From Policy to Practice: Research Directions for Trustworthy and Responsible AI "By Design"

Ramayya Krishnan
AI Measurement Science & Engineering Center
Heinz College of IS and Public Policy
Carnegie Mellon University
rk2x@cmu.edu

John P. Lalor
Human-centered Analytics Lab
Department of IT, Analytics, and Operations
University of Notre Dame
john.lalor@nd.edu

Nicolas Prat
Department of IS, Data Analytics, and Operations
ESSEC Business School
prat@essec.edu

Ahmed Abbasi
Human-centered Analytics Lab
Department of IT, Analytics, and Operations
University of Notre Dame
aabbasi@nd.edu

Abstract

Rapid advancements in the development and adoption of artificial intelligence (AI) have accelerated the need for trustworthy and responsible AI. National/international AI governance and risk management policies and frameworks have identified a core set of tenets for trustworthy and responsible AI, including but not limited to, fairness, safety, privacy, security, transparency, explainability, and responsible deployment. Responsible AI processes/tools (RAPs) are solutions designed to operationalize and implement the tenets, serving as a middle-layer between the tenets and real-world AI-embedded processes. In recent years, the design of RAPs has emerged as an important avenue for computational and social science researchers, practitioners, and policy-makers. We highlight six important research directions for the design of RAPs. Using a real-world case study, we describe the importance of each research direction and illustrate current challenges.

1 Overview

The impact of artificial intelligence (AI) can be viewed from the perspective of people, process, and technology. The rise of state-of-the-art (SOTA) foundation models capable of assessment (i.e., predictive inference) and generation (i.e., multimodal GenAI for text, image, video, audio, etc.) has opened up a bevy of opportunities for processes conceived or enriched by AI. AI-embedded processes are ones where advancements in AI's ability to assess/infer and/or generate are used to *automate* or *augment* existing, traditionally human-guided, processes. The role of – and impact on – people in AI processes cannot be overstated. As depicted in the bottom part of Figure 1, AI processes are disrupting *labor supply chains* [1] with implications for the future of work, the role of the human-in-the-loop (HITL), and the economic and humanistic implications of *exposure* to AI versus *substitution* due to AI [2].

These advancements underscore the importance of trustworthy and responsible AI (TRAI). Guided by normative goals and national/international AI governance and risk management frameworks and policies [3],[1] the tenets of TRAI include: *fairness* and mitigation of harmful bias; *safety* and alignment with legalities and human values; *privacy* in protecting personal data; *security* and resiliency against adversarial attacks; *responsible deployment* to increase opportunity, access, and productivity; *transparency* and accountability in design/training/alignment data and mechanisms; and *explainability* and interpretability of specific model decisions, and underlying decision-making processes, respectively. Notably, this list of TRAI tenets (depicted in the top part of Figure 1) is illustrative, not exhaustive.

[1] For example, NIST AI Risk Management Framework: https://www.nist.gov/itl/ai-risk-management-framework; EU AI Act: https://artificialintelligenceact.eu/

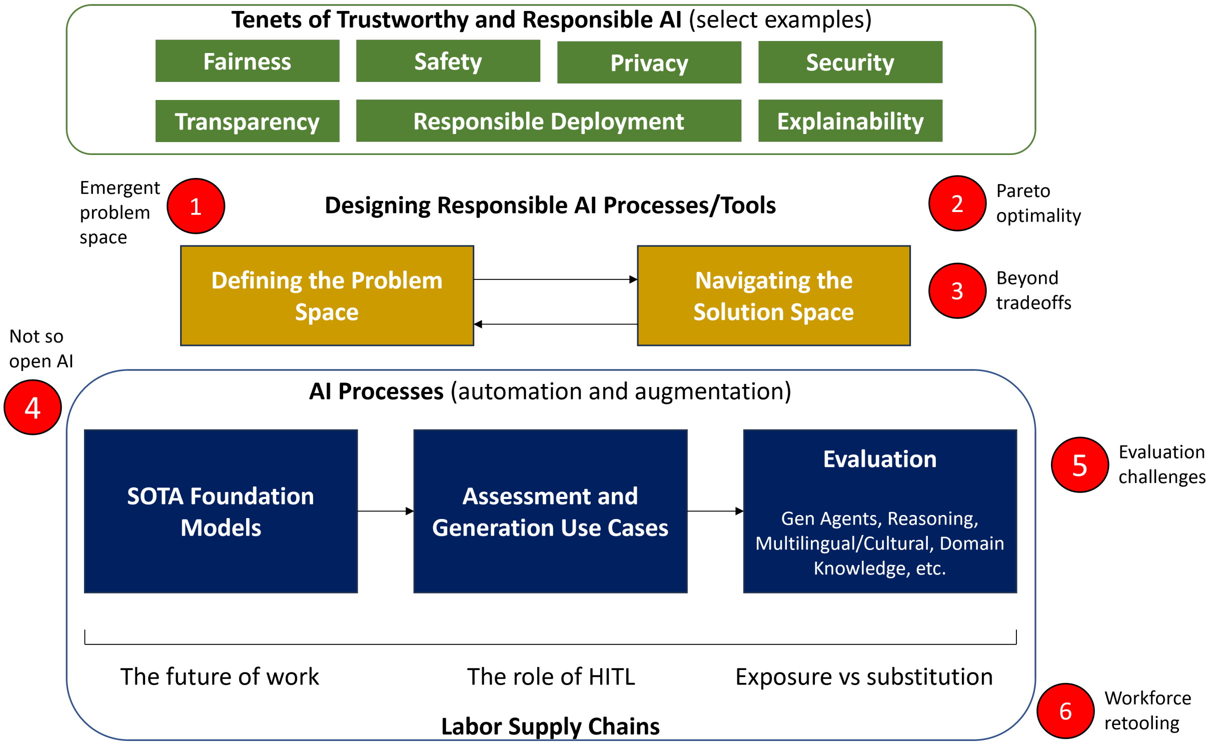

Figure 1: Six Research Directions for Responsible AI "By Design". The figure depicts AI processes (bottom), some example tenets of Responsible AI (top), the need for Responsible AI processes (middle), and six challenging research directions (red numbered circles).

Responsible AI processes/tools (RAPs) are solutions designed to operationalize and implement the tenets of TRAI in AI processes. More specifically, RAPs are intended to serve as a middle-layer between the tenets and the real-world settings in which AI manifests, by supporting key governance functions such as mapping, measuring, and managing risks (adapted from NIST[1]). In recent years, the design of RAPs has emerged as an important avenue for computational and social science researchers. From Simon's classical "sciences of the artificial" perspective, design can be considered a problem-solving paradigm [4] comprising the proposal of novel solutions to well-defined problems. In this case, the problems of interest being how best to design RAPs to operationalize (i.e., map, measure, manage, and govern) the tenets of TRAI (middle of Figure 1). The purpose of this article is to highlight six important research directions for the design of RAPs (numbered circles in Figure 1). Using a real-world case study, we describe the importance of each research direction and illustrate current challenges.

2 Six Important Research Directions

We use a real-world healthcare example to guide our discussion, and to illustrate some of the nuanced challenges and opportunities pertaining to each of the six research directions. Based on the mantra that "prevention is better than cure," the AI process depicted in Figure 2 relates to the use of text-message based nudges to encourage proactive health behaviors [5], such as not canceling an upcoming annual checkup appointment. Trained AI models are used to: (1) predict those most likely to cancel an appointment; (2) to send messages based on users' levels of anxiety visiting the doctor's office, with message content varying based on their predicted level of health literacy (anxiety and literacy are inferred based on their prior mobile activity, survey responses, text, and/or clinical data). Lower health literacy and high anxiety have been found to be important impediments to future doctor visits [6]. In this example, the desired RAP is to send the AI-based nudges in a manner where the messaging is best aligned with users such that appointment cancellations are minimized, while also adhering to the tenets of TRAI. For brevity, in our example, we mostly use the tenet of fairness (with some discussion of privacy) to illustrate the six research directions. The overarching objective of fairness is to reduce variance in model performance across protected attributes such as demographics (e.g., age, gender, etc.). We use fairness as our focal TRAI tenet in part because it has garnered considerable attention in the literature [7, 8]. All results presented for this health analytics example are based on 8,502 users.

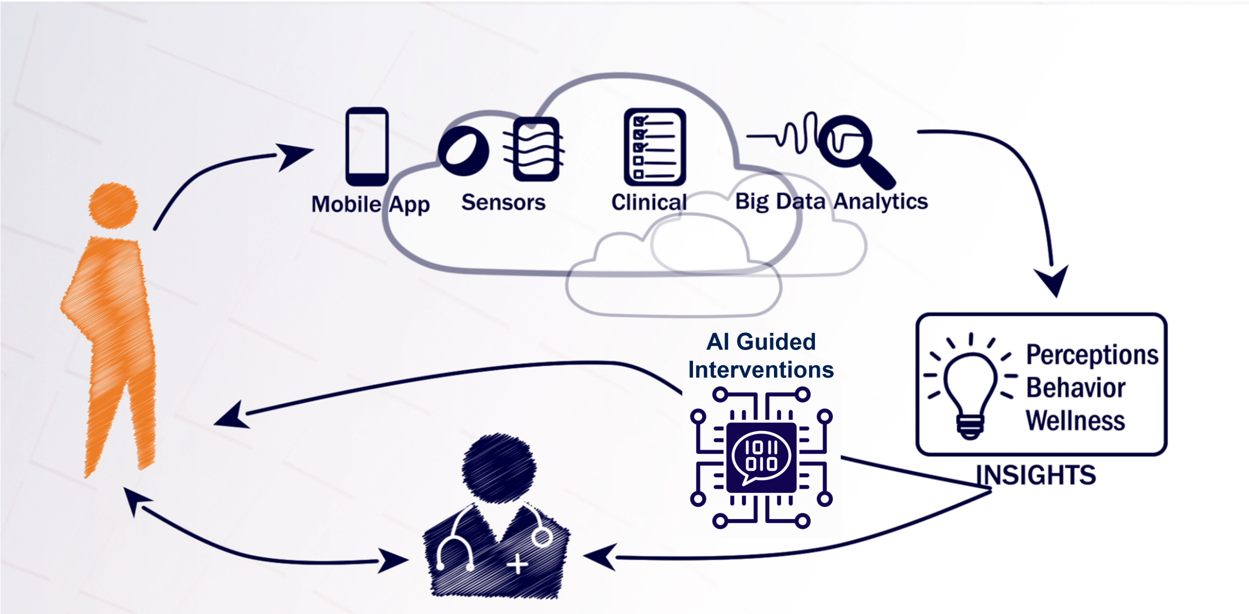

Figure 2: Health Analytics Example to Illustrate the Six Research Directions.

2.1 Emergent Problem Space

Traditionally, design research has focused on proposing solutions to well-defined problems [4]. A problem may be characterized as the difference between an existing state and a desired state [4], and the goal is to get from the existing state to the desired one. How should we problematize the tenets of TRAI? If the goal (desired state) of RAPs is to support the governance functions of mapping, measuring, and managing risks, what does the problem space look like? When it comes to TRAI, we argue that the problem space is highly complex, emergent, and ill-defined. In regards to fairness, one survey identified 23 types of bias, 10 definitions of fairness, and noted that reconciling and synthesizing different perspectives of fairness into a single definition/problem space remains a top challenge [7]. Furthermore, fairness measures of biases materializing upstream – model representational harm due to pretraining or fine-tuning – do not correlate well with downstream allocational harm due to unfair allotment of resources or opportunities [9]. This issue is depicted in Figure 3a, which shows gender-related fairness metrics for two anxiety and literacy inferring language models (BERT and DeBERT) across three stages: stereotyping in pretraining, representational harm in (upstream) fine-tuning, and allocational harm in (downstream) decision-making. Importantly, the pretraining fairness metrics (average of SEAT-6 and SEAT-8 scores), and most of the seven upstream fairness metrics (disparate impact, etc., in the middle of each chart), consider the biases to be in the opposite direction (positive values) relative to the downstream allocational harm (negative values). More specifically, the pretrained stereotype and upstream representational harm fairness metrics are suggesting that the LLMs are overly associating anxiety with female patients, however the downstream suggests that in fact, the female patients are not receiving sufficient anxiety-alleviation text message nudges. Any upstream fairness processes/algorithms would further exacerbate downstream misallocation (i.e., sending increasingly misaligned quantity and types of text-message nudges to women versus men). Additionally, changes in the environment may alter the existing and desired states and related goals [4], such as advancements in the SOTA (e.g., masked, autoregressive, mixture-of-expert language models). The multi-faceted, non-stationary, and amorphous nature of the problem poses challenges for the design and development of solutions.

2.2 Pareto Optimality – Is Satisficing Possible?

If the problem space were well-defined, to account for the complexity of problems, Simon [4] defined the concept of satisficing (as opposed to optimal) solutions. A solution is satisficing if it meets aspirations along all criteria (i.e., Pareto optimality). However, in the case of TRAI, the highly multi-faceted nature of the problem space challenges the very notion of satisficing solutions. How should the satisfactory thresholds be defined for different criteria? Thresholds may not be overly difficult to define for economic or technological criteria, but what about other dimensions like ethicality – when can we consider a solution to be “ethical enough?” In the case of fairness, the protected attributes may include demographics such as gender, race, and age, resulting in two-way and three-way interaction effects – often referred to

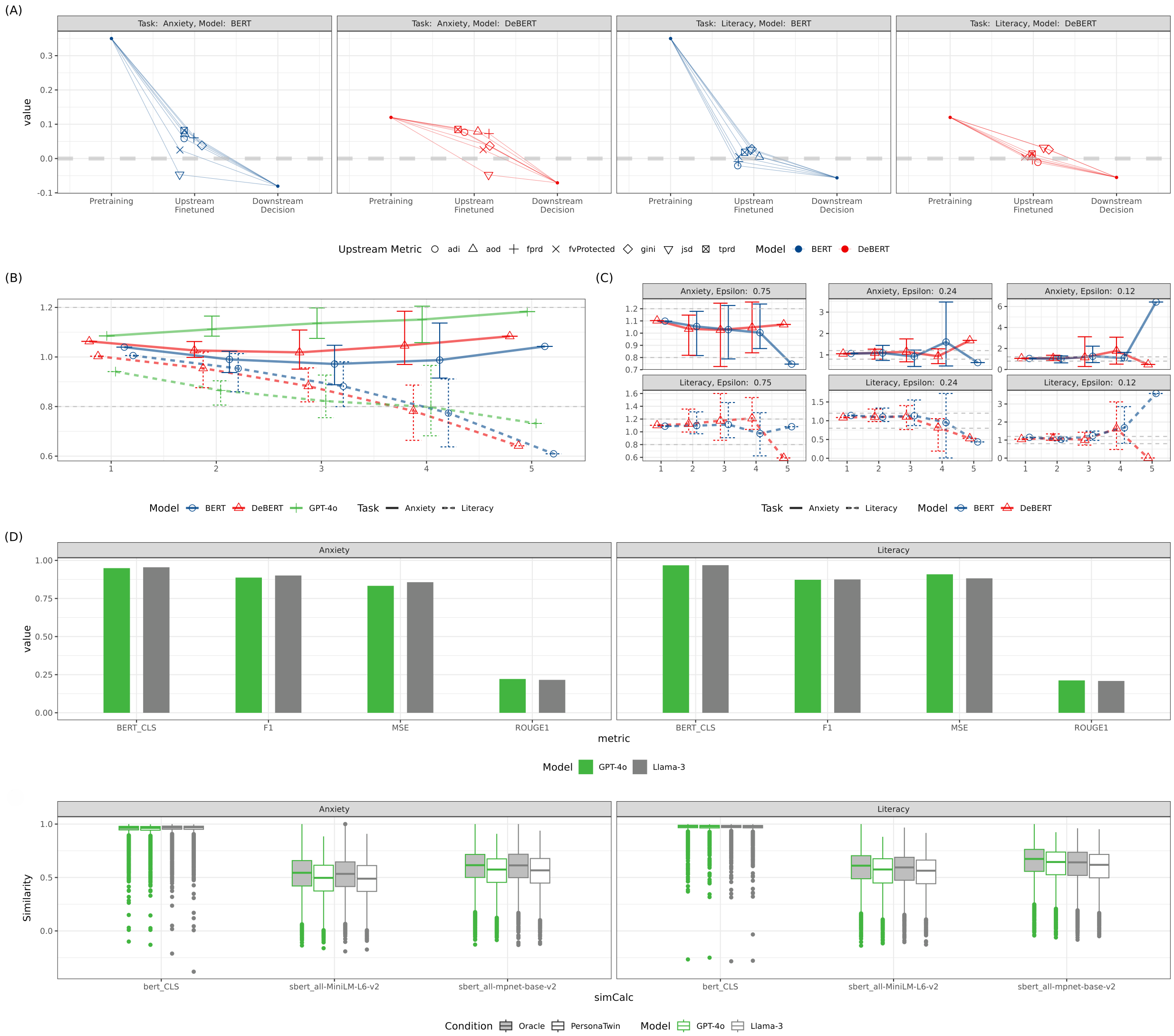

Figure 3: Results from Health Analytics Example Related to Select Research Directions: (a) emergent problem space; (b) pareto optimality; (c) beyond tradeoffs; (d) evaluation challenges for generative AI

as intersectional bias [10, 11]. As the interaction combinations increase, so do the number of needed thresholds. Should we be fairer to older men or younger women? Consequently, the potential range of biases can be amplified, whereas the effectiveness of debiasing methods degrades [10, 11]. For the anxiety and literacy scoring AI models in our health analytics example, this issue is illustrated in Figure 3b – as we add multiple protected attributes (e.g., demographics such as age, gender, race, education, and income along the x-axis), mean disparate impact (DI), and range of DI both increase considerably (y-axis). Importantly, this holds for fine-tuned language models (BERT), debiased language models (DeBERT), and in-context learning LLMs (GPT-4o).

2.3 Beyond Tradeoffs – Fairly Private or Privately Fair?

The prior discussion (and illustrative example) was within a single TRAI tenet, that is, satisficing within fairness. When adding an additional tenet to the equation, the notion of satisficing breaks down completely. We intentionally use privacy to illustrate this point because the notion of privacy is, in some respects, at least in practice, antithetical to fairness. Fairness requires some knowledge of protected attributes (to ensure debiasing, fairer models, better alignment in LLMs, etc.). However, privacy relates to having the ability to protect disclosure of said protected attributes. Revisiting our health analytics example, this tension between privacy and fairness is illustrated in Figure 3c. As differential privacy values go up from 100 (no privacy - Figure 3b) to 0.75 (very little privacy – left charts in Figure 3c) to 0.12 (fairly private – rightmost charts in Figure 3c), the range and mean DI increases 3 to 5 fold for the non-debiased language models (BERT) and doubles for the debiased one (DeBERT). The example underscores the fact that although the principles of privacy and fairness are important and complementary tenets of TRAI, their operationalizations produce unintended and undesirable tradeoffs.

2.4 Not So Open AI

Foundation models can be grouped into three categories: open-source, open, and closed. Open-source models are ones where the complete training data, alignment code, weights, and inference code are readily available [12, 13]. Examples of open-source LLMs, which are few and far between, include Olmo, GPT-Neo, and GPT-J. Open models are ones where the weights and inference code are available (e.g., Llama, Qwen, Deepseek). Closed models, such as GPT-4, do not provide training weights. Under the common-task framework, science advanced considerably these past 25 years because of open source. This is especially true for rapid advancements in deep learning over the past 15 years [12, pp. 10-11] where "it increasingly became the norm to publicly release code and datasets...". The implications of this reversal to not-so-open AI are evident in Figure 3a of our health analytics example, where it is difficult to surmise the extent of bias in embeddings using SEAT scores in the pretrained GPT-4 LLM (it is absent from the chart), or the absence of a debiased GPT-4 in Figure 3b. Researchers and practitioners would have to rely on self-reported white papers or alternative benchmarks, as well as downstream inference-based analysis.

2.5 Evaluation Challenges for Generative AI

Whereas evaluation criteria and metrics are well-established for many inference/assessment tasks [14], and for general-purpose text generation tasks (such as question-answering and language modeling capabilities), evaluation of generative AI effectiveness and risks in domain and task-specific contexts remains challenging [15, 16]. In our health analytics example, as part of the piloting phase, let us assume we want to consider the use of generative agents to help simulate how actual users might respond to our AI-guided nudges. Generative agents could be useful for amplifying statistical minority samples in our testbed. For each of the 8,502 users in our testbed, we trained an agentic digital twin – an LLM-based generative agent provided with the demographic, behavioral, and psychological attribute/trait information of the human counterpart. The agentic LLM counterparts were trained using GPT-4o and Llama-3. We then compared the similarity between the generative agent's responses to anxiety and literacy-related prompts relative to those provided by the human counterpart, and the implications of using the human versus digital twin data in downstream prediction models. Figure 3d shows the average similarity of responses between each agent and their human counterpart, including embedding distance and ROUGE scores, relative MSE and F1 performance of anxiety and literacy text classifiers using agent text, relative to human text (top row), and the distribution of individual text response distance scores across the 8,502 human-agent tuples (box plots in bottom row). Looking at the results in the top row, the BERT-CLS embedding similarities are high, and the ROUGE scores are high, suggesting high average semantic similarity between the humans and their agentic digital twins. Similarly, replacing the human text with that of their digital twin does not overly degrade performance for the BERT-based fine-tuned text classifiers (as evidenced by the relative F1 and MSE percentage scores). However, when looking at the individual pair-wise similarities (bottom row), we do see considerable variance in the effectiveness of the agentic digital twins, with a performance long tail near the bottom of the box-plots. This raises many questions. How should we evaluate the effectiveness and risks for this use case? How do we define success? What is acceptable variance at the individual "twin" level? How do we ensure that well-intended generative AI use cases do not lead to unintended consequences?

2.6 Workforce Retooling - Blurring Boundaries Between Human-in-the-loop and AI-in-the-process

When workers are exposed to AI, the outcome could be enhanced augmentation and productivity gains, or worker substitution through automation [2]. AI automation reduces worker demand in an occupation, necessitating new alternative occupations [1]. In contrast, AI augmentation necessitates new skills required to undertake the modified tasks – resulting in human-AI integrated workflows [2]. From a TRAI perspective, the (in)ability to re-skill, whether because of AI automation or to leverage and keep pace with AI augmentation, could be an important and obvious inclusiveness consideration. Less apparent are the potentially profound implications of AI augmentation for AI processes, and consequently, for designing RAPs. As traditionally human-in-the-loop tasks become AI augmented tasks, the boundaries and delineations between AI tasks and human activities become less clear. In the health analytics example (Figure 2), the AI-guided interventions were the focus of our discussion of TRAI research directions related to designing RAPs. However, according to Anthropic's Economic Index report,[2] three of the occupations with the

[2] https://www.anthropic.com/news/anthropic-economic-index-insights-from-claude-sonnet-3-7

highest usage of LLMs (including the extended thinking models) are computer/information research scientists, software developers, and bioinformatics technicians. All three roles were/are central to the design of the AI processes depicted in Figure 2. As AI becomes more ubiquitous and omnipresent – augmenting the design of models, software, systems, and pipelines such as the big data analytics in Figure 2 – what will the design of RAPs look like when the AI-in-the-process cannot be depicted using neat little boxes?

3 Conclusion

The purpose of this article was to shed light on important challenges and opportunities for research related to the design of responsible AI processes/tools. Using a health analytics case study, we presented six research directions for designing responsible AI processes (RAPs). Table 1 summarizes the research directions and some associated, concrete research questions/avenues. Given the important role of RAPs in operationalizing trustworthy and responsible AI, by supporting implementation of the governance functions of mapping, measuring, and managing, these directions are important for creating bridges from policy to practice. Our coverage of the tenets of trustworthy and responsible AI are intentionally meant to be illustrative as opposed to exhaustive. Similarly, the six research directions identified are not intended to holistically capture all challenges and opportunities. Rather, our hope is to motivate rich research streams that usher in a new wave of ideas and thought-leadership on the design of RAPs such that we can move closer towards realizing trustworthy and responsible AI "by design."

4 Acknowledgments

This research was funded by the National Institute of Standards and Technology (ror.org/05xpvk416) under Federal Award ID Number 60NANB24D231 and Carnegie Mellon University (https://ror.org/05x2bcf33) AI Measurement Science and Engineering Center (AIMSEC). Members of Notre Dame's Human-centered Analytics Lab (HAL) were funded in part through U.S. NSF grants IIS-2403438 and IIS-2039915.

References

[1] Kartik Hosanagar and Ramayya Krishnan. Who profits the most from generative AI? *MIT Sloan Management Review*, 65(3):24–29, 2024.

[2] Erik Brynjolfsson. The turing trap: The promise & peril of human-like artificial intelligence. In *Augmented education in the global age*, pages 103–116. Routledge, 2023.

[3] Ricardo Baeza-Yates and Usama M Fayyad. Responsible AI: An urgent mandate. *IEEE Intelligent Systems*, 39(1):12–17, 2024.

[4] Herbert A Simon. *The Sciences of the Artificial*. The MIT Press. The MIT Press, third edition [2019 edition]. edition, 2019.

[5] Katherine L Milkman, Mitesh S Patel, Linnea Gandhi, Heather N Graci, Dena M Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Modupe Akinola, John Beshears, et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20):e2101165118, 2021.

[6] Richard G Netemeyer, David G Dobolyi, Ahmed Abbasi, Gari Clifford, and Herman Taylor. Health literacy, health numeracy, and trust in doctor: Effects on key patient health outcomes. *Journal of Consumer Affairs*, 54(1):3–42, 2020.

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[9] John P Lalor, Ahmed Abbasi, Kezia Oketch, Yi Yang, and Nicole Forsgren. Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, 42(4):1–41, 2024.

[10] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[11] John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. Benchmarking intersectional biases in NLP. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 3598–3609, 2022.

[12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[13] Kezia Oketch, John P Lalor, Yi Yang, and Ahmed Abbasi. Bridging the LLM accessibility divide? performance, fairness, and cost of closed versus open LLMs for automated essay scoring. In *In Proceedings of the 4th GEM Workshop: Generation, Evaluation & Metrics (GEM @ ACL)*, 2025.

[14] Nicolas Prat, Isabelle Comyn-Wattiau, and Jacky Akoka. A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, 32(3):229–267, 2015.

[15] Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. In *Findings of Annual Meeting of the Association for Computational Linguistics (ACL Findings)*, 2025.

[16] Nicolas Prat, John P Lalor, and Ahmed Abbasi. Galea–leveraging generative agents in artifact evaluation. In *International Conference on Design Science Research in Information Systems and Technology*, pages 83–98. Springer, 2025.

Research Direction	Description	Example Research Questions/Avenues
Emergent Problem Space	TRAI as a problem space is highly complex, emergent, and ill-defined, posing challenges for the design and development of solutions.	Can computational researchers and ethicists work together to provide guidelines for when a solution can be considered to be "ethical enough?"
		Can alignment/reinforcement learning with human-feedback (RLHF) be extended to accommodate deeper ethical dilemmas and/or moral reasoning?
		How can designers reconcile or align perspectives on representational versus allocational harm?
Pareto Optimality	The highly multi-faceted nature of the problem space, even within a single tenet of TRAI, challenges the very notion of satisficing solutions.	Should more methods/benchmarks be designed to test interactions between different facets of a TRAI tenet (e.g., fairness)?
		How might different designs improve satisficing across facets?
Beyond Tradeoffs	When adding an additional tenet to the equation, the notion of satisficing breaks down completely.	Can utility-risk frameworks be extended, and perhaps integrated into design of models and RAPs, to allow satisfactory outcomes across TRAI tenets?
Not So Open AI	The trend away from open-source models impedes researchers and practitioners from developing and evaluating TRAI capabilities.	How can we develop digital twins of closed and open-weight foundation models that approximate training data, alignment code, and model weights?
		Can open-source models be scaled up to the performance levels of their closed/open counterparts?
Evaluation Challenges for Generative AI	Evaluation of generative AI effectiveness and risks in domain and task-specific contexts remains challenging.	How can we better design generative AI evaluations that consider real-world settings and interactions, such as sequential evaluation and other longitudinal, dynamic field contexts?
		When should such evaluations consider average effects versus individual or sub-group level heterogeneity?
Workforce Retooling	As traditionally human-in-the-loop tasks become AI augmented tasks, the boundaries and delineations between AI tasks and human activities become less clear.	How can TRAI researchers design RAPs for AI processes where the delineations between humans and AI are less clear?
		What should the interplay between exposure versus substitution and TRAI look like when designing RAPs?

Table 1: Example Research Questions/Avenues Pertaining to the Six Directions