People Can Accurately Predict Behavior of Complex Algorithms That Are Available, Compact, and Aligned

ANONYMOUS AUTHOR(S)

ACM Reference Format:
Anonymous Author(s). 2025. People Can Accurately Predict Behavior of Complex Algorithms That Are Available, Compact, and Aligned. In *STAIG*. ACM, New York, NY, USA, 10 pages. https://doi.org/X

1 Introduction

Users' ability to understand and predict the behavior of algorithms is integral to user agency in interactions with algorithmic systems [49]. When humans use algorithms as a tool in pursuit of a goal, being able to predict the tool's behavior is central in deciding whether and how to use them [3, 5, 45, 49]. When algorithms are experienced largely by those they are applied *upon*—in contexts like hiring [37], bail and parole [14], rent-setting [47], information seeking [4], and displaying political content online [34], users and the broader public have a vested interest in being able to predict how algorithms will behave, to know how they will be impacted and shape their interactions and advocacy.

In order to model the behavior of designed systems such as the algorithmic ones we consider, users construct mental models: cognitive representations of system behavior [9]. Since users typically struggle to build predictive mental models of complex algorithms [18, 19], one obvious response is to just use simpler algorithms. An extremely simple algorithm, such as ranking a social media feed reverse-chronologically, is trivially easy to understand. However, such a decision means entirely ruling out higher-complexity algorithms, which can achieve aims the simple ones are simply incapable of. However, we argue that people can predict the behavior of even highly complex artificial intelligence algorithms when as those algorithms align with concepts they can understand and replicate. For example, you can likely accurately predict the behavior of a large language model (LLM) that classifies whether a social media post is about politics, despite the LLM relying on a complex attention network and hundreds of billions of parameters.

We propose that people can create an accurate predictive mental model of an algorithm if and only if the algorithm is *Available*, *Compact*, and *Aligned*. These ACA criteria together capture what it means for a person to be able to map algorithm behavior into an *existing* cognitive schema: (1) Availability, a reference to availability bias [43], captures the recognizability of the underlying concept that the algorithm is modeling. (2) Compactness, drawing on the literature of cognitive chunking [10, 41], refers whether the algorithm's behavior can be synthesized into a single cohesive concept: is the algorithm representing a single concept or fitting together multiple concepts into a greater whole that people understand? (3) Finally, alignment tests whether the algorithm's execution of its concept agrees with the person's execution of that concept, similar to representational alignment [40]. To present a first test our prediction that ACA is necessary and sufficient, we report two experimental studies ($N = 1200$ and $N = 600$) where we vary the algorithm and test whether people can predict the algorithm's behavior. We situate these studies in the domain of social media feed algorithms, first exposing participants to a randomly selected algorithm and then asking them to repeatedly predict which of two previously unseen posts the algorithm would rank higher. We measure whether participants correctly predict the algorithm's decision and find that, as predicted by our proposed theory, participants rank algorithms that satisfy all three criteria most accurately.

User understanding of algorithm behavior is of high import for the broader question of designing and implementing algorithm governance. Since algorithm behavior itself is the subject of governance, algorithm governors necessarily must reason over their behavior. Thus, the stakeholders involved need to understand what constitutes the algorithm behavior to make informed decisions. To communicate effectively as part of a governance process involving multiple stakeholders, people must be able to create abstractions with which to reason over and describe algorithm behavior: existing or desired. Theory that describes when users can understand algorithm behavior therefore may help determine algorithm desirability and inform the process of governance itself by warning of potential communication breakdowns.

2 Related Work

As complex algorithms play a larger role in day-to-day life, algorithm understandability has been raised as a major concern [3, 31, 49]. Without sufficiently understanding algorithms, harms emerge from errors, overreliance, and misinterpretation [23, 39, 50]. If we take the perspective that complex algorithms are fundamentally opaque and difficult to understand or predict, then the only way to reduce these harms is to replace complicated algorithms with simpler, more transparent ones, and accept whatever performance losses correspond to the change. In this paper, we explore an alternative perspective that the behavior of some complex algorithms *is* easy to predict for humans, because there is a difference between what is computationally complex for an algorithm to execute and what is cognitively complex for a human to understand. We aim to better explain the difference between technical complexity of an algorithm and the cognitive complexity of its behavior, and develop a framework for predicting the latter. Our theory builds on the extensive theory developed within the domains of explainable and interpretable artificial intelligence: we are both concerned with how well people can reliably understand and interpret the behavior of algorithms.

Explainable and interpretable AI research is focused on creating and evaluating methods that operate over AI models: by offering post-hoc explanations for decisions made by algorithms (e.g., LIME [38]) and evaluating their efficacy [20]; manipulating the models directly such that they operate over human understandable concepts (e.g., Concept Bottleneck Models [26]) and evaluating the efficacy of these inherently more interpretable models [27]; and giving global overviews of models' abilities, as in the case of a number of HCI frameworks or systems [7, 8, 24, 33, 35], evaluating the efficacy of these systems [7, 8, 24]. Because work in explainable and interpretable AI and our work are concerned with helping people understand AI systems, we draw heavily on this work to create our framework, especially research concerning how to produce better human understanding of AI systems.

Explainable AI (XAI) theory on what qualifies as an effective explanation helps inform our own theory about people's understanding and ability to form mental models. Our criteria are informed by several developments in this area. We are not the only researchers to note that simply switching to a simpler model architecture is neither necessary nor sufficient to produce user understanding. Research in XAI has noted that linear models can be more opaque than even deep learning models due to the more convoluted features used in them [31]. Additionally, we take inspiration from the explainability work that demonstrates how people's understanding is impacted by factors beyond the specifics of the technical system, including factors like the social context and users' intuition formed from experience [11, 12, 16].

Another line of XAI work has demonstrated that cognitive effort plays a major role in whether people verify and override AI's erroneous decisions, as well as in whether people will attempt to understand explanations [2, 5, 6, 45].

Particularly, even explanations that are simple compared to the original model may not be simple enough for the human to accept. A theme in XAI research therefore involves lowering the amount of cognitive effort required for people to understand explanations or use them to verify the model output [1, 38, 44]. Much of our theory focuses on what characteristics of algorithms allow users to form, hold, and use mental models with minimal cognitive effort. We are influenced especially by model architectures that use concepts as a building component to render themselves more interpretable to users [26, 28]. Our theory uses concepts as the building block of user mental models.

Our method is also informed by the way that XAI as a field has demonstrated that user understanding of explanations is task, context, and user-dependent, and therefore demands a human-centered approach [17, 30]. The field of interpretable AI has similarly contended that interpretability efforts should be directed at and evaluated with respect to specific end-goals and end-users [5, 13, 36, 42, 46, 48]. We therefore focus our theory development and evaluation toward a specific context and task. We are motivated by the XAI insight that *contrastive* reasoning is used for answering questions of *why* something happened, the most complex type of reasoning about and understanding behavior [32]. Our domain of social media is ripe with this type of reasoning, as social media algorithms are used to decide what users see and what they do not.

3 ACA: Availability, Compactness, and Alignment

We propose that, in order for users to be able to form and deploy a mental model of algorithm behavior, the algorithm's behavior must satisfy Availability, Compactness, and Alignment. If the algorithm fulfills these criteria, then users can enlist existing mental representations to accurately predict its behavior.

Availability: In our context, availability refers to the cognitive availability of the algorithm's objective. An available concept is one that the person is primed to expect or would readily leap to, as a user of a social media platform expects their algorithm to optimize for engagement signals. Likewise, a less available concept, for example the (very real) social media algorithm objective of showing you content that is predicted to produce replies for other users who otherwise have no feedback on their posts [15], is far less likely to lead users to produce a predictive mental model.

Compactness: For an algorithm to be *compact*, it must consist of few concepts (ideally, a single one), or multiple that can be unified into fewer. Compactness anchors on the cognitive psychology concept of *chunking*, the "recoding of smaller units of information into larger, familiar units" [41]. A chunk unifies features together so long as they have stronger associations with each other than with other potential features [21]. An algorithm meets compactness criteria if it can be effectively chunked, or represented, into cognitive concepts. Simple algorithms are often compact by default: e.g., the decision criterion for a credit card might be a specific minimum credit score. On the other hand, when an algorithm adds other concepts that cannot be represented (chunked) together effectively, for example combining engagement with political balance, then the algorithm becomes less compact.

Alignment: For an algorithm to be *aligned*, the user's understanding of the concept must agree with the algorithm's execution of the concept. Simple algorithms are often easily aligned: if we sort a social media feed reverse chronologically, the algorithm's implementation and the person's implementation are very likely to match. For more complex algorithms, alignment may simply mean a high-performing algorithm that achieves high accuracy. However, algorithms with subjective concepts—for instance, perceptions of humor or toxicity, which may differ on a person-by-person basis—may also be misaligned for a given viewer even when aligned optimally for the average viewer [22].

We argue that when all three criteria are met, then a person can predict an algorithm's behavior, no matter the algorithm's underlying complexity. Simple algorithms are often available, compact, and aligned, like chronological feed algorithms. We also claim that even extremely complex algorithms can be ACA compliant. For example, the vision models that accurately distinguish photos of cats and dogs are extremely computationally complex, while also exhibiting behavior that is very intuitive and therefore predictable for everyday people familiar with the animals.

4 Experimental Study

We measure people's ability to predict behavior for a variety of algorithms to test our theory. We hypothesize that participants will predict algorithms that fulfill all ACA criteria with higher accuracy than those failing at least one.

Task and Study Procedure. We recruit 1200 participants across Prolific and Cloud Research Connect. We ask them to observe a randomly assigned social media feed algorithm for two minutes, using posts collected from X. After viewing the feed, we ask them to arrange two posts according to how they think the algorithm would rank them: ten times for training with feedback on their correctness and then thirty times without feedback. We collect their accuracy during this second phase to calculate their aggregate accuracy, where random is 50%.

Algorithmic Conditions. We construct algorithms with varying levels of ACA compliance. Five of the algorithms are **ACA**, meeting all three criteria. Each other possible selection of criteria met versus not (e.g. available but not compact or aligned...). Descriptions of each and justifications for each assignment are in Appendix A.1.

Results. We visualize our results in Figure 1 and the include the output of our mixed effect logistic regression model: `accurate ~ available * compact * aligned + rank_difference + (1 | participant)` [1] in Table 1.

Our data partially support our hypothesis: participants performed stronger on prediction for algorithms fulfilling all three conditions, with a statistically significant positive coefficient for the three-way interaction effect of availability, compactness, and alignment in our mixed effect logistic regression. The coefficient (log odds) is 1.0, which signifies an odds ratio of 2.8 for correct predictions when satisfying all three criteria versus none. However, we also note that the coefficients for the post rank difference and availability-alignment interaction are statistically significant.

	Mixed-effect Logistic Regression	
	DV: Correctness	
Constant	-0.518^{*}	(0.237)
availability	-0.264	(0.152)
compactness	-0.036	(0.152)
alignment	0.057	(0.144)
post rank difference	**0.655^{**}**	**(0.245)**
availability:compactness	0.296	(0.215)
availability:alignment	**0.806^{***}**	**(0.211)**
compactness:alignment	0.120	(0.210)
availability:compactness:alignment	**1.034^{***}**	**(0.285)**

Table 1. In the mixed-effect logistic regression, we see significant positive coefficients for the ACA interaction, as well as the AA interaction. All other interactions between ACA criteria do not show significance. Note: *p<0.05; **p<0.01; ***p<0.001

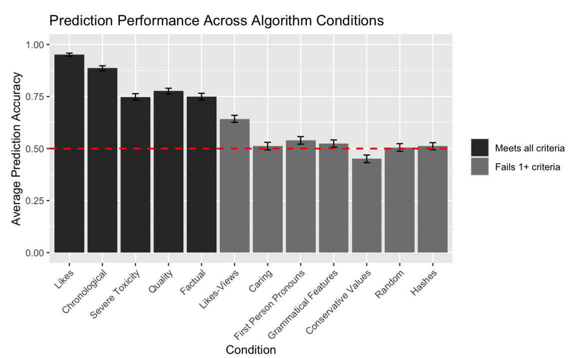

Fig. 1. Participants predicted the behavior for algorithms that satisfied all ACA criteria with the highest accuracy. They predicted most algorithms that failed 1+ criteria at close to baseline rates.

[1]This model specification is designed to test our hypothesis that all three criteria are needed, thus the three-way interaction (including two-way interaction and individual terms). Rank difference represents the absolute value of the difference in sorted rank for the given algorithm condition across all possible posts in our sample, normalized by the number of posts—in other words, whether two posts are near or far from each other according to this ranking. If a participant understands the underlying algorithm behavior, they should have an easier time classifying which post is higher ranked when the ranking would be very different, and thus "rank difference" would be large. The three criteria–available, compact, and aligned–are coded as binary variables.

While the post rank difference coefficient is consistent with out theory,[2] the positive and significant coefficient for the availability-alignment interaction is a concern. We attribute it to the higher-than-expected score for the likes-views ratio algorithm, which had ~65% performance. Notably, all other non-ACA algorithms had accuracies below 55%, where 50% would be the random guessing baseline, leaving likes-views ratio as a significant outlier. To explain this anomalous result, we analyzed the mental models that participants described holding for their algorithm conditions and noted that participants in the likes-views ratio condition commonly mentioned simpler algorithms that were more compact (e.g., information quality, political leaning). We hypothesized that their high prediction accuracy was enabled by latent correlations in the posts used and ran simulations to estimate prediction accuracy for participants who based their predictions on what the behavior of the most highly correlated of the alternate mental models would be. We found that average prediction accuracy above 65% was feasible through this method, which could explain our study result.

To further test this hypothesis, we performed a follow-up study using the same methodology as the first, with two new conditions: using likes-views ratio but choosing a set of posts without large correlations with our other algorithms, and a more complicated combination of engagement features (likes-retweets-views ratio) that did not have the same latent correlations. Both of these new conditions were intended to show that a non-compact algorithm will lessen user prediction accuracy when without strong correlations with ACA algorithms. In this second study ($N = 600$), the de-correlated version of likes-views ratio corresponded to a prediction accuracy of 57% and the likes-retweets-views ratio to 54%, rather than 63% for regular likes-views ratio (replicated to within 1% of our original result). This result, with the non-compact algorithms at near-random-guessing prediction performance, helps to account for the unexpectedly high prediction accuracy we earlier saw in an AcA condition. We saw evidence that failing compactness will hurt prediction accuracy when these helpful correlations are not present. Together, these two studies demonstrate that available, compact, and aligned algorithms can be predicted by users with higher accuracy than others.

5 Discussion

We propose that available, compact, and aligned algorithms are easier for people to predict. Through experiments, we demonstrate that users can form accurate predictive mental models for algorithms if and only if they fulfill these criteria. We demonstrate the promise of a new class of algorithms which are understandable to users while still complex.

The ACA criteria should not only inform algorithm design and evaluation in the general context, but have profound implications around algorithm governance. If algorithm stakeholders struggle to conceptualize how the algorithm behaves, then this prevents them from making informed decisions in the context of governance. We therefore believe that the ACA criteria offer insights into desirable characteristics of algorithm, as well as into the deliberation process for algorithm governance. If user understanding of algorithms is desirable, then the ACA characteristics provide a foundation from which to dictate achieving that. In terms of the governance process, our work can help predict possible communication breakdowns in the governance process when discussing desired algorithm behavior. Our work suggests the need for carefully considered abstractions while deliberating and discussing algorithm behavior so that they are mutually intelligible and enable accurate communication.

Continuations of this work will verify the theory with a more diverse array of algorithms, and then investigate possible interventions into existing algorithms according to the ACA criteria. Future work should also replicate these experiments outside the feed algorithm context to test generalizability.

[2] The log odds for post rank difference is 0.66, meaning that there is an associated 1.9 times greater odds of prediction correctness on a question when the posts are ranked maximally differently versus exactly the same. Such a relationship is consistent with participants who are basing their predictions on an understanding of the algorithm: a larger difference in the algorithm rating should make the distinction of which should be ranked higher more apparent.

References

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. arXiv:2006.14779 [cs.AI] https://arxiv.org/abs/2006.14779

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[4] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1-7 (1998), 107–117.

[5] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169. https://doi.org/10.1109/ICHI.2015.26

[7] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. 2023. Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 419, 14 pages. https://doi.org/10.1145/3544548.3581268

[8] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 136 (apr 2023), 21 pages. https://doi.org/10.1145/3579612

[9] John M Carroll and Judith Reitman Olson. 1988. Mental models in human-computer interaction. *Handbook of human-computer interaction* (1988), 45–65.

[10] William G Chase and Herbert A Simon. 1973. Perception in chess. *Cognitive psychology* 4, 1 (1973), 55–81.

[11] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. arXiv:2202.04092 [cs.AI] https://arxiv.org/abs/2202.04092

[12] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (oct 2023), 32 pages. https://doi.org/10.1145/3610219

[13] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML] https://arxiv.org/abs/1702.08608

[14] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[15] Dean Eckles, René F Kizilcec, and Eytan Bakshy. 2016. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7316–7322.

[16] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. https://doi.org/10.1145/3411764.3445188

[17] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 94, 6 pages. https://doi.org/10.1145/3411763.3441342

[18] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like"' it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[19] Megan French and Jeff Hancock. 2017. What's the folk theory? Reasoning about cyber-social systems. *Reasoning About Cyber-Social Systems (February 2, 2017)* (2017).

[20] Damien Garreau and Ulrike Luxburg. 2020. Explaining the explainer: A first theoretical analysis of LIME. In *International conference on artificial intelligence and statistics*. PMLR, 1287–1296.

[21] Fernand Gobet, Peter CR Lane, Steve Croker, Peter CH Cheng, Gary Jones, Iain Oliver, and Julian M Pine. 2001. Chunking mechanisms in human learning. *Trends in cognitive sciences* 5, 6 (2001), 236–243.

[22] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[23] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–44.

[24] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[25] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4459–4471. https://doi.org/10.18653/v1/2022.acl-long.306

[26] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. arXiv:2007.04612 [cs.LG] https://arxiv.org/abs/2007.04612

[27] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.

[28] Michelle S. Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A. Landay, and Michael S. Bernstein. 2023. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 741, 24 pages. https://doi.org/10.1145/3544548.3581290

[29] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. arXiv:2202.11176 [cs.CL] https://arxiv.org/abs/2202.11176

[30] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv:2110.10790 [cs.AI] https://arxiv.org/abs/2110.10790

[31] Zachary C. Lipton. 2017. The Mythos of Model Interpretability. arXiv:1606.03490 [cs.LG] https://arxiv.org/abs/1606.03490

[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM. https://doi.org/10.1145/3287560.3287596

[34] Arvind Narayanan. 2023. Understanding Social Media Recommendation Algorithms. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

[35] Gregory Plumb, Nari Johnson, Angel Cabrera, and Ameet Talwalkar. 2023. Towards a More Rigorous Science of Blindspot Discovery in Image Classification Models. *Transactions on Machine Learning Research* (2023).

[36] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. arXiv:1802.07810 [cs.AI] https://arxiv.org/abs/1802.07810

[37] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG] https://arxiv.org/abs/1602.04938

[39] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2016. Automation, algorithms, and politics| when the algorithm itself is a racist: Diagnosing ethical harm in the basic components of software. *International Journal of Communication* 10 (2016), 19.

[40] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2024. Getting aligned on representational alignment. arXiv:2310.13018 [q-bio.NC] https://arxiv.org/abs/2310.13018

[41] Mirko Thalmann, Alessandra S Souza, and Klaus Oberauer. 2019. How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45, 1 (2019), 37.

[42] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

[43] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology* 5, 2 (1973), 207–232.

[44] Berk Ustun and Cynthia Rudin. 2015. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (Nov. 2015), 349–391. https://doi.org/10.1007/s10994-015-5528-6

[45] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[46] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Perspectives on Dependable AI*. The MIT Press. https://doi.org/10.7551/mitpress/12186.003.0014 arXiv:https://direct.mit.edu/book/chapter-pdf/2249916/c006600_9780262366212.pdf

[47] Heather Vogell, Haru Coryne, and Ryan Little. 2022. Rent going up? One company's algorithm could be why. *Pro Publica* (2022).

[48] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[49] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (may 2019), 70–79. https://doi.org/10.1145/3282486

[50] Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41, 1 (2016), 118–132.

A Appendix

A.1 Conditions

Below, we shorthand the criteria met by the algorithm: a capital letter to signify that they meet that criterion, and a lowercase letter to signify that they do not meet that criterion. For example, an "ACA" algorithm is available, compact and aligned; whereas an "acA" algorithm is not available nor compact but it is aligned.

A.1.1 Likes: ACA. Sorting by likes involves only features that are numeric and visible on each post, making it available. It uses only a single feature, making it compact. And since likes are reported exactly, there is no ambiguity in calculating the number, making it aligned.

A.1.2 Reverse Chronological: ACA. Sorting by recency involves only features that are reported on each post, making this algorithm available. It only uses one feature, making it compact. And since the time since posting is reported on the post, there is little[3] ambiguity in calculating the number, making it aligned.

A.1.3 Severe Toxicity: ACA. Sorting by extreme toxicity involves only the textual content, which humans are skilled at processing and synthesizing in contrast to many numeric factors. Extreme toxicity stands out since it is frequently shocking or unusual, making it highly available. Toxic or antisocial behavior is a known concept that does not need to be subdivided, making it compact. And since the toxicity classifier works quite well on political posts, the algorithm is well-aligned. Ratings of severe toxicity come from Perspective API [29].

A.1.4 Factual: ACA. Sorting by factual-presenting content involves only the post text, which humans are skilled at processing and synthesizing in contrast to many numeric factors. Factual posts and opinion-based posts look very different in the political domain, making this algorithm available. People are taught to distinguish fact from opinion, so they do not have to memorize all of the individual characteristics of how facts versus opinions are expressed, making this distinction a compact concept. This algorithm is well-aligned due to its high accuracy and the strong distinction between posts that the participants are asked to compare[4]. Factual ratings are obtained by prompting GPT 4o[5] and are given from 0 to 1 with 0.1 increments.

A.1.5 Quality: ACA. Sorting by writing quality involves only the post text, which humans are skilled at processing and synthesizing in contrast to many numeric factors. Features like proper grammar and lengthy descriptions clearly distinguish quality writing, which makes this algorithm available. People are taught how to write well, making writing quality a compact concept. This algorithm is well-aligned due to its high accuracy and the strong distinction between

[3]If two posts were from very similar times, they could be within the rounding error of the time reported. However, since our task involves comparing very differently ranked posts, this should not be an issue in principle.

[4]Our "factual" algorithm codes on whether posts are presented as factual content (not the actual truth of the content). Thus, issues with detecting misinformation are not relevant to alignment for this algorithm.

[5]All prompts are reported in the appendix.

posts that the participants are asked to compare. Quality ratings are obtained by prompting GPT 4o and are given from 0 to 1 with 0.1 increments.

A.1.6 Likes-Views ratio: AcA. Both likes and views are presented clearly on the post, making the features of this algorithm available. However, the combination of likes divided by views makes it not compact, since multiple features are being combined in an unclear way. Note that while division is not a particularly complicated way to combine factors, combining factors (especially non-additively) makes the algorithm very difficult to detect. Likes divided by views does not collapse to a single recognizable concept that can be processed as a unit. Since the algorithm is exactly likes divided by views, there is no ambiguity, so it remains highly aligned.

A.1.7 First-person pronouns: aCA. The count of first-person pronouns is not easily recognizable from looking at or even reading a post. This algorithm is therefore not available. The idea of first-person pronouns is a taught concept, making it compact. And the algorithm is calculated only by counting the number of these words, preventing any ambiguity since posts with different numbers are being compared. First-person pronouns is therefore aligned.

A.1.8 Caring: ACa. Sorting by how caring a post is relies only on the post text, which humans are good at processing. A very caring post sounds very different from an uncaring one in the political domain, making it available. Caring as a concept is well known to people, making it compact. However, this algorithm does not perform accurately enough among the political tweets we use, making it unaligned. This algorithm is implemented using a BERT-based architecture that can report the presence of different values [25].

A.1.9 Grammatical features: acA. This algorithm uses different grammatical features (10 times the ratio of second person to first person pronouns, plus the average word length, plus the number of punctuation marks) which are all not available upon looking at the post. By combining multiple different features without a clear reason for the association, it is not compact. However, when applying this algorithm, there is no ambiguity or errors, making it aligned.

A.1.10 Conservative values: Aca. This algorithm, which combines ratings for several concepts (tradition, achievement, personal security, and conformity to rules), uses values that are familiar to people and salient in the political context, making them available. However, by combining so many separate concepts, this algorithm becomes non-compact. Due to the lower than necessary performance of each individual classifier for the different values, this algorithm is also unaligned. The value concept rating is performed by the same BERT-based model as in the Caring condition.

A.1.11 Random: aCa. Since randomness does not involve any particular features, there is nothing to be rendered available. The concept of randomness is well known, making this algorithm technically compact. However, knowing that an algorithm is random does not help a user to make ranking predictions, making this algorithm not aligned. The user cannot make effective ranking predictions without in-depth knowledge of what the random ranking value is, which they do not have.

A.1.12 Hashing: aca. This algorithm applies a hashing function to the post text, creating a not-quite random algorithm. The hash value is in no way available, but also not compact, since the user cannot see the value used by the ranking, nor uncover how it was reached. The user cannot make effective ranking predictions without in-depth knowledge of how the function works, which they do not have, making this algorithm not aligned[6].

[6]Note: hashing has a very similar effect to randomization, but involves a subtly more involved process due to the complexity of how it is computed. Average people have no pre-existing schema for what hashing is, the way they do for randomness, which is the only distinguishing factor between these algorithms. And even though hashing has the appearance of randomness, it is not random—the same posts will always result in the same ranking—so a

conceptualization around randomness would not suffice. However, the difference in having an existing schema of what the algorithm behavior is does not make much of a qualitative difference to the user, due to the lack of availability and alignment of said schema for randomness.