A Multidomain Relational Framework to Guide Institutional AI Research and Adoption

Vincent J. Straub[1,*], Deborah Morgan[1,2], Youmna Hashem[1], John Francis[1], Saba Esnaashari[1] and Jonathan Bright[1]

[1]Alan Turing Institute, British Library, 96 Euston Rd., London NW1 2DB, United Kingdom
[2]Department of Computer Science, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom

Abstract

Calls for new metrics, technical standards, and governance mechanisms to guide and evaluate the adoption of ethical Artificial Intelligence (AI) in institutions are now commonplace. Yet, most research and policy efforts do not fully account for all the different approaches and issues potentially relevant to the institutional adoption of AI. In this position paper, we contend that this omission stems, in part, from what we call the 'relational problem': the persistence of differing value-based terminologies to categorize and assess institutional AI systems, and the prevalence of conceptual isolation in the fields that study them including ML, human factors, and social science. After developing this critique, we propose a basic ontological framework to bridge ideas across fields—consisting of three horizontal, discipline-agnostic domains for organizing foundational concepts into themes: Operational, Epistemic, and Normative.

Keywords

Multidomain approach to AI, socio-technical topics, institutions, conceptual framework

Extended Abstract

Research related to the social, political, and legal implications of algorithms and computing is now commonplace [1]. However, most work arguably still fails to account for the diverse potential advantages and consequences of institutional AI adoption. Instead, many contributions tend to foreground only a handful of topics, such as mathematical formulations of outcomes' fairness in ML [2], at the expense of others. How then do we ensure that research on new metrics, technical standards, and governance mechanisms better accounts for all the topics, issues, and methods potentially relevant to the institutional adoption of ethical AI? In this position paper, we focus on one conceptual issue that has arguably hindered existing efforts within the algorithmic fairness community and elsewhere to comprehensively study institutional AI : fundamental ontological questions about the field have not yet been settled—contributing to semantic ambiguity problems, such as a lack of agreed upon definitions for key terms and differing standard terminologies across subcommunities [3]. We contend that this omission stems, in part, from the use of differing value-based terminologies to assess AI systems and the

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
*Corresponding author.
vstraub@turing.ac.uk (V. J. Straub)
https://www.turing.ac.uk/people/researchers/vincent-straub (V. J. Straub)
0000-0003-3393-6027 (V. J. Straub); 0000-0002-0142-9736 (D. Morgan)

prevalence of conceptual isolation in the fields that study institutional AI [4], which we call the 'relational problem'.

While a myriad ways of categorizing issues related to AI development and adoption have been proposed, many ultimately rely on unidimensional thinking. That is, they rely heavily on a single viewpoint or concept—understood both as an abstract idea that offers a point of view for understanding some aspect of experience (e.g., bias), and, relatedly, a mental image that can be operationalized (e.g., measurement bias)—to discuss institutional AI. Much work in the social and policy sciences stresses the ethical and legal challenges at stake in the adoption of AI; while research in computer science tends to highlight the computational and operational aspects that need to be considered. Yet, as pointed out by [5] in discussing scholarship on AI ethics and governance, addressing this shortcoming and uniting the field requires sustained interdisciplinary effort and a richer consideration of the multi-faceted *relation* between concepts.

To address this relational gap, we propose a basic ontological framework, described briefly below, to help bridge terms across fields—consisting of three discipline-agnostic domains for organizing relevant concepts: Operational, Epistemic, and Normative. Our framework aims to achieve two key aims: (1) disciplinary reach, i.e., bridge different subcommunities (ML, human factors, social science, policy etc.), and (2) provide impetus for an intellectual shift that reframes how researchers and key stakeholders (decision-makers, policy creators, advocates, etc.) think about institutional AI systems. Our framework is ontological in the sense that it is composed of three simple domains or meta-concepts that aim to act loosely as semantic fields [6] to guide researchers engaged in studying and conceptualizing institutional AI systems (Figure 1).

Operational Domain The first field is the 'operational domain', which aims to represent the topics, issues, and methods related to the routine activities and functionality of institutional AI systems. It is meant to capture terms that are mainly but not exclusively defined, operationalized, and studied in a technical, applied context. More specifically, it is meant to enable researchers to categorize into a single category all relevant concepts that can be employed both as an abstract idea (e.g., 'accuracy') and easily operationalized to quantitatively measure a specific attribute of a particular institutional AI system (i.e., 'percentage of correct predictions').

Epistemic Domain The epistemic domain aims to capture knowledge-related topics and issues connected to a particular AI system or institutional AI in general. That is, the epistemic domain is meant to help researchers group together concepts that seek to describe properties which pertain to the interface between AI applications and human actors. Both in terms of the knowledge, beliefs, and intentions of those using AI applications (e.g., a desire for transparency), and the internal properties of the system itself (e.g., its interpretability).

Normative Domain The meaning and uses of concepts in the normative domain, the final domain we propose, collectively relate to the entitlements, values, and principles of political morality that stakeholders and affected parties hold towards a particular algorithmic system or institutional AI in general. The term 'political morality' is used here to refer to normative principles and ideals regulating and structuring the political domain [7].

Taken together, the framework's utility derives from the fact that it is discipline-agnostic. More specifically, it aims to be instructive for the individual researcher studying institutional AI in both helping with organizing concepts used to study AI systems and, perhaps more importantly, by drawing attention to whether all potential topics—by virtue of being relevant to one or more of the three proposed domains—have been accounted for. Overall, our contribution

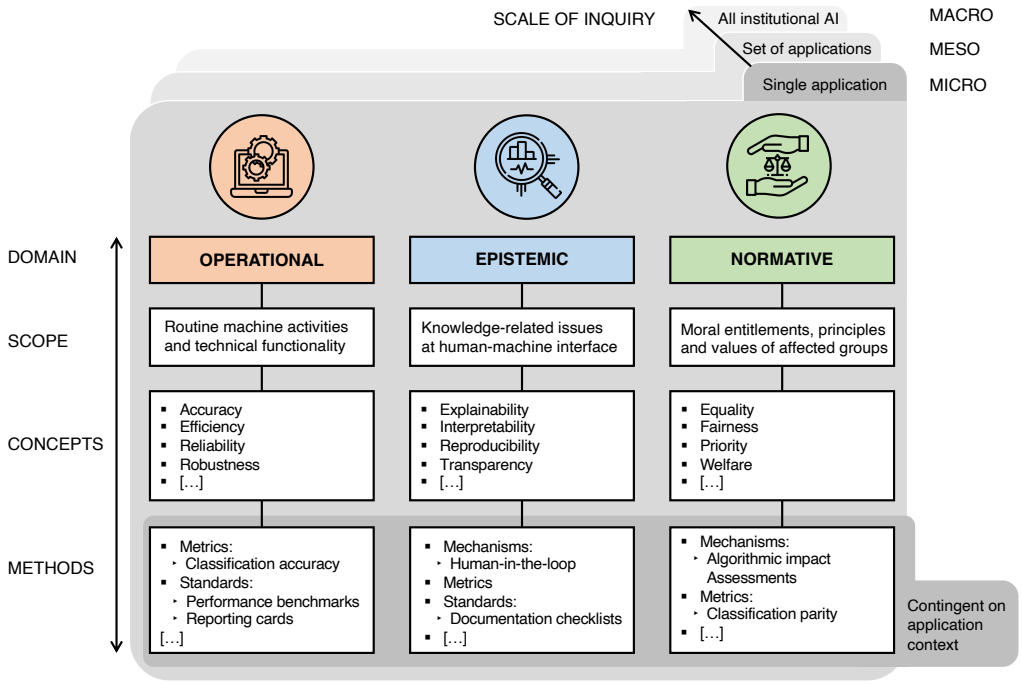

Figure 1: Graphic representation of our framework for organizing concepts relevant to institutional AI research and adoption into three domains: Operational, Epistemic, and Normative.

aims to benefit the algorithmic fairness community by facilitating a constructive dialog around the challenges we face as a diverse, interdisciplinary field.

References

[1] H. Margetts, C. Dorobantu, Rethink government with ai, Nature 568 (2019) 163–165.

[2] B. Laufer, S. Jain, A. F. Cooper, J. Kleinberg, H. Heidari, Four years of facct: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 401–426.

[3] J. Mökander, M. Sheth, D. S. Watson, L. Floridi, The switch, the ladder, and the matrix: Models for classifying ai systems, Minds and Machines (2023) 1–28.

[4] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, M. Bao, The values encoded in machine learning research, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 173–184.

[5] C. Burr, D. Leslie, Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies, AI and Ethics (2022) 1–26.

[6] A. Wierzbicka, Semantic primitives and semantic fields, Frames, fields, and contrasts (1992) 209–227.

[7] E. Erman, M. Furendal, Artificial intelligence and the political legitimacy of global governance, Political Studies (2022) 00323217221126665.