EDITOR: Ahmed Abbasi, aabbasi@nd.edu

DEPARTMENT: TRUSTWORTHY AND RESPONSIBLE AI

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

Ramayya Krishnan, *Carnegie Mellon University, Pittsburgh, PA, 15213, USA*

John P. Lalor, *University of Notre Dame, Notre Dame, IN, 46556, USA*

Nicolas Prat, *ESSEC Business School, 95021, Cergy-Pontoise, France*

Ahmed Abbasi, *University of Notre Dame, Notre Dame, IN, 46556, USA*

Rapid advancements in the development and adoption of artificial intelligence (AI) have accelerated the need for trustworthy and responsible AI (TRAI). National/international AI governance and risk management policies and frameworks have identified a core set of tenets for TRAI, including 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 policymakers. 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.

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 generative AI for text, image, video, audio, and so forth) 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 Figure 1(c), 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,a] 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,

1541-1672 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Digital Object Identifier 10.1109/MIS.2025.3598960
Date of current version 2 October 2025.

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

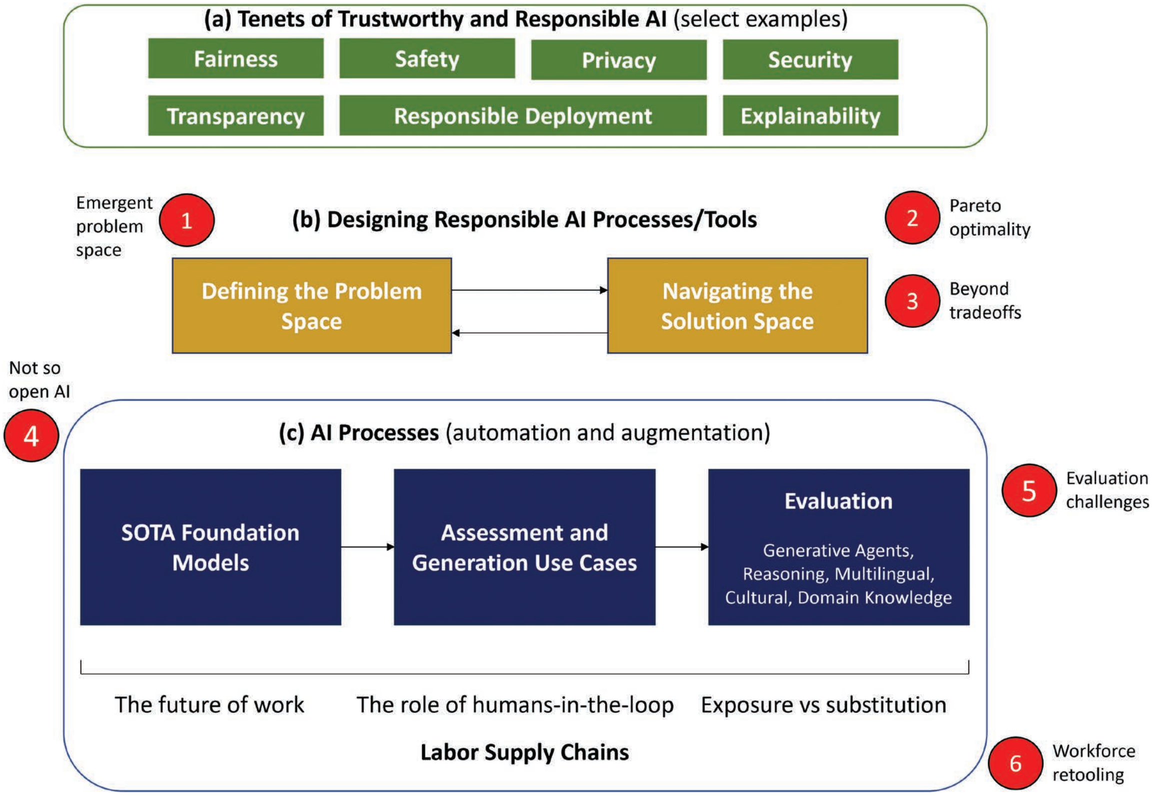

(a) Tenets of Trustworthy and Responsible AI (select examples)
Fairness
Safety
Privacy
Security
Transparency
Responsible Deployment
Explainability
Emergent problem space
1
(b) Designing Responsible AI Processes/Tools
2
Pareto optimality
Defining the Problem Space
Navigating the Solution Space
3
Beyond tradeoffs
Not so open AI
4
(c) AI Processes (automation and augmentation)
5
Evaluation challenges
SOTA Foundation Models
Assessment and Generation Use Cases
Evaluation
Generative Agents, Reasoning, Multilingual, Cultural, Domain Knowledge
The future of work
The role of humans-in-the-loop
Exposure vs substitution
Labor Supply Chains
6
Workforce retooling

FIGURE 1. Six research directions for responsible AI "by design." The figure depicts (a) some example tenets of responsible AI, (c) AI processes, and (b) the need for responsible AI processes, and six challenging research directions (red numbered circles).

and underlying decision-making processes, respectively. Notably, this list of TRAI tenets [depicted in the top part of Figure 1(a)] is illustrative, not exhaustive.

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 the National Institute of Standards and Technology[a]). In recent years, the design of RAPs has emerged as an important avenue for computational and social science researchers. From Simon's[4] classical "sciences of the artificial" perspective, design can be considered a problem-solving paradigm that comprises the proposal of novel solutions to well-defined problems. In this case, the problems of interest are how best to design RAPs to operationalize (i.e., map, measure, manage, and govern) the tenets of TRAI [Figure 1(b)]. The purpose of this article is to highlight six important research directions for the design of RAPs (red numbered circles in Figure 1). Using a real-world case study, we describe the importance of each research direction and illustrate current challenges.

SIX IMPORTANT RESEARCH DIRECTIONS

We use a real-world health-care example to guide our discussion and illustrate some of the nuanced challenges and opportunities that pertain 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 and 2) 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

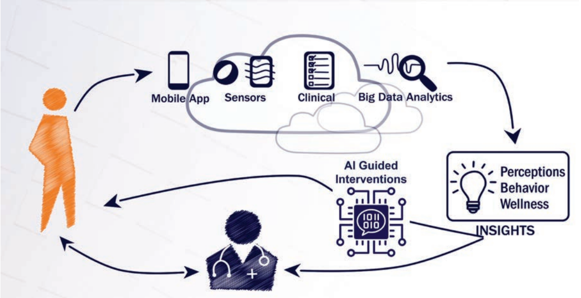


FIGURE 2. Health analytics example to illustrate the six research directions.

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, and so on). We use fairness as our focal TRAI tenet, in part because it has garnered considerable attention in the literature.[7,8] All the results presented for this health analytics example are based on 8502 users.

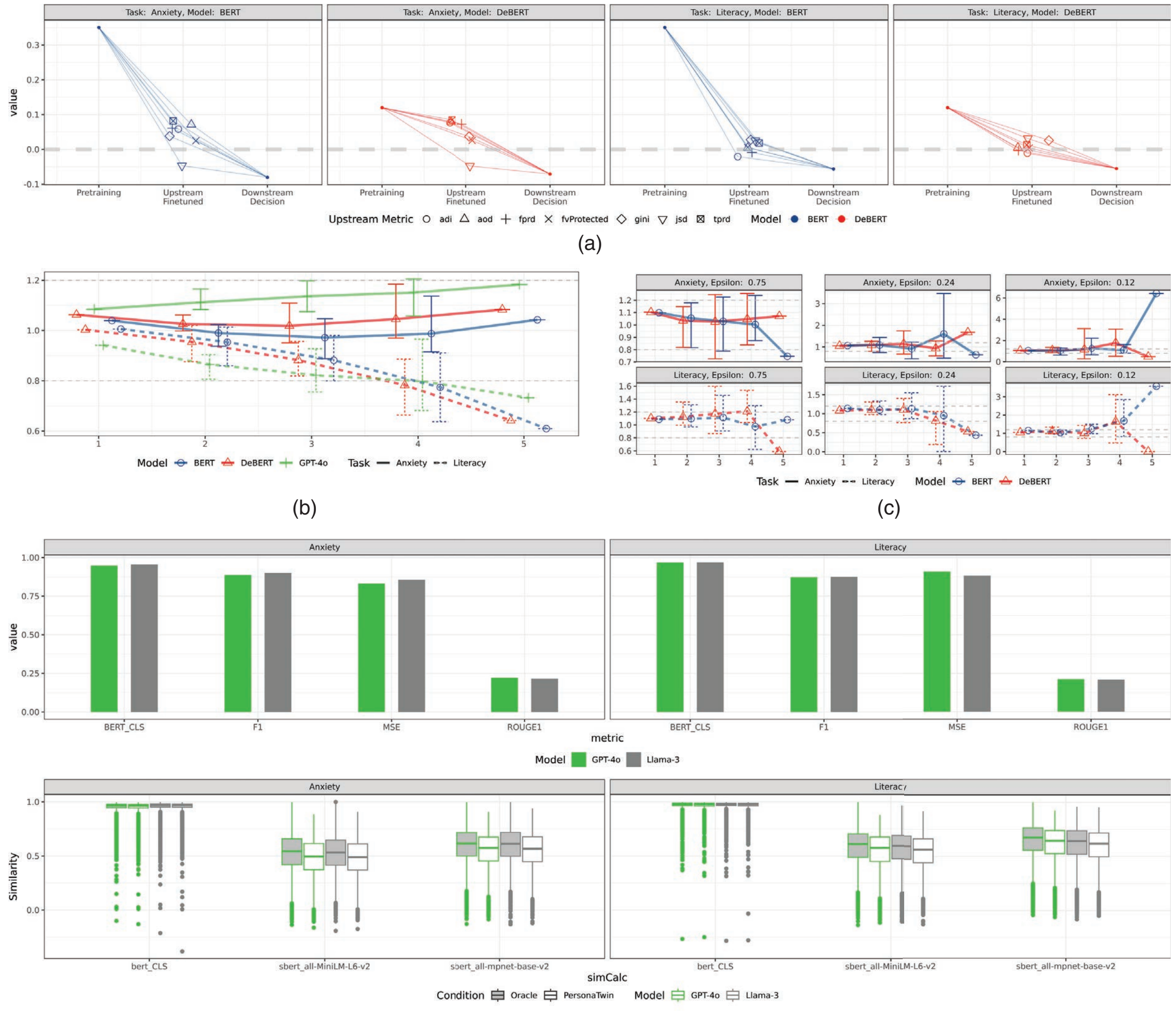
Emergent Problem Space

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 one,[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 regard 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 3(a), which shows gender-related fairness metrics for two anxiety- and literacy-inferring language models [bidirectional encoder representations from transformer (BERT) and debiased BERT (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 sentence embedding association test (SEAT)-6 and SEAT-8 scores], and most of the seven upstream fairness metrics [disparate impact (DI) and so forth 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 suggest that the large language models (LLMs) overly associate anxiety with female patients, however, the downstream allocation harm 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, and mixture-of-expert language models). The multifaceted, nonstationary, and amorphous nature of the problem poses challenges for the design and development of solutions.

Pareto Optimality: Is Satisficing Possible?

If the problem space is 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 multifaceted 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- and three-way interaction effects, often referred to as *intersectional bias*.[10,11] As the interaction combinations increase, so too 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



(a)

(b)

(c)

(d)

FIGURE 3. Results from the health analytics example related to select research directions. (a) Emergent problem space. (b) Pareto optimality. (c) Beyond tradeoffs. (d) Evaluation challenges for generative AI. adi: adjusted disparate impact; aod: average odds difference; fprd: false positive rate difference; fvProtected: fairness violation (protected class); jsd: Jensen-Shannon divergence; tprd: true positive rate difference.

analytics example, this issue is illustrated in Figure 3(b); as we add multiple protected attributes (e.g., demographics such as age, gender, race, education, and income along the *x*-axis), mean DI, and range of DI both increase considerably (*y*-axis). Importantly, this holds for fine-tuned language models (BERT), DeBERT, and in-context learning LLMs [generative pretrained transformer (GPT-4o)].

Beyond Tradeoffs: Fairly Private or Privately Fair?

The prior discussion (and illustrative example) are 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, and so on). 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 3(c). As differential privacy values go up from 100 [no privacy, Figure 3(b)] to

0.75 [very little privacy, the left charts in Figure 3(c)] to 0.12 [fairly private, the right-most charts in Figure 3(c)], the range and mean DI increases three- to fivefold for the nondebiased language model (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.

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, and DeepSeek). Closed models, such as GPT-4, do not provide training weights. Under the common-task framework, science has 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 and 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 3(a) 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 3(b). Researchers and practitioners would have to rely on self-reported white papers or alternative benchmarks as well as downstream inference-based analysis.

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 that 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 8502 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 3(d) shows the average similarity of responses between each agent and their human counterpart, including embedding distance and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores, relative mean square error (MSE) and F1 performance of anxiety and literacy text classifiers using agent text, relative to human text [Figure 3(a)], and the distribution of individual text response distance scores across the 8502 human–agent tuples [box plots in Figure 3(d)]. Looking at the results in Figure 3(a), both the BERT-CLS-embedding similarities and 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 pairwise similarities [bottom row in Figure 3(d)], 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?

Workforce Retooling: Blurring Boundaries Between HITL 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, which are required to undertake the modified tasks, resulting in human–AI integrated workflows.[2] From a TRAI perspective, the (in)ability to reskill, 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

TABLE 1. Example research questions/avenues that pertain to the six research directions.

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 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 multifaceted 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 when 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 subgroup-level heterogeneity?
Workforce retooling	As traditionally HITL 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?

of AI augmentation for AI processes, and consequently, for designing RAPs. As traditionally HITL 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,[b] three of the occupations with the 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 pipeline in Figure 2—what will the design of RAPs look like when AI-in-the-process cannot be depicted using neat little boxes?

CONCLUSION

The purpose of this article was to shed light on important challenges and opportunities for research related to the design of RAPs. Using a health analytics case study, we presented six research directions for designing RAPs. Table 1 summarizes the research directions and some associated concrete research questions/avenues. Given the important role of RAPs in operationalizing TRAI, by supporting implementation of the governance functions of mapping, measuring, and managing, these directions are important for creating bridges from

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

policy to practice. Our coverage of the tenets of TRAI 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 toward realizing TRAI "by design."

ACKNOWLEDGMENTS

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

REFERENCES

1. K. Hosanagar and R. Krishnan, "Who profits the most from generative AI?" *MIT Sloan Manage. Rev.*, vol. 65, no. 3, pp. 24–29, 2024.
2. E. Brynjolfsson, "The turing trap: The promise & peril of human-like artificial intelligence," in *Augmented Education in the Global Age*, D. Araya and P. Marber, Eds., New York, NY, USA: Routledge, 2023, pp. 103–116.
3. R. Baeza-Yates and U. M. Fayyad, "Responsible AI: An urgent mandate," *IEEE Intell. Syst.*, vol. 39, no. 1, pp. 12–17, Jan./Feb. 2024, doi: 10.1109/MIS.2023.3343488.
4. H. A. Simon, *The Sciences of the Artificial*, 3rd ed. Cambridge, MA, USA: The MIT Press, 2019.
5. K. L. Milkman et al., "A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, no. 20, 2021, Art. no. e2101165118, doi: 10.1073/pnas.2101165118.
6. R. G. Netemeyer, D. G. Dobolyi, A. Abbasi, G. Clifford, and H. Taylor, "Health literacy, health numeracy, and trust in doctor: Effects on key patient health outcomes," *J. Consum. Affairs*, vol. 54, no. 1, pp. 3–42, 2020, doi: 10.1111/joca.12267.
7. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021, doi: 10.1145/3457607.
8. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2021, pp. 610–623.
9. J. P. Lalor, A. Abbasi, K. Oketch, Y. Yang, and N. Forsgren, "Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 1–41, 2024, doi: 10.1145/3641276.
10. Y. Chern Tan and L. E. Celis, "Assessing social and intersectional biases in contextualized word representations," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13,230–13,241.
11. J. P. Lalor, Y. Yang, K. Smith, N. Forsgren, and A. Abbasi, "Benchmarking intersectional biases in NLP," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 3598–3609.
12. R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
13. K. Oketch, J. P. Lalor, Y. Yang, and A. Abbasi, "Bridging the LLM accessibility divide? Performance, fairness, and cost of closed versus open LLMs for automated essay scoring," in *Proc. Assoc. Comput. Linguistics Workshop Gener., Eval. Metrics (GEM^2)*, 2025, pp. 655–669.
14. N. Prat, I. Comyn-Wattiau, and J. Akoka, "A taxonomy of evaluation methods for information systems artifacts," *J. Manage. Inf. Syst.*, vol. 32, no. 3, pp. 229–267, 2015, doi: 10.1080/07421222.2015.1099390.
15. Y. Li, Y. Miao, X. Ding, R. Krishnan, and R. Padman, "Firm or fickle? Evaluating large language models consistency in sequential interactions," in *Proc. Findings Assoc. Comput. Linguistics*, 2025, pp. 6679–6700.
16. N. Prat, J. P. Lalor, and A. Abbasi, "GALEA – Leveraging generative agents in artifact evaluation," in *Proc. Int. Conf. Des. Sci. Res. Inf. Syst. Technol.*, Cham, Switzerland: Springer-Verlag, 2025, pp. 83–98.

RAMAYYA KRISHNAN is the W. W. Cooper and Ruth F. Cooper Professor of Management Science and Information Systems, Carnegie Mellon University, Pittsburgh, PA, 15213, USA. Contact him at rk2x@cmu.edu.

JOHN P. LALOR is an assistant professor with the Human-centered Analytics Lab, Department of IT, Analytics, and Operations, University of Notre Dame, Notre Dame, IN, 46556, USA. Contact him at john.lalor@nd.edu.

NICOLAS PRAT is an associate professor with the Department of IS, Data Analytics, and Operations, ESSEC Business School, 95021, Cergy-Pontoise, France. Contact him at prat@essec.edu.

AHMED ABBASI is Joe and Jane Giovanini Professor of IT, Analytics, and Operations, University of Notre Dame, Notre Dame, IN, 46556, USA. Contact him at aabbasi@nd.edu.