Automation and Mercy

Kiel Brennan-Marquez

This chapter explores the idea that machines are incapable of adopting a "merciful attitude" toward decision-making. If that is true, I argue it supplies a reason to be sceptical of many forms of legal automation - regardless of how powerful or computationally complex the instruments of automation become. To make this argument, I connect longstanding debates about the link between justice and the mercy, inspired by the scholastics, to contemporary literature on "algorithmic governance."

When automated systems replace human decision-makers, what is lost? Over the last few decades, scholars have developed two answers to this question: one focused on distributional accuracy,[1] the other on procedural integrity.[2] This chapter offers a different sort of answer. In some domains, I argue, the most salient drawback of automation is neither distributional nor procedural. Rather, it concerns the absence of a particular kind of attitude—a merciful disposition—on the part of those responsible for executing decisions.[3] Even at their most callous, human decision-makers tend to exercise some degree of forbearance. Judges dismiss charges. Executives grant pardons.[4] Guards unlock doors.[5] Not often, certainly not in every case —perhaps too little, perhaps too much. But whatever the exact calibration

[1] *See* Ryan Calo & Danielle Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 Emory L. J. 797 (2021); Andrea Roth, *Trial By Machine*, 104 Geo. L. J. 1245 (2016).

[2] *See* Hannah Bloch-Wehba, *Visible Policing: Technology, Transparency, and Democratic Control*, 109 Cal. L. Rev. 917 (2021); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 Admin. L. Rev. 1 (2019).

[3] The chapter builds on past (co-authored) work in this vein. *See* Kiel Brennan-Marquez & Stephen E. Henderson, *Role-Reversibility, AI, and Equitable Justice – Or: Why Mercy Cannot Be Automated*, 114 J. Crim. L. & Criminology Online 1 (2023); Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. Crim. L. & Criminology 137 (2019).

[4] This function is typically associated with heads of state—presidents, governors, and the like. But the logic may reach further. *See* Lee Kovarsky, *Prosecutor Mercy*, 24 New Crim. L. Rev. 326 (2021).

[5]

of mercy, its *possibility* forms the backdrop of all juridical decision-making, regardless of institutional particulars, across space and time.

To "automate away" forbearance, then, would be to discard an essential ingredient of the administration of justice among human beings, as it has been practiced for millennia.[6] Would this change be welcome? In what follows, I explore this question by drawing a link between (1) longstanding puzzlement about the relationship between justice and mercy and (2) today's "algorithmic governance" debates. Those debates typically unfold by asking what someone poised to suffer adverse treatment might say about the legitimacy of human judgment, on one hand, or robotic artifice, on the other. This emphasis on the "perspective of the condemned" is understandable, even virtuous. But it has obscured a different question, one of potentially greater importance, which requires taking the perspective of *the executioner*. How do decisions look from the vantage point of those responsible for carrying them out? From an "internal point of view," what does it mean to be the one charged with wielding the axe—not the one awaiting its blade?

Does it matter, in short, if executioners are free, until the very last moment, to lay down their arms?[7] Does this freedom change the moral quality of law? Some might say no. Others might say yes, but in a troubling way. There is, after all, a concept of legality—quite alive in our institutional practices today—that sees "mercy" as a euphemism for arbitrariness and caprice, which might suggest that automation stands to *perfect*, not to imperil, human legal institutions.[8] My goal is not to dislodge either of these positions. It is far more modest. I aim to explore the implications of the "pro-mercy" view for the enterprise of legal automation. Here is the argument on offer:

If the executioner's freedom is a welcome aspect of legal systems—if mercy is integral to law's moral quality—then all legal automation, regardless of specifics, should be cause for concern.

The inverse is not necessarily true. Automation may *still* be cause for concern even if mercy is irrelevant (or inimical) to law's moral quality. But

[6] *See* FERNANDA PIRIE, THE RULE OF LAWS: THE 4000 YEAR QUEST TO ORDER THE WORLD (2021); Martha Nussbaum, *Equity and Mercy*, PHIL. AND PUB. AFFAIRS (1993).

[7] For present purposes, I count judges—and many other state officials who are not literally responsible for administering capital punishment—in the "executioner" category. *See* Robert Cover, *Violence and the Word*, 95 YALE L. J. 1601 (1986).

[8] For an argument along these lines, see Jane Bambauer, *Filtered Dragnets and the Anti-Authoritarian Fourth Amendment*, 97 SO. CAL. L. REV. (forthcoming 2024).

we should be clear, either way, about what is at stake in today's "algorithmic governance" debates. In the end, those debates are not about technical specifics or jurisprudential minutiae. They are about the fundamental status of moral agency—the sense of freedom, or lack thereof—in our public life.

Since the scholastics, and perhaps long before, the relationship between mercy and justice has been uneasy. If mercy represents a departure from the requirements of justice—if mercy is "beyond" justice—how is it distinct from injustice? If justice supplies an answer, in principle, to all relevant cases, what role is left for mercy? For thinkers like Anselm and Aquinas, the urgency of this question was metaphysical: they sought to reconcile the promise of natural law—the notion that God's will is intelligible to human reason—with the idea that salvation is an act of grace, freely given and irreducible to law. Their solution, broadly speaking, was to imagine grace as a supplement to law. "God acts mercifully," Aquinas famously wrote, "not indeed by going against His justice, but by *doing something more than justice*."[9]

Modern legal systems have inherited a version of this solution. We frequently imagine mercy as something "more than" justice: not counteracting or overriding the latter's requirements, but improving upon them. On this view, justice becomes a necessary but insufficient condition of legal perfection. In their ideal form, the thought goes, legal institutions will be just, but they will not be *exclusively* just. They will also be merciful; they will also make room for forbearance.[10]

The "supplemental" idea of mercy is not without dissent. For some observers, mercy is a structural pathology: a bug, not a feature, of modern legal systems.[11] Whatever its appeal in particular cases, the argument goes, mercy—as an act of radical discretion—is antithetical to the rule of law. All mercy, regardless of moral valence, represents the triumph of personal will over law's impersonal majesty.

For the scholastics, to be clear, this was exactly the point. God's will, manifest as merciful salvation, was supposed to take precedence over natu-

[9] Thomas Aquinas, Summa Theologica (Part I).

[10] For an example of this form of argument, see Rachel Barkow, *The Ascent of the Administrative State and the Demise of Mercy*, 121 Harv. L. Rev. 1332 (2008).

[11] *See* Aziz Z. Huq, *The Difficulties of Democratic Mercy*, 103 Cal. L. Rev. 1679 (2015); Dan Markel, *Against Mercy*, 88 Minn. L. Rev. 1421 (2004).

ral law; or at any rate, natural law was not supposed to *preclude* merciful salvation. For modern sceptics, on the other hand, transplanting the puzzle to the realm of human legal institutions causes its solution to invert. Mercy, the modern sceptics insist, is no longer the thing that needs protecting; rather, it is what justice must be protected *from*. Foreclosing the space of mercy—taking institutional steps necessary to ensure that reason, to the maximal extent possible, *does* preclude will—is a central aspiration of "law's empire."[12]

From here, the debate has many subtle turns. Some have argued, for example, that counterposing justice and mercy is too simple—the wrong frame on the problem. Instead, mercy is best understood as a *continuation* of justice: a complement, not a supplement, to the application of discrete rules, especially in cases whose particularity, idiosyncrasy, or pathetic quality make them difficult to categorize ex ante.[13] Others, meanwhile, have argued that acts of mercy are no more (or less) unaccountable than ordinary acts of sovereign decision—which is certainly a problem to be managed in practice, but hardly a challenge to the rule of law in principle. In fact, advanced legal systems embed many "states of exception" into their everyday workings; forbearance is not special.[14]

Not surprisingly, these nuanced reconstructions of mercy have spawned equally nuanced rejoinders. Some observers argue, for example, that even if mercy harmonizes with rule-of-law principles, it tends, in practice, to be deployed regressively—and becomes objectionable for all the usual reasons that regressive aspects of the legal system are objectionable.[15] Along similar lines, other observers worry that forbearance mechanisms stunt the dynamic evolution of legal rules, causing doctrine to atrophy over time. Why bother refining normally-applicable rules, the thought goes, when mercy is there to "clean up" the exceptions?[16]

For present purposes, the bottom-line is that even though (1) mercy can be celebrated for many different reasons, (2) most observers attribute *some* value to mercy—such that its wholesale elimination from human legal

[12] *See* Ronald Dworkin, Law's Empire (1986).

[13] *See* Linda Meyer, "The Merciful State," in FORGIVENESS, MERCY, AND CLEMENCY (Hussain & Sarat, eds. 2007); ROBIN WEST, CARING FOR JUSTICE (1997).

[14] *See* Giordana Campagna, *The Miracle of Mercy*, 41 OXFORD J. LEGAL STUDIES 1096 (2021).

[15] *See* Markell, note 9.

[16] *See* Mary Sigler, "Equity, Not Mercy," in THE NEW PHILOSOPHY OF CRIMINAL LAW (Flanders & Hoskins, eds. 2016).

systems would register as a loss. Furthermore, even those who express scepticism about mercy often do so in relative terms: they argue that mercy's value is, in certain contexts, not worth prioritizing over *other* values, not that it lacks value at all.

Some observers, to be clear, *do* argue that mercy is categorically unworthy of prioritization, insofar as they take mercy to conflict necessarily with the rule-of-law.[17] But that is by far the minority position: an outlier the rest of this chapter will set aside. If one believes that mercy is, ultimately, just a temptation to avoid—a contingent feature of legal systems that, under the right conditions, could and should be eliminated—the "mercilessness" of automated decisions will not be cause for concern. It may be cause for cheer. If, on the other hand, mercy has *some kind* of moral worth, the question becomes: can the radical freedom that mercy instantiates be replicated by non-humans means? Are the values served by mercy—whatever their exact content and contours—susceptible to automation?

Let us begin with what we know about *human* mercy. For one thing, mercy is inextricably linked to grace. Mercy is never compelled; its receipt is never a matter of right or entitlement, and its dispensation is never a matter of duty. Rather, mercy is, by necessity, "freely given."[18] We also know, moreover, that nothing about the conceptual structure of mercy—as grace —makes it the exclusive province of God. One *could* conceive of mercy that way; indeed, this is a plausible reconstruction of the modern sceptical view (discussed above), which wants to insist on the impermissibility of mercy within human institutions. But other positions are available. It is perfectly coherent—and familiar—to speak about human beings dispensing grace to one another. Furthermore, this is true whether or not grace is thought to have any connection to divinity. Even if grace is an inclination of divine origin, that hardly precludes human beings from sharing in its spirit. It may, indeed, embolden that outcome.

In other words, it is possible to imagine human officials as *agents* of grace, vested with the authority to decide, case by case, that otherwise-just punishment ought to be set aside. Premodern political theories made this connection literal, casting sovereign grace as a matter of divine delegation,

[17] I imagine even these observers would be open, at least in principle, to attributing value to mercy in *other* settings—e.g., mercy exercised between soldiers on warring sides of a battle, or mercy exercised by a healer in the face of medical suffering. For present purposes, however, I leave this point to one side.

[18] Paul Twambley, *Mercy and Forgiveness*, 36 Analysis 84, 87 (1976).

whereas modern political theories take a more figurative approach to the "agency" question. Both, however, reach the same end. Grace is an inclination—perhaps of divine origin, perhaps not—to which humans can plausibly aspire, and around which human institutions can be built.

Does this logic extend to machines? Can we imagine machines as "agents of grace" in the same way that we imagine human beings in that mode? No, I want to suggest—because grace, like the mercy it occasions, is *attitudinal or dispositional* in nature. Seneca, the first great defender of mercy as a political virtue, defined it as an "inclination of the soul to mildness in exacting penalties."[19] In practice, this plays out, phenomenologically, as regard for "each particular case as a complex narrative of human effort in a world full of obstacles."[20] Abiding this inclination, the "merciful judge will not fail to judge the guilt of the offender," but she "will also see the many obstacles this offender faced... imagin[ing] what it was like to have been this particular offender, facing those particular obstacles with the resources of [their particular] history."[21]

This operation is not reducible to information-processing, the dry application of abstract rules to concrete facts. It requires imagination and, more importantly, the ability to self-conceive *as an agent*—because it requires the judge to be capable of considering how the world might have seemed from the offender's perspective, whether the judge herself might (or might not) have acted differently in the offender's shoes, and how the offender's moral frailty is connected to the moral frailty of human beings, writ large, simply in virtue of being human.[22] As Martha Nussbaum once put the point:

The merciful attitude requires, and rests upon, a new attitude toward the self. The retributive attitude has a we/them mentality, in which judges set themselves above offenders, looking at their actions as if from a lofty height and preparing to find satisfaction in their pain. The [merciful] judge, by contrast, has both identification and sympathetic understanding.[23]

The merciful judge, in other words, not only regards the offender in a particular way; she also regards *herself* in a particular way. She looks both outward and inward—to the offender, to herself, to all of humanity—in deciding whether lenience is warranted in the particular case. This process

[19] Seneca, De Clementia (Book II Chap. 3).

[20] Nussbaum, note 4 at 102.

[21] *Ibid.*

[22] *See* Brennan-Marquez & Henderson, *Artificial Intelligence and Role-Reversible Judgment*, note 1 (elaborating these dynamics at greater length).

[23] Nussbaum, note 4 at 103.

may benefit from heuristics and parameters, but it admits of no shortcuts. The decision may be easy or difficult, pleasant or painful. But whatever its other qualities, the decision is always—and irreducibly—particular. It is truly about whether lenience *is warranted*, not whether lenience is compelled. For mercy, unlike justice, is never compulsory. It is always a free act.

None of this means, of course, that mercy is always exercised wisely or legitimately in practice. The form of mercy sketched above is a stylized aspiration—not a sociological description. On the ground, especially with respect decisions made "at scale," mercy is typically non-existent, and when it *does* transpire, it often looks more routine than majestic. Worse still, as sceptics like to remind us, the motivation behind particular instances of mercy can be venal, nepotistic, or vindictive. Forbearance can be bargained for and weaponized. It can be made into a political commodity. In short, not every exercise of mercy deserves celebration—far from it. If the possibility of mercy enhances the moral quality of law, it is not because mercy always bespeaks virtue, but in spite of the fact that it sometimes—too often —does not.

At some level, however, the "dark side" of mercy only underscores why the merciful attitude, as an attitude, is likely to elude machines. Mercy requires an inner life, mediated by a sense of frailty that unifies human experience across place, time, and context. This sense of frailty is what allows judges to be "inclined mildly" toward the moral shortcomings of others. And it is also on display when human mechanisms of mercy are corrupted or abused. In that case, frailty is what weakens the moral will of decision-makers, not what allows them to sympathize with the moral weakness of others. Either way, however, the upshot is the same. Mercy requires a kind of self-understanding (1) that machines are unlikely, in principle, to be capable of, and (2) that real-world efforts toward automation tend, in any case, to eliminate. This should give us pause. For it suggests that, in some contexts, the challenges that artificial intelligence pose to public life are more existential than practical—and that familiar fixes, centred on transparency, intelligibility, and democratic process, are unlikely to solve the core problem.