A Moral Agency Framework for Legitimate Integration of AI in Bureaucracies (Extended Abstract)

Chris Schmitz and Joanna Bryson

Centre for Digital Governance, Hertie School, Berlin
ch.schmitz@hertie-school.org, jjb@alum.mit.edu

Abstract

AI systems are presented as intransparent, yet increasingly ascribed varieties of agency implying moral status. Such a context can produce credible, and even self-fulfilling, threats to the rationality, stability, and process legitimacy of bureaucracies deploying AI. We argue here that bureaucracies can *always* adopt AI systems in ways that even *increase* the bureaucracy's own legitimacy, rationality, and even agency, but also that this is by no means a given. Legitimate use of AI in governance mandates careful design across both AI system and institution. We provide a candidate design framework.

Full Paper — https://arxiv.org/abs/2508.08231

The Dual Purpose of Bureaucracy

Bureaucratic public-sector organizations serve two purposes, which (i) are in trade-off, and (ii) each require the moral agency of human bureaucrats. The first is the legitimate implementation of legislation, which Weber (1991) describes as the "compliance model" of bureaucracy. Here, the moral agency of bureaucrats — specifically, their responsiveness to systemic reward and punishment — enables constraining them in an "iron cage," making them predictable, and in aggregate, making the bureaucracy legitimate. The second purpose is the provision of "stewardship", the maintenance of long-term stability for the general public and other branches of government, including resilience to bouts of policy extremism (Heath 2020). This stewardship is enabled by the diversity and plurality of individual bureaucrats, including their moral dispositions (Zacka 2017).

Situating AI in Bureaucracy

We consider two conceptions of AI: as tool and as moral subject. We argue neither fully captures the role of AI in bureaucracy. We resolve the apparent resulting dilemma via an alternative conception of AI as an element of the *institutional structure* of Weberian bureaucracy.

- The "tool" conception falls short because details of AI systems are too intransparent, autonomous, and general to meet the Heidegger (2008) definition of a tool requiring full understanding by individual operators. (Gunkel 2017). Therefore, the conception cannot be used as justification to allocate all responsibility to humans, yet such allocation *is* possible in principle (Lazar 2024).
- The "moral agent" conception is inapplicable because digital systems can't be guaranteed to experience an aversion to punishment (Dennett 1978), or even to maintain sustained identity. These are necessary for structures of responsibility attribution and accountability.
- We formulate the "institutional structure" conception: we conceptualize AI systems as artifacts with no moral agency of their own, which nonetheless amplify, alter, and interact with the moral agency of humans. This is the exact status already ascribed to elements of bureaucratic structure like hierarchy and documented procedure.

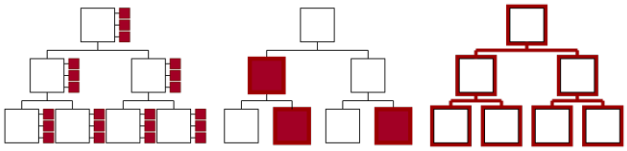

Figure 1: Three conceptions of AI systems (in red) in bureaucracy: (i) as tool, (ii) as a moral subject, (iii) or in our recommended conception, as part of the *institutional structure*, amplifying human agency and accountability.

A Moral Agency Framework

1. **Maintain clear and just human lines of accountability** for any decisions involving or affected by AI systems.
2. **Ensure that humans can verify** the correctness and appropriateness of AI system outputs within their operational context.
3. **Introduce AI only where it does not inhibit** a bureaucracy's ability to pursue legitimacy and stewardship.

In conclusion, the potential threats from AI systems to bureaucratic legitimacy and stewardship are substantial, but procurement of AI deploying a suitable moral framework reduces such threats while enabling access to AI's benefits.

References

Dennett, D. C. 1978. Why you can't make a computer that feels pain. *Synthese*, 38(3): 415–456.

Gunkel, D. J. 2017. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press. ISBN 978-0-262-53463-5.

Heath, J. 2020. *The Machinery of Government: Public Administration and the Liberal State*. Oxford University Press. ISBN 978-0-19-750961-6.

Heidegger, M. 2008. *Being and Time*. HarperCollins. ISBN 978-0-06-157559-4.

Lazar, S. 2024. Legitimacy, Authority, and Democratic Duties of Explanation. In Sobel, D.; and Wall, S., eds., *Oxford Studies in Political Philosophy Volume 10*, 28–56. Oxford University PressOxford, 1 edition. ISBN 978-0-19-890946-0 978-0-19-890949-1.

Weber, M. 1991. *From Max Weber: Essays in Sociology*. Psychology Press. ISBN 978-0-415-06056-1.

Zacka, B. 2017. *When the State Meets the Street: Public Service and Moral Agency*. Harvard University Press. ISBN 978-0-674-54554-0.