Research Proposal

Machine Learning offers unprecedented potential to automate tasks - but at the same time poses risks such as reinforcing historical biases [1], introducing unwanted feedback loops [2], exacerbating power inequalities, stifling human autonomy [3, 4] and disrupting economic sectors [5]. To systematically address these risks, I propose to advance the debate around democratic governance of deployed machine learning systems. I seek out to build and test participative tools and in parallel advance the theoretical understanding of governance of said systems. Specifically, I want to focus on participative methods to govern reinforcement learning systems in the area of work-management on work platforms.

The need for participative governance:

Machine Learning and reinforcement learning are increasingly used in the management of common pool resources [6], e.g. in ecological control [7] [8], public health [9], smart grids [10], and managing content and attention in online platforms [11]. Planning, deploying and maintaining these systems involves value laden design decisions, resulting in arbitrating between stakeholder interests [12], thus constituting profoundly political processes. Existing and planned governance encompasses product safety regulations like the AI-act as well as various norms and specifications [13]. However, these forms of governance are slow, static, unresponsive, oftentimes barely democratically legitimised [14] and seldomly allow for stakeholder participation or even undermine current provisions [15]. However, stakeholder participation might be the preferential way to accommodate for the non-stationarity of systems, hard to operationalize preferences of stakeholders [16] and dependency of deployment contexts on risk profiles [17, 18].

Therefore, people have been arguing for [12] and exploring [19] participative methods for machine learning [20] and also reinforcement learning systems [21]. However, legal, technical, political and philosophical questions arise: e.g. values and trade-offs are oftentimes hidden in training data, technical opacity, or may emerge over time. Significant design decisions may be distributed along the supply chain, i.e. through pre-trained models, datasets or computing infrastructure. These issues also interact with classical prerequisites for participative governance like transparency, accountability and capacity building.

Work-management as a subject area

Participatory governance in work-management is a) of special importance as an area of application and b) suitable for generalisable research for governance of AI:

a) Work and work-conditions are existential for human livelihood, wellbeing and personal development, as also acknowledged by the “EU-AI act” [13] proposal.

b) in some jurisdictions (e.g. Germany) participatory governance is institutionalised through law and social infrastructure, i.e. works councils and unions [22]. This also means that there already exists a high degree of organisation providing the political infrastructure for further participatory governance.

Within the field of work-management, work-platforms might be a suitable target for research as they exhibit a high level of automatisation and are at the same time notorious for poor working conditions [23]. To address these, worker-owned platforms emerged, e.g. *The Driver Cooperative, CoopCycle.* These organisations would be ideal collaboration partners to test participative governance in practice.

Operationalisation in reinforcement learning systems

A significant body of research exists on how different components of reinforcement learning systems can be adapted: e.g. by specifying action- and state-space [12], testing simulations of the environment [24], defining requirements of the model space [25], auditing learned policies [26] or adjusting the objective function [27].

However, beyond technical questions, political and organizational questions remain: How can political components be reliably defined and identified? Can participatory practices be structured along certain dimensions, e.g. time (planning, deployment, maintenance) or generalizability (singular decisions vs. broader standards)? Here, I would like to explore how political theory of governance, newer experiences with democratic innovations and computer science literature, e.g. on Marr's Three levels of analysis [28], might be usefull for deriving a more general framework for *what*, *how* and *when* to govern those systems using participatory methods.

Toolkit and Evaluation

The toolkit's design and functioning should be informed by these theoretic considerations and be used to evaluate derived hypotheses. The empirical research will encompass a study on how the governance toolkit influences key economic and psychological indicators as well as qualitative analyses. It will take practical lessons from prior studies into account, e.g. from the Algorithmic Equity Toolkit [19], as well as research on the management of non-technical resources and infrastructure [29]. While a significant part of the toolkit will certainly be software, social aspects could play an important part in its functioning.

My reasons for applying to Oxford University

Oxford University offers a unique set of research groups with whom potential collaborations could be explored: At the *Oxford Internet Institute* (*OII*), Divya Siddarth is doing inspiring research on collective intelligences, governance of digital commons and improving low-income workers working conditions. The *FairWork for AI* working group at the *OII* is working on standards and benchmarks for platform work, which will be highly relevant for the empirical analysis. The *AI Governance* Research Group at the *Future of Humanity Institute* explores cooperative AI from the perspective of game-theory and multi-agent systems, which might be critical for understanding the general dynamics socio-technical system exhibit.

Reuben Binns is a Professor of Human Centred Computing. His research connects political philosophy and law with lower-level technical operationalization. This will also be the main challenge of the research I seek out to do and he would therefore be my preferred supervisor.
Legal aspects are a further crucial factor of bringing participative governance methods into practice. Jeremias Adams-Prassl is a professor of law at oxford university and focuses among other things on algorithmic work-management in work platforms. With him as my second supervisor, I would like to explore the legal foundations and implications of participative governance.

My academic and professional experience in human-, machine- and hybrid-learning systems puts me into a unique position to conduct the proposed research. I have experience engineering and analysing the reinforcement learning systems in question, as well as connecting them to higher-level legal concepts within interdisciplinary research teams. Beyond this, I gathered practical experience in the domain in question, by doing interviews with unions and works councils on organisational constraints, work-management systems and collective bargaining. I'm committed to both understanding sociotechnical systems and designing them in a way that increases human welfare and agency.

994 Words

Bibliography

[1] S. U. Noble, Algorithms of Oppression: How Search Engines Reinforce Racism, New York: NYU Press, 2018.

[2] S. Barocas, M. Hardt and A. Narayanan, "Feedback and feedback loops," in *Fairness and Machine Learning*, Nips tutorial, 2017, pp. 13-15.

[3] C. Prunkl, "Human autonomy in the age of artificial intelligence," *Nature Machine Intelligence,* vol. 4, pp. 99-101, 2022.

[4] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao and N. Shadbolt, "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions," in *Conference on Human Factors in Computing Systems*, 2018.

[5] A. Korinel and J. E. Stiglitz, "Artificial Intelligence and Its Implications for Income Distribution and Unemployment," in *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019, pp. 349-390.

[6] E. Ostrom, "Institutions and Common-Pool Resources," *Journal of Theoretical Politics,* vol. 4, no. 3, pp. 243-245, 1992.

[7] C. J. Fonnesbeck, "solving dynamic wildlife resource optimization problems using reinforcement learning.," *Nature Resource Modeling,* vol. 18, no. 1, pp. 1-40, 2008.

[8] D. Silvestro, S. Goria, T. Sterner and A. Antonelli, "Improving biodiversity protection through artificial intelligence," *Nature Sustainability,* vol. 5, pp. 415-424, 2022.

[9] H. Bastani, K. Drakopoulos, V. Gupta, I. Vlachogiannis, C. Hadjichristodoulou, P. Lagiou, G. Magiorkinis, D. Paraskevis and S. Tsiodras, "Efficient and targeted COVID-19 border testing via reinforcement learning," *Nature,* vol. 599, pp. 108-113, 2021.

[10] D. Zhang, X. Han and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems,* vol. 4, no. 3, pp. 362-370, 2018.

[11] J. Stray, I. Vendrov, J. Nixon, S. Adler and D. Hadfield-Menell, "What are you optimizing for? Aligning Recommender Systems with Human Values," *arXiv:2107.10939 ,* 2021.

[12] R. Dobbe, G. T. Krendle and Y. Mintz, "Hard choices in artificial intelligence," *Artificial Intelligence,* vol. 300, 2021.

[13] M. Veale and F. Borgesius, "Demystifying the Draft EU Artificial Intelligence Act," *Computer Law Review International,* vol. 22, no. 4, 2021.

[14] A. Fung and E. O. wright, "Deepening Democracy: Innovations in Empowered Participatory Governance," *Politics & Society,* vol. 29, no. 1, pp. 5-41, 2001.

[15] A. Cefaliello and M. Kullmann, "Offering false security: How the draft artificial intelligence act undermines fundamental workers rights," *European Labour Law Journal,* vol. 13, no. 4, pp. 542-562, 12 2022.

[16] B. Wine, S. Gilroy and D. A. Hantula, "Temporal (In)Stability of Employee Preferences for Rewards," *Journal of Organizational Behavior Management,* vol. 32, no. 1, pp. 58-64, 2012.

[17] J. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," in *Equity and Access in Algorithms, Mechanisms, Optimization*, Ney York, 2021.

[18] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji and T. Gebru, "Model Cards for Model Reporting," in *Conference on Fairness, Accountability, and Transparency*, 2019.

[19] P. M. Krafft, M. Young, M. Katell, J. E. Lee, S. Narayan, M. Epstein, D. Dailey, B. Herman, A. Tam, V. Guetler, C. Bintz, D. Raz, P. O. Jobe and F. Putz, "An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists," *Conference on Fairness, Accountability, and Transparency,* pp. 772-281, 2021.

[20] A. Zhou, D. Madras, D. Raji, S. Milli, B. Kulynych and R. Zemel, "Workshop: Participatory Approaches to Machine Learning," International Conference on Machine Learning Workshop, 7 2020. [Online]. Available: https://icml.cc/virtual/2020/workshop/5720. [Accessed 4 12 2022].

[21] T. Gilbert, S. J. Russell, T. O. Zick, A. Snoswell and M. Dennis, "Workshop: Political Economy of Reinforcement Learning Systems," 14 12 2021. [Online]. Available: https://neurips.cc/virtual/2021/workshop/21864. [Accessed 4 12 2022].

[22] R. Hyman, "How can trade unions act strategically?," *Transfer: European Review of Labour and Research,* vol. 13, no. 2, 2007.

[23] M. Graham, J. Woodcock, R. Heeks, P. Mungai, J.-P. Van Velle, S. Fredman, A. Osiki, A. van der Spuy and S. M. Silberman, "The Fairwork Foundation: Strategies for improving platform work in a global context," *Geoforum,* vol. 112, pp. 100-103, 2020.

[24] J. Huang, H. Oosterhuis, M. de Rijke and H. van Hoof, "Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning Based Recommender Systems," *Conference on Recommender Systems,* p. 190–199, 2020.

[25] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence,* vol. 1, pp. 206-215, 2019.

[26] K. Zhang, H. Wang, J. Du, B. Chu, A. Robles Arévalo, R. Kindle, L. A. Celi and F. Doshi-Velez, "AN interpretable RL framework for pre-deployment modeling in ICU hypotension management," *npj Digital Medicine,* vol. 5, no. 173, 2022.

[27] E. J. Michaud, A. Gleave and S. Russel, "Understanding Learned Reward Functions," in *Deep RL Workshop, NeurIPS*, 2020.

[28] D. Marr, Vision: A computational investigation into the human representation and processing of visual information, San Fransico: Freeman, W.H., 1982.

[29] J. Hinkel, P. W. G. Bots and M. Schlüter, "Enhancing the Ostrom social-ecological system framework thorugh formalization," *Ecology and Society,* vol. 19, no. 3, pp. 51-70, 2014.

[30] M. Lapeyrolerie, M. S. Chapman, N. K. E. A. and C. Boettiger, "Deep reinforcement learning for conservation decisions," *Methods in Ecology and Evolution,* vol. 13, no. 11, pp. 2649-2662, 2022.