to the work of other workers or a known "gold standard" correct response [**?**]. However, often the most popular answer isn't actually the correct one [**?**], and a gold standard answer may not be the only correct answer [**?**]. If the task becomes more complex, for example writing, algorithmic systems fall back to evaluating for syntactic features that, paradoxically, both make it easy to game and frustratingly difficult to succeed [**?**]. This general characterization of an algorithmic agent — one that essentially seeks agreement in some form — is not designed to evaluate entirely novel work. With all of the mistakes these systems make, and with the additional work that crowd workers have to do to make these systems work [**?**], it should come as little surprise that crowd workers are hesitant to attempt work from unknown and unreliable requesters [**?**].

The problem is that these algorithmic foremen can't distinguish novel answers from wrong answers. Rare responses do not necessarily mean that the worker was not paying attention — in fact, we prize experts for unique insights and answers nobody else has produced. However, where the foreman would evaluate unusual work relative to its particular constraints, the algorithm at best can only ask if this work resembles other work. Crowd workers might then receive a mass rejection as a result, harming their reputation on the platform. Again as in other cases we've discussed, gathering more training data is not a feasible path to fix the problem: crowd work is often carried out exactly in situations where such data does not yet exist. Machine learning algorithms that evaluate worker effort also require substantial data for each new task [**?**]. The street–level algorithm is stuck with a cold–start problem where it does not have enough data to evaluate work accurately.

Algorithmic bias in justice

American courts have, in recent years, turned to algorithmic systems to predict whether a defendant is likely to appear at a subsequent court date, recommending the level at which bail should be set. The idea is to train these systems based on public data such as whether historical defendants who were let out on bail actually showed up at their court date. These systems take into account dimensions such as the charges being levied, the defendant's history, their income level, and much more, in the hopes of yielding outcomes that increase public welfare, for example by reducing jailing rates by 40% with no change in resulting crime rates [**?**], all while being less biased and more empirically grounded.

Instead, observers have seen patterns of bias that either reflect or amplify well–documented prejudices in the criminal justice system. Researchers have found bail–recommendation systems replicating and exacerbating racial and gender biases — recommending against offering bail to black men disproportionately more than for white men, for example [**?**]. In some cases, it seems that problems stem from the data that informs models [**?**]; in others, recommendation systems are, as AI researchers say, reflecting a mirror back at our own society, itself steeped in racial prejudice [**? ?**].

In this case, the analogical street–level bureaucrat is probably clear. It is the person whose work the algorithm seeks to replicate: the judge. These algorithms are often even trained on judges' prior decisions [**?**]. However, as street–level bureaucrats, judges have struggled to answer the question of "which defendants secure release before trial?" for most of the 20th century [**?**]. While constitutional law protects people from "excessive bail", **?** points out that ultimately this decision is left to the discretion of the judge [**?**]. A judge hears preliminary information about the case, reviews information about the defendant (such as past criminal record, assets, and access to means of travel), and sets bail that should be sufficiently high that a defendant will appear for their court date without being inaccessible.

In this third case study, we observe something new: a street–level bureaucrat interacting with a street–level algorithm. This interaction can be fraught: bureaucrats in the judicial system resist, buffer, and circumvent the algorithmic

recommendations, especially as those algorithms attempt to subsume the work of those bureaucrats. Indeed, **?** explores some of the tensions that emerge when algorithms begin to absorb bureaucrats' responsibilities and shift the latitude that bureaucrats enjoyed, finding that bureaucrats work around and subvert these systems through foot-dragging, gaming, and open critique as a way of keeping their autonomy [**?**]. **?** go further to illustrate some of the ways that designers of algorithmic systems can better support street–level bureaucrats given these and other tensions [**?**].

Researchers have contributed many valuable insights about bail recommendation algorithms from the perspective of fairness, accountability and transparency (reviewed in [**?**]); the literature of street–level bureaucracies adds a reminder that each case may involve novel circumstances and deserves thoughtful consideration about which humans in particular are well–equipped to reason. As **?** writes, "street–level bureaucrats … *at least [have] to be open to the possibility* that each client presents special circumstances and opportunities that may require fresh thinking and flexible action." [**?**]. Otherwise, why bother having judges or trials at all? Why not articulate the consequences directly in the law, feed the circumstances of the crime into a predefined legal ruleset (e.g., `{crime: murder}, {eyewitness: true}, {fingerprints: true}`), and assign whatever conclusion the law's prescriptions yield? Largely the reason that our society insists on the right to a trial is that there may be relevant characteristics that cannot be readily encoded or have not been foreseen in advance.

If street–level algorithms are liable to make errors in marginal and novel situations, it suggests that the problem is not just how to handle biased data, but also how to handle missing data. Increased training data is insufficient: for important cases at the margin, there may be no prior cases. Intersectionality is growing as an area of focus within HCI [**?**]; intersectionality fundamentally calls attention to the fact that combinations of traits (e.g., being a woman and a person of color) need to be treated as a holistically unique constellation of traits, rather than as some sort of sum of the individual traits. As a matter of probability, each additional dimension in the intersection makes that constellation less likely. While similar cases in the mainstream may well have been seen before by the algorithm, when the case is at the margin, its particular intersection of traits may be completely novel. Adding training data is almost a waste of resources here, as the combination may be so rare that even increasing dataset size tenfold or one hundredfold may only add a single additional instance of that combination.

In practice, this intersectional challenge is one reason why many democracies use a form of case law, allowing an individual to argue to a judge that their circumstances are unique and should be examined uniquely, and with discretion. Many cases are straightforward; however, when they're not, the court system must re–examine the case and the law in this new light. How could an algorithm identify a situation that needs to be treated as a novel interpretation of a policy, as opposed to one that is only a small variation on a theme that has been seen before?

Much of the discussion of judicial bail recommendation algorithms today is focused on the goals of fairness, accountability and transparency, or *FAT*. We argue that FAT is a necessary, but not sufficient, goal. Even a perfectly fair, transparent, and accountable algorithm will make errors of generalization in marginal or new cases.

4 DESIGN IMPLICATIONS

We've discussed the demonetization on YouTube, the management of crowd work, and the bias of algorithmic justice, but the same undercurrent moves all of these cases. The inability of algorithmic systems to *reflect* on their decision criteria appears across these diverse cases. A human has the capacity to recognize the substantive injustice of a judicial system that targets and disenfranchises people of color when faced with a new situation; an algorithm can only see a pattern. Not a good or bad pattern — just a pattern. And even the goodness or badness of that pattern must itself be taught.

What should designers of street–level algorithms do? The question we hope to answer is how to find and identify circumstances for which algorithmic systems would not yield problematic outcomes, assuming that designers want to create a prosocial, fair system. The defining goal, then, should be to identify cases requiring discretion and flexibility. In some cases that will be easy — some classifications' confidence will be low, or the result will be ambiguous in some predictable way. In much the way we already do, we should divert those cases to human bureaucrats. However, often the system performs these erroneous classifications with high confidence, because it does not recognize that the uniqueness of the input is different than other unique tokens or inputs.

? argues that street–level bureaucrats must exercise reflexivity, recognizing the underlying purpose of the tasks at hand, to be effective. If this is the substantive goal of designing effective street–level algorithms, then we need to figure out how to get there. Today's most advanced AIs cannot reflect on the purpose or meaning of the tasks for which they're optimizing results. We turn, then, to **?** 's work yet again, where he argues that appeals and recourse — and the ability for people to recognize the uniqueness of a situation when it's explained to them — are necessary features of street–level bureaucrats who fail to recognize the marginality of a case at hand the first time around. Developing more robust mechanisms for recourse may be the path to sufficient, if not yet effective, street–level algorithms.

We argue that system–designers need to develop ways for people to get *recourse* if the system makes a mistake, much like citizens of a bureaucratic institution can mitigate harm due to mistakes. Our theory suggests to look to best practices in bureaucratic design as inspiration. Crafting a fair appeals process is one common lever: for example ensuring that the person reviewing any appeal does not have conflicting interests or misaligned incentives — or in this case, perhaps not the same software developers, human bureaucrats, or machine learning model as the socio-technical system that made the original decision. Another approach is a predefined process for recourse, for example compensating lost income. Third, since bureaucracies can be as opaque as algorithms, many bureaucracies are required by law or design to publish materials describing peoples rights in plain language.

Recourse and appeals require grounds for the individual to understand precisely where and how the system made a mistake. How, for instance, can a person prove that they have been misjudged by the algorithm? One solution might be to represent the embeddings generated in classifying that case by showing similar points in the embedding space, suitably anonymized. If, for instance, the classification system figured that the current case was very similar to a number of other cases, presenting the user's case in the context of some of those closely–aligned cases can give the user sufficient context to articulate why their situation is marginal relative to the projection the system has of it.

For example, YouTube's demonetization system could communicate its judgments about videos to YouTubers, giving those performers an opportunity to argue that they've been misjudged. If a YouTuber uploads a video discussing gender identity, and the system thinks that content is sexual content, it might present a handful of similar videos nearby in the embedding space. By comparing their video to these, a YouTuber can identify whether they're being misjudged as similar to videos from which they're substantively different. This kind of information is crucial for systems such as these to develop depth in the sense that they understand the difference between gender and sex, for instance.

Bail recommendation systems could offer similar insights to stakeholders and help both judges and defendants better understand the intuition the algorithmic system has developed. A judge might see that the embeddings generated for a defendant are categorically wrong in some way that the algorithm either doesn't understand or can't measure. In this kind of scenario, information about the other defendants might be sensitive for various reasons, but some representation of the present case and its neighbors in the embedding space can reveal whether a categorical error has been made.

These examples may prove untenable due to any number of hurdles, but by sparking a conversation along this dimension — one that calls to attention how the choices of designers manifest in their systems as features of a bureaucracy which billions of people may have to navigate — we hope to encourage thinking about these problems along the same general lines that political scientists have been thinking for the better part of half a century. In doing so, we may be able to find and leverage stopgap solutions — developed and improved by social scientists over decades — that mitigate some of the harms done to marginalized communities.

This discussion has focused on individual–level recourse — what about institutional checks? In other words, how can designers of algorithmic systems ensure that the systems will self–police, or if they can at all? Our approach seeks grounding in the literature via the mechanisms that evaluate and manage bureaucracies to identify possible design directions. Street–level bureaucracies experience oversight via a number of channels, including internal audits, external review, and the judicial system. These methods are variably effective, which is to say that public administration has no perfect answer to accountability. However, these models have worked and likely will continue to work as ways that designers can build oversight into algorithms, for example peer juries of disruptive behavior online [**?**]. This approach provides us with a structure for reasoning about algorithmic systems with the added benefit of decades of theoretical refinement.

The more distant question, of course, is what needs to change before algorithms can reliably subsume the roles of street–level bureaucrats as **?** described. **?** argues that programs will never be able to take the roles of human beings: *"the essence of street–level bureaucrats is [that] they cannot be programmed"* [**?**], because they can't think deeply about the circumstances of the case in front of them or their role in the decision being rendered. That certainly is true today, but advances in AI since **?** 's time may have surprised even him. It may be possible that algorithmically intelligent systems will reach a point where they can believably take up the roles of human bureaucrats; what they will need to demonstrate is some capacity to reflect on the novelty of novel or unusual cases, and what the implications of their decisions might be *before* a decision is made.

5 DISCUSSION

Street–level bureaucracies are not a perfect metaphor for the phenomena we've discussed. We didn't address in any capacity the fact that street–level bureaucrats sometimes diverge in unexpected ways from the prerogatives of their managers. This becomes the source of tension in **?** 's treatment of street–level bureaucracies, but in our discussion of the relationships between street–level algorithms and their stakeholders, we avoided the relationship between engineers and designers and the systems themselves. Suffice it to say that while there is a disconnect between intent and outcome, the nature of that relationship is so different that it warrants

much further discussion. We also avoided a unique quality of machines in this nuanced tension: algorithmic systems operate at far greater speed than humans can [**?**], precipitating what **?** characterized as a "technical infeasibility of oversight" [**?**].

Nor are street–level bureaucrats paragons of justice, fairness, accountability, transparency, or any particular virtue. Street–level bureaucrats have historically been agents of immense prejudice and discrimination: writing insuring guidelines specifying that racially integrated neighborhoods are inherently less safe than white ones [**?**], for instance. **?** 's ethnography of organized crime in Boston, and of a corrupt police force that took payoffs to exercise their discretion more favorably toward criminal enterprises [**?**], illustrates in a different way how discretion can be applied contrary to our values. Street–level bureaucracies are loci of immense power — and power can be abused by those who have it.

Perhaps least certain of all the questions that emerge as a result of this discussion of street–level algorithms is that of the relationship between conflicting agents. What happens when street–level bureaucrats collide with street–level algorithms? The theory of street–level bureaucracies doesn't offer much to mitigate this tension. **? ?** have traced the landscape of challenges that may emerge and ways to mitigate those conflicts [**? ?**]. This area in particular needs further study: the fault lines are here to stay, and we need to reflect on this shifting of discretion from the bureaucrat to the engineer [**?**]. A value-sensitive approach [**?**] would ensure that engineers be careful of how the algorithms may support or undermine bureaucrats' authority.

Algorithms and machine learning may yet introduce new methods that override the observations made here. We are assuming, as is the case currently, that algorithms require feedback or additional training to update their learned models. Likewise, we are assuming that algorithms will continue to make errors of confidence estimation, and will make mistakes by labeling marginal, novel cases with high confidence. Nothing about the emerging architectures of modern machine learning techniques challenges these assumptions, but should it happen, the situation might improve.

Despite these limitations, we suspect that the lens of "street–level algorithms" gives us a starting point on many questions in HCI and computing more broadly. We've discussed the ways that street–level bureaucracies can inform how we think about YouTube content moderation, judicial bias, and crowdwork, but we could take the same framing to task on a number of other cases:

- *Moderation of forum content:* For many of the same reasons that we see trouble with the application of algorithmic classification systems on YouTube, we should expect to see problems applying algorithmic systems to textual forums.
- *Self-driving cars:* Cars will make algorithmic decisions all the way from Level 1, where vehicles decide to break when we get too close to others, to Level 3, where they will need to decide when to hand control back to drivers, to Level 5, where systems may decide which routes to take and thus how late we should be to our destinations. Self-driving cars are literal "street–level" algorithms.
- *Parental controls:* Algorithms that lock children out after a certain amount of screen time elapses will need to learn how to handle unforeseen situations when the device should remain functional, such as a threat or emergency.
- *AI in medicine:* When decisions are life-or-death, how does a patient or doctor handle an algorithm's potentially error-prone recommendations?

6 CONCLUSION

In this paper we've explored a framework for thinking about algorithmic systems mediating our lives: one that leans substantially on the scholarship on street–level bureaucracies. We've conducted this exploration by discussing three cases of particular salience in the scholarly and public discourse, but our overarching goal is to motivate the use of this framing to ask questions and even develop lines of inquiry that might help us better understand our own relationships with these systems — and hopefully to design better systems.

While we have alluded only briefly to the dangers of bureaucratic organizations and their histories reaffirming prejudice and biases, it's our hope that the underlying narrative — that these institutions, and in particular the agents "on the street", carry overwhelming power and should be regarded accordingly. It's not our intention to imply that bureaucratic organizations are in any sense a panacea to any problems that we've discussed; instead, we hope that people can take this discussion and begin to apply a vocabulary that enriches future conversations about algorithmic systems and the decisions they make about us. Indeed, by reasoning about street–level algorithms with the benefit of theoretical and historical background afforded by **?** 's discussion of street–level bureaucracies and the body of work that followed, we are confident that we (designers, researchers, and theorists) can make substantial progress toward designing and advancing systems that consider the needs of stakeholders and the potent influence we have over their lives.

ACKNOWLEDGMENTS

We would like to thank Os Keyes and Jingyi Li, among others, for volunteering their time and patience as they provided input to help us understand and discuss several of the sensitive topics with which we engaged.

This work was supported by a National Science Foundation award IIS-1351131 and the Stanford Cyber Initiative.