Philosophy & Technology (2024) 37:82
https://doi.org/10.1007/s13347-024-00771-7

COMMENTARY

The Limits of Calibration and the Possibility of Roles for Trustworthy AI

Ulrik Franke[1,2]

Received: 13 June 2024 / Accepted: 21 June 2024

Abstract
With increasing use of artificial intelligence (AI) in high-stakes contexts, a race for "trustworthy AI" is under way. However, Dorsch and Deroy (*Philosophy & Technology* 37, 62, 2024) recently argued that regardless of its feasibility, morally trustworthy AI is unnecessary: We should merely rely on rather than trust AI, and carefully calibrate our reliance using the reliability scores which are often available. This short commentary on Dorsch and Deroy engages with the claim that morally trustworthy AI is unnecessary and argues that since there are important limits to how good calibration based on reliability scores can be, some residual roles for trustworthy AI (if feasible) are still possible.

Keywords Trustworthy AI · Reliable AI · Calibration · Second-level bias · Generative AI

1 Introduction

With increasing use of artificial intelligence (AI), the ways in which it can fail have come under scrutiny. In particular, AI systems have turned out to be insecure, biased, and brittle. Such worrying findings have propelled a race for "trustworthy AI", a concept used for example by the EU High-Level Expert Group on AI (2019) and in the EU Artificial Intelligence Act. However, this has been criticized for misusing the concept of trust and unnecessarily anthropomorphizing AI (Ryan, 2020).

Dorsch and Deroy (2024), in their contribution to this critical literature, observe that AI systems trained on appropriate datasets with ground truth are often reliable and thus *prima facie* trustworthy. However, they argue, trust is a moral concept which requires more than mere reliability, namely, knowing that the trusted agent is morally rational

✉ Ulrik Franke
ulrik.franke@ri.se

1 RISE Research Institutes of Sweden, SE-164 29, Kista, Sweden

2 KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden

Published online: 01 July 2024

in the sense that it can be expected to do the right thing for the right reasons (p. 9). Terms such as "trust" or "trustworthy" should be reserved only for agents exhibiting moral rationality, avoiding unwarranted AI-anthropomorphization.

However, while Dorsch and Deroy are clearly skeptical about whether it is even possible to build morally rational machines (see their sections 3.1 and 3.2), their main argument against trustworthy AI is not that it is *infeasible*, but that it is *unnecessary*. They observe that trust is a solution to a particular problem of epistemic asymmetry between what we ought to know and what we actually know when deciding whether to rely on another agent for a particular novel task. We want to know how well the agent performs that task—such calibration enables informed decisions, without any need for trust. However, if we do *not* know this, e.g., because of task novelty, we may still trust the agent with the task, "by appealing to the agent's moral trustworthiness in lieu of her task-relevant reliability" (p. 13). *This* is where trust is needed. Now, argue Dorsch and Deroy, in machine learning applications trained on datasets with ground truth, trust is not relevant. The task is not novel and there are plenty of reliability scores such as precision, recall, false positive rate, etc. which suffice for calibration. Thus, Dorsch and Deroy conclude that trustworthy AI "is unnecessary for promoting optimal calibration of the human decision-maker on AI assistance. Instead, reliability, exclusive of trust, ought to be the appropriate goal of ethical AI" (p. 18).

This short commentary on Dorsch and Deroy (2024) does not address the feasibility of morally trustworthy AI. Instead, its purpose is to argue that calibration on reliability scores—though we should certainly use this to the full extent possible—does not exhaust what a human deliberating whether to use AI for some particular task should know. Some residual needs persist after calibration, and trustworthy AI (if feasible) would go some way towards addressing them. This does not amount to the strong claim that trustworthy AI is necessary, merely to the refutation of the claim that it is unnecessary.

2 The Limits of Calibration on Reliability Scores

Dorsch and Deroy (2024) delimit their argument to "machine learning applications that have been trained on a dataset about which there is a ground truth" (p. 4), used in "high-stakes decision-making environments" (p. 6). Clearly, many important applications fall within this scope. For example, algorithmic fairness is often discussed in the context of classification or prediction systems trained on datasets with ground truth for use in high-stakes situations, e.g., finding out who should be employed, treated, or granted bail. In such cases the argument of Dorsch and Deroy is compelling: calibration with respect to reliability seems enough and trust over and above it unnecessary.

But reliability scores also have well-known limitations. Confronted with a score such as F_1, a user should ask additional task-specific questions. For example, is the training data old? If so, its properties may have changed, and special methods are needed (Ditzler & Polikar, 2012). Is the training data imbalanced? This also requires special methods (Jeni et al., 2013). Is the training data representative? Many models

perform well for some groups and poorly for others (Cavazos et al., 2020; Koenecke et al., 2020). Did the training include adversarial examples? If not, performance may be much worse than indicated by the F_1 score, since classifiers are more robust to random noise than to adversarial perturbations (Fawzi et al., 2018).

A unifying explanation of the limitations of reliability scores is that all tasks exhibit *some* novelty. Since the actual data encountered in the task is not identical to the training data, we may say with Heraclitus that the AI never steps into the same data twice. While calibration is certainly possible, it is not exhaustive. These limitations of reliability scores open for the possibility of a minimal residual role for trust.

3 A Minimal Residual Role for Trustworthy AI: Selection of Reliability Scores

We may agree with Dorsch and Deroy that whenever we have appropriate reliability scores, we should use them for calibration, thus minimizing the need for trust. But even so, a question of *second-level bias* (Franke, 2022) is raised: How do we select appropriate reliability scores in an unbiased and impartial way? To illustrate the relevance of the question, observe (i) that using the wrong reliability scores can be misleading (see, in addition to the literature cited in Section 2, also Chicco and Jurman, 2020; Fourure et al., 2021; Yao and Shepperd, 2021), and (ii) that second-level bias can be very difficult to avoid or detect (see Franke, 2022, Section 4).

Ground truth training data and reliability scores notwithstanding, in any practical application there is some task novelty (the training data is not the same as the actual data), casting at least a shadow of a doubt on the relevance of the scores (sure, the model performs well according to these scores, but do we measure the right things?), entailing some remaining epistemic asymmetry between what we know and what we ought to know.

This asymmetry can be overcome in several, not necessarily mutually exclusive, ways. One possibility is to trust humans—engineers, lawyers, marketers, regulators—to select the right scores. A second possibility is to calibrate ourselves to how good institutions such as a legal system or market competition are at finding the right scores. A third possibility, however, is to (try to) build what Dorsch and Deroy call morally rational artificial agents which we could trust. To be clear, such trust need not *replace* calibration to the scores at hand, but rather *complement* it (hence the *residual* role for trust). It is one thing to know that (i) "this AI has $F_1 = x$ on this task and the F_1 score may or may not be a good measure", another that (ii) "this AI has $F_1 = x$ on this task, the F_1 score may or may not be a good measure, and this AI is sensitive to moral norms".

This is not to say that this trust solution is, generally or ever, the preferred one (for one thing, morally trustworthy AI may not be feasible). But the claim that "such machines would not help mitigate the vulnerability that deploying such technology creates" (Dorsch & Deroy, 2024, p. 14) seems too strong—trustworthy AI could at least play this minimal residual role .

4 A Larger Residual Role for Trustworthy AI: Systems for Open-Ended Problems

Could there be more than this minimal role for trust? There is an increasing number of AI systems outside the delimitations of Dorsch and Deroy: they do *not* undergo supervised training to solve specific problems using datasets with ground truths. Instead, as the name suggests, a generative pretrained (GPT) model is largely built on unsupervised pretraining on large datasets (sometimes but far from always followed by supervised fine-tuning) to solve more general open-ended problems. Such generative AI represents some of the most spectacular recent advances in AI—ChatGPT, LLaMA, DALL-E, Midjourney, etc.

How can we to calibrate ourselves to generative AI? We may have plenty of scores available. Consider GPT-3, a large language model (LLM): There are scores such as accuracy, perplexity, and F_1 available for a range of cloze, completion, question-answering, and translation tasks (Brown et al., 2020). However, the generality of the model make such scores much less suitable for calibration than corresponding scores for specific problems. Knowing F_1 or accuracy scores on a standardized benchmark such as SuperGLUE (Wang et al., 2019) only gives a *general* idea of how well the LLM will perform in the vast range of possible tasks it could meaningfully address; not a *specific* measure for the concrete task at hand. Thus, epistemic asymmetry seems an inescapable feature of generative AI.

For particular tasks, of course, the asymmetry can be dispelled. Employing an LLM for medical diagnosis, we can measure scores such as F_1 for particular diagnostic tasks (Yang et al., 2022), rather than relying on general scores. (Still, as discussed in Section 3, the question of second-level bias remains, which is why finding the *right* scores is an ongoing endeavor; see, e.g., Abbasian et al. 2024.) But finding such specific scores is not possible for *all* the possible tasks which generative AI such as LLMs can address. Indeed, their power stems from their ability to generalize—to *not* need specific training with ground truth for each and every task. Such generality means precisely that there will not be specific scores available for calibration with respect to every specific task.

Dorsch and Deroy suggest that we should simply not use AI to address open-ended, novel problems:

> Considering the above interpretation of the significance of trustworthiness, one could interject that designing AI to be morally trustworthy would mean that we would be justified in deploying it in novel situations, which, in this context, would mean deploying it without the required training. To our knowledge, this would violate ethical codes as well as be simply a bad idea from an engineering standpoint, and so the point is somewhat moot, since its practical significance is unclear. (Dorsch and Deroy, 2024, p. 14, footnote 11)

In many cases, this makes sense. If there is available training data, it is clearly a bad idea to deploy AI in a high-stakes classification problem without training it on the data beforehand. Similarly, it is a bad idea to deploy general purpose systems such as ChatGPT to well-defined closed problems where more specialized, better trained systems do better (Kocoń et al., 2023), at least when the stakes are high.

But abstention will hardly be universal. Even assuming *ethical* commitment to using generative AI only in low-stakes situations, this is complicated by *epistemic* difficulties: First, the nature of open-ended problems is such that the stakes are not obvious. (If you build an AI model to assess child mistreatment allegations, it is clear that the stakes are high. But if you build an AI model with a general capacity to process natural language, it can be used for all kinds of purposes with all kinds of stakes.) Second, the fact that so many different actors with different roles are involved in a modern AI system (Barclay & Abramson, 2021) means that no single individual may be in a good position to understand *both* the nature of the AI system and the nature of the problem it is used to solve.

The general scores available for AI designed to address open-ended problems are not as good a basis for calibration as are the specific scores available for AI trained for classification problems. The epistemic asymmetry persists, suggesting a possible residual role for trustworthy AI.

5 Concluding Remarks

Dorsch and Deroy (2024) conclude that the idea of "developing morally trustworthy AI is fundamentally wrongheaded" because it is "unnecessary for promoting optimal calibration of the human decision-maker on AI assistance" (p. 18). However, in Section 2 we observed that there are important limits to how good such calibration based on reliability scores can be, suggesting that trust may have a residual role to play, even after we have made the most of calibration. More precisely, we have argued in Section 3 that morally trustworthy AI may have a role to play in overcoming concerns about second-level bias in the selection of reliability scores. This minimal role is possible even when reliability scores come from training an AI model on the particular problem at hand, using ground truth data. Furthermore, many (generative) AI models are capable of addressing more general problems, so that reliability scores from their training or testing are too general to be a good basis for calibration in a particular case at hand. In Section 4, we have argued that in such cases—which are beyond the delimitations set by Dorsch and Deroy—morally trustworthy AI may have a somewhat larger role to play.

That said, the project of Dorsch and Deroy remains an attractive one. We should certainly strive to make the most of calibration using reliability scores (broadly construed to include XAI techniques and other effective means to communicate the strengths and weaknesses of AI models). But even if we do so, we have argued that there will remain some residual needs, not necessarily met by calibration. Here, some possible roles for morally trustworthy AI (if feasible) remain.

Author Contributions Not applicable (single author).

Funding Open access funding provided by RISE Research Institutes of Sweden. The author received no external funding for this work.

Availability of Data and Material Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Yes.

Competing Interests The author declares no conflict of interest.

References

Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Abad, Shakeri Hossein, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin. B, et al. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine, 7*(1), 82. https://doi.org/10.1038/s41746-024-01074-z

Barclay, I., Abramson, W. (2021). *Identifying roles, requirements and responsibilities in trustworthy AI systems*. Association for Computing Machinery, Inc, pp. 264–271. https://doi.org/10.1145/3460418.3479344

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science.* https://doi.org/10.1109/TBIOM.2020.3027269

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 1–13. https://doi.org/10.1186/s12864-019-6413-7

Ditzler, G., & Polikar, R. (2012). Incremental learning of concept drift from streaming imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 25*(10), 2283–2301. https://doi.org/10.1109/TKDE.2012.136

Dorsch, J., Deroy, O. (2024). Quasi-metacognitive machines: Why we don't need morally trustworthy AI and communicating reliability is enough. *Philosophy & Technology*, *37*(62). https://doi.org/10.1007/s13347-024-00752-w

Fawzi, A., Fawzi, O., & Frossard, P. (2018). Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning, 107*(3), 481–508. https://doi.org/10.1007/s10994-017-5663-3

Fourure, D., Javaid, M.U., Posocco, N., Tihon, S. (2021). Anomaly detection: How to artificially increase your F1-score with a biased evaluation protocol. In: Joint European conference on machine learning and knowledge discovery in databases (pp. 3–18). Springer. https://doi.org/10.1007/978-3-030-86514-6_1

Franke, U. (2022). First- and second-level bias in automated decision-making. *Philosophy & Technology*, *35*(21). https://doi.org/10.1007/s13347-022-00500-y

High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Jeni, L.A., Cohn, J.F., De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction (pp. 245–251). IEEE. https://doi.org/10.1109/ACII.2013.47

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion, 99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems 32*

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., et al. (2022). A large language model for electronic health records. *NPJ Digital Medicine, 5*(1), 194. https://doi.org/10.1038/s41746-022-00742-2

Yao, J., & Shepperd, M. (2021). The impact of using biased performance metrics on software defect prediction research. *Information and Software Technology, 139*, 106664. https://doi.org/10.1016/j.infsof.2021.106664