Decision-Maker Alignment: Benchmark Datasets

Anik Sen
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
as5867@drexel.edu

Rosina O Weber
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
rosina@drexel.edu

Christopher B. Rauch
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
cr625@drexel.edu

Mallika Mainali
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
mm5579@drexel.edu

JT Turner
Knexus Research
National Harbor, MD, USA
jt.turner@knexusresearch.com

John Meyer
Knexus Research
National Harbor, MD, USA
john.meyer@knexusresearch.com

Michael W. Floyd
Knexus Research
National Harbor, MD, USA
michael.floyd@knexusresearch.com

Matthew Molineaux
Parallax Advanced Research
Beavercreek, OH, USA
matthew.molineaux@parallaxresearch.org

2025 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3315-2400-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/CAI64502.2025.00212

***Abstract*—While alignment in artificial intelligence (AI) is broadly concerned with how AI systems align with human values, *decision-maker alignment* refers to how algorithms align with the values of individual decision makers. Consequently, the values targeted in *decision-maker alignment* are attributes that influence the decision-making process employed by humans. Example of a cognitive attribute is risk tolerance. The ideal environments for investigating *decision-maker alignment* are those in which the optimal decision may not be available, forcing humans to compromise and select a suboptimal decision. An optimal decision may not be available due to environment constraints such as limited resources, uncertainty, or some source of pressure. In contrast with supervised machine learning, *decision-maker alignment* can be investigated with labeled data in which decisions are made by different decision makers who are influenced by cognitive attributes. The problem is that datasets to study *decision-maker alignment* are difficult and expensive to create and they result in a small number of samples. For this reason, this paper proposes an approach to extend existing datasets for the purpose of studying decision-maker alignment. We exemplify our proposed approach by extending a dataset of health insurance alternatives.**

***Index Terms*—decision making, ITM, decision-maker alignment, cognitive attributes, AI alignment, synthetic data**

I. Introduction

The goal of artificial intelligence (AI) alignment is to ensure that the behavior of the models is consistent with human-intended goals and preferences [1], [2]. Although there has been recent research on AI alignment, much of it has mainly focused on specific societal issues, such as racial and gender bias [3], [4]. This paper is concerned with decision-maker alignment (DMA), the problem of aligning algorithmic decisions with individual decision makers; it is a subset of the research in AI alignment. DMA is concerned with environments where the optimal decision may not be available. Because it is unavailable, human decision makers must select a suboptimal decision and compromise on certain aspects of the decision. In rational decision making [5] (*i.e.*, where an optimal decision is available), it does not matter who makes the decision because any decision maker, human or algorithmic, is expected to be rational and thus select the optimal decision. An optimal decision that would be considered optimal for all may not be accessible for different reasons [6]. It may be due to limited resources; for example, it may be clear what the ethical action is but there are not enough resources to be ethical. Sometimes decision makers are under time constraints, and it becomes impossible to take all the elements of a problem into consideration, so a suboptimal decision is made [7]. Pressure may also be of political nature. Information uncertainty is another condition that renders optimal decisions impossible to identify, just due to the lack of information [8]. In these situations, it matters who makes the decision because humans rely on their inherent cognitive attributes [9] to find a compromise with which they can be comfortable. Studies in cognitive science suggest that cognitive attributes such as risk tolerance are stable attributes that influence human decision making [10]–[12]. Consequently, data to study DMA needs to entail such cognitive attributes.

Benchmark datasets to investigate algorithms for and properties of DMA are expensive to build because they typically require interviewing users to capture decisions influenced by cognitive attributes. Datasets designed for other learning and alignment problems are not suitable. There are datasets

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA), contract number FA8650-23-C-7317. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

979-8-3315-2400-5/25/$31.00 ©2025 IEEE
DOI 10.1109/CAI64502.2025.00212

available for general AI alignment (*e.g.*, [13]–[15]), but those do not distinguish decision makers or characterize them with influential attributes. In this paper, we address this gap by exploring how existing data can be modified and extended in support of investigations of *DMA*. In the next section, we describe DMA and introduce a notation for it. In Section III, we propose an approach to generate datasets, and in Section IV, we discuss how to compute alignment and demonstrate the performance of our algorithm that learns to align with decision makers. We conclude in Section V.

II. Decision-Maker Alignment

DMA refers to the building of algorithms that align with the values of individual decision makers. As introduced earlier, the environment where this type of alignment is studied forces decision makers to select a suboptimal decision because the optimal decision, although it could be somehow envisioned, is not accessible or available for some reason. This forces decision makers to compromise and select a suboptimal decision; this is when they might disagree because they rely on individual principles that characterize their decisions, which may differ. Cognitive science studies a class of cognitive attributes [16] that are believed to be responsible for variations in decisions when the optimal decision is not available [8], [17], [18].

The DARPA ITM program [19] selected combat triage as an environment where soldiers often face the characteristics of uncertainty, time pressure, and limited resources, where optimal decisions are not available. These decisions are often referred to as *impossible decisions* [20].

A typical environment for learning DMA can be represented through a dataset consisting of four components, namely, *probes* $p \in P$, their respective *decisions* $y \in Y$, *contextual categories* CC, and attributes $u \in U$ for *decision-maker characterizations*. A *probe* is a decision prompt $p \in P$ that requires a decision or action $y \in Y$ that the decision maker must make or select to respond to the probe. Hence, the probe entails the dilemma faced by the decision maker. For example, "How to tag a casualty?" or "Which level of deductibles could I afford?" The alternative *decisions* for a probe represent the output of the alignment task. The *contextual categories* are described through a set of real-valued features $f \in F$. These features are equivalent to the features in ML training instances and the contextual categories are analogous to conceptual classes. The context features that make the contextual categories help decision makers assess the current state as they provide context for them to envision potential outcomes and select a decision. For example, in the combat triage domain, features include the type of injury suffered by a casualty, as well as their vital signs. For the health insurance domain, contextual features include whether family members have chronic conditions and the age of their children.

The fourth component is the *decision-maker characterization*. The working hypothesis is that decision makers can be characterized by cognitive attributes $u \in U$ [21]–[23]. A decision maker can be characterized with one or multiple attributes, and the characterization may represent an aggregate of decision makers. When a probe is given as training data to learn DMA, an alignment score Ψ must be computed. The notion is that an algorithmic decision maker that makes the exact same decision as the ground-truth decision would receive the maximum score of $\Psi = 1$ for that probe. The problem posed by DMA may involve training data about probes that differ from those in the testing data. For an example of an alignable algorithmic decision maker, see [24], [25].

The environments for DMA differ from those in ML in that every instance represents a decision made by a human. Furthermore, there must be a characterization of human decision makers along the attributes that influence their decisions. In supervised ML, the space of instances $x \in X$ and labels $y \in Y$ are used to learn a function $\Omega(X) = Y$ whereas the space of instances for DMA consists of instances $x \in X$, labels $y \in Y$, and attributes $u \in U$.

III. Generating Datasets to Evaluate the Decision-Maker Alignment Problem

Datasets to investigate DMA can be built with human studies but they can be expensive and result in a dataset of limited size (*e.g.*, [26]). Another direction to create benchmark datasets would be using benchmark models of human behavior [27] but there are not enough data to build such models. Therefore, we propose an approach where we start from the association between *probes*, their alternative *decisions*, and applicable *cognitive attributes*, and only then generate and populate values for context features that represent *contextual categories*. The advantage of this approach is that we start by identifying *probes* that reflect the required dilemma for decision makers and directly associate each alternative *decision* of each *probe* with a *decision maker characterization* such as a risk tolerance level of *low* or *high*. In this section, we provide details of the approach and illustrate it by implementing it from health insurance plan specifications.

The proposed approach to generate DMA datasets considers the main components in DMA introduced in Section II: *probes* $p \in P$, *decisions* $y \in Y$, *contextual categories* C, and *attributes* $u \in U$ for *decision-maker characterizations*. The approach consists of five steps, listed next. We discuss computing alignment in the next section.

A. Identify a source of plausible *probes* $p \in P$ and respective decisions $y \in Y$ that replicate the conditions of DMA.
B. Identify *attributes* $u \in U$ that influence decision-maker behavior for the probes in Item 1 for *decision-maker characterizations* and alternative probe *decisions*.
C. Identify *features* $f \in F$ that characterize different *contextual categories* CC where the probes in Item 1 can occur.
D. Establish relations that associate decision-maker characterizations with probe decisions for all contextual categories. These relations guide the population of values for the features.
E. Stratify the data by creating testing and training sets.

A. Identify Plausible Probes

As previously discussed, DMA can be investigated in environments where humans make suboptimal decisions. Therefore, the dataset must include probes that simulate the dilemmas humans should face to reveal where they lie in cognitive attributes, which they do by selecting where to compromise. In these environments, we refer to probes that simulate said dilemma as *plausible* for investigating DMA.

TABLE I
HEALTH INSURANCE PLAN EXAMPLE PROBES

Deductible Tier 1 Network (Cost in $)
Deductible In-Network (Cost in $)
Deductible Out-of-Network (Cost in $)
Out-of-Pocket Maximum Tier 1 Network (Maximum Cost)
Out-of-Pocket Maximum In-Network (Maximum Cost)
Out-of-Pocket Maximum Out-of-Network (Maximum Cost)
Preventive Care Services Out-of-Network (Percent Plan Pays)
Primary Care Physician (PCP) In-Network (Co-pay in $)
Primary Care Physician (PCP) Out-of-Network (Percent Plan Pays)
Tele-Medicine In-Network (Co-pay in $)
Specialist Office Visit Tier 1 Network (Co-pay in $)
Specialist Office Visit In-Network (Co-pay in $)
Specialist Office Visit Out-of-Network (Percent Plan Pays)

Health insurance plans are often presented to customers through lists of items with their associated costs and coverage specifying different plans. Table I lists examples of those items that we use as plausible probes for the environment. The decisions of these probes are either dollar amounts or percentages. The uncertainty is in the inability of a family to estimate its future healthcare needs. The domain of cognitive attributes is financial given that health insurance plans describe their value as providing financial risk reduction. We started building the health insurance dataset by selecting 20 such probes. The combination of the probe and the alternative decisions simulates the dilemma, but the alternative decisions depend on the selected cognitive attributes.

We emphasize that, in reality, humans who seek to select an insurance health plan do not have to make individual decisions about each of those probes. All they do is select one health plan. When they select a plan, the decision for each of the probes is entailed. Consequently, the dataset we are generating does not simulate a decision that humans are required to make in any way. The environment and the datasets created from it are not meant to be used by humans but to train and evaluate algorithmic decision makers (*e.g.*, TAD [24]) that have learned the decision-making behavior of target decision makers from these data.

B. Identify Attributes to Characterize Decision Makers

The identification of cognitive attributes that influence decision-maker behavior is an area of study on its own (*e.g.*, [28], [29]. We adopted two attributes for the health insurance plan environment. The first, *risk tolerance* [30] is widely studied and validated. Risk tolerance describes the level of tolerance for risk that a decision maker is willing to take. This is an obvious selection given that insurance plans claim to reduce customer's risk, hence risk tolerance is directly related. For example, a decision maker with a high level of risk tolerance might be more inclined to take a chance by selecting a lower premium that comes with higher deductibles because they are comfortable with such risk.

The second attribute, *choice*, is straightforward and does not require complex interpretation. We refer to this as informal because we did not find it in the literature. In these data, *choice* refers to the number of alternatives of service that a decision maker has within a health insurance plan. A decision maker highly influenced by choice will be interested in plans with low costs for out-of-network services. All probes are influenced by both attributes where one places high influence and the other minimal. The decision-maker characterizations in this data are valued with labels *high* and *low* rather than numbers, where the first is for risk tolerance and the second for choice. Consequently, there are four possible characterizations for target decision makers in the health insurance environment, namely, attribute risk tolerance *low*, attribute risk tolerance *high*, attribute choice *low*, and attribute choice *high*.

TABLE II
FEATURE NAMES AND DATA TYPES

Feature name	Type
children_under_4	Integer
children_under_12	Integer
children_under_18	Integer
children_under_26	Integer
employment_type	Salaried, Bonus, Hourly
distance_dm_home_to_employer_hq	Integer
travel_location_known	Binary
owns_rents	Owns, Rents
no_of_medical_visits_previous_year	Integer
percent_family_members_with_chronic_condition	Integer
percent_family_members_that_play_sports	Integer

C. Identify Features to Characterize Contextual Categories

The third step is to define the features that describe the state for each probe. For this environment, we identified features that may be strongly or loosely associated with the reasons a human would be inclined to select a given health insurance plan. We created 11 features to characterize different contextual categories, they are in Table II. For example, strongly related features would include the number of family members and whether they engage in activities that make them prone to accidents, such as playing sports or having chronic diseases. Loosely related features include whether the customer owns or rents and whether they earn a salary, are eligible for a bonus, or are paid hourly. There are features that are relevant for both attributes (*i.e.*, risk tolerance and choice). One feature related to choice is how far the customer lives from the headquarters of the company for which they work. This is to characterize a potential demand for out-of-network services.

Once the features were defined, we created four contextual categories CC. Each CC represents an expected level of

need for health services, which is the counterpart of the premium[1]. Then, we define probability distributions $P(X)$ that correspond to probability measures that assign values to contextual features $f \in F$ to create subsets of $X' \subseteq X$ for each CC. The selection of probability distributions has the purpose of representing different contextual states. Hence, these contextual categories determine the specific values and weights used in each probability distribution. As an example, the number of children under 12 years of age is populated with random values $\{0, 1, 2, 3, 4\}$ with a variable weight determined for each contextual category CC.

D. Establish Relations and Alternative Probe Decisions to Populate the Data

Relations between DMA components combined with probability distributions is what enables the creation of datasets with high number of instances without the need for human studies. Each new dataset to be created from an environment such as the health insurance plan is defined by relations between the decision makers' attributes $u \in U$, probe decisions $y \in Y$, and subsets of $X' \subseteq X$ characterizing a finite set of contextual categories CC. See (1).

$$R = \{(u_i, y_i, x'_i) \in U \times Y \times X'\} \tag{1}$$

In other words, with the comprehension of the meaning of each attribute, we can create values for decisions that align with a given decision-maker characterization and a contextual category for each probe. Table III shows relations between these three components for a given probe. For example, at a contextual category where feature values indicate few or no children, a salary with bonus, and an owned home for a decision-maker characterized as high-low, the decision value of the highest deductible is represented for CC 3 and Val1 at the Table III.

TABLE III
CONTEXTUAL CATEGORIES AND DECISION VALUES

	Val1	**Val2**	**Val3**	**Val4**
CC 1	hh	hl	lh	ll
CC 2	na	hh	hl	lh
CC 3	hl	lh	ll	hh
CC 4	lh	ll	na	hh

Legend: CC[x] refers to four different contextual categories; hh: high-high, lh: low-high, hl: high-low, ll: low-low, na: not applicable to any characterization.

The relations in (1) can be considered both a key for the dataset and also a kind of knowledge representation. They represent knowledge because the relation between a decision makers' attributes $u \in U$, probe decisions $y \in Y$, and contextual categories CC must follow the true sense of the cognitive attribute selected with respect to how they would guide decisions based on each different contextual category. The attributes selected for this environment (*i.e.*, risk tolerance and choice) refer to the core of the dilemma in the probes. They are either about cost versus utility (*i.e.*, use of health services) or about cost and available services. Therefore, a high cost would have to be compensated for by a high expectation on the use of health services. Thus, although an optimal decision is not easy to determine, the relationship between risk tolerance and alternative decisions can be estimated. The same happens when choice is the prominent attribute, where the probes associated with choice can be established by the contents of the probes, such as out-of-network services. The relations for a given probe can be visualized in a table such as Table III.

Based on the (2) and (3) below, each of the 20 probes are built with 16 variations. The population of the feature values with probability distributions was done 100 times each. Hence, the total number of instantiated probes is 32,000. This way, we aim to generate a dataset that can be used to learn the decision-making behavior of four target decision makers.

$$U_i \quad i = 1, \ldots, 4 \qquad CC = 1, \ldots, 4 \tag{2}$$

$$R = U \times Y \times X' = 4 \times 4 \times X' = 16 \times X' \tag{3}$$

E. Stratify the Data

We stratified this dataset in four splits to evaluate how well algorithmic decision makers learn the behavior of different decision-makers. In the first split, *50-50*, we randomly selected half of the instances for testing and the other half for training. We made sure the distribution of the two splits is i.i.d. (*i.e.*, independent and identically distributed). This data can be used as a conventional ML supervised dataset.

We created other three splits because environments to investigate DMA are likely not going to be i.i.d.. Therefore, we removed one of the main components studied in DMA in each split. The second split, *context*, had one of the four contextual categories removed from the test set. The third split, *target*, set does not include one decision maker in training. The fourth split, *probe*, does not have in its training data five types of probes out of 20. All test sets include the values for the two attributes for the decision maker the data is asking the algorithms to emulate. The four datasets are available via this link: https://github.com/Rosinaweber/ITM-datasets/tree/main.

IV. COMPUTING ALIGNMENT

This section serves both to illustrate how to compute alignment and to demonstrate the quality of the first dataset we built with the approach described herein. The main goal of investigating DMA is to compute how well algorithmic decision makers learn from data to align with individual decision makers. Therefore, we hypothesize that a quality dataset will support an algorithmic decision maker designed for DMA to learn and align well with the decision makers in that dataset. Because we do not have a baseline to determine what *well* means, we propose to hypothesize that an algorithmic decision maker designed for DMA will perform better than synthetic decision selectors that receive only the

[1]Please note we do not include the premium for each probe because that would make a different type of learning problem given the premium would be the same for a set of selected decisions.

four alternative decisions for each probe and make no use of any training data.

As an algorithmic decision maker designed for DMA, we use the Trustworthy Algorithmic Delegate (TAD) [24], [25]. As synthetic decision selectors (SDS), we use three functions based on either a random function or a simple heuristics to select one of the alternatives. The first SDS is *Random*, which always selects one of the alternative decisions randomly for each probe. The second SDS, the *Spender*, always selects the most expensive alternative decision among those available. The third is *Average*, which always takes the average of all available decisions and selects the one closest to it.

We calculate alignment scores Ψ based on the distance between the selected decision from the ground-truth decision of the target decision maker in the dataset. Given that all probes have a maximum of four available decisions to select, there is a maximum of four discrete distance values that we can compute. When there are exactly four decisions, the alignment score values are mapped to the set $\Psi = 1, 0.67, 0.34, 0$. There are also probes where there are three decisions, those are mapped to $\Psi = 1, 0.5, 0$. (see Table V and Table VI in Appendix B for details).

The methodology to create the dataset we use was described in Section III and a link to its stratified versions made available. We remind the reader that there are three non i.i.d. datasets whose names receive the name of the component that was removed from its training data, namely, *context*, *target*, and *probe*. There are four target decision makers, and each probe is assigned one target. We show the results in charts with percentages of probes that resulted in one of three alignment categories, namely, good alignment, medium alignment, and low alignment. We use percentages because each dataset has a different total number of testing probes. Given the majority of probes are mostly influenced by the cognitive attribute risk tolerance, we show the results for only those probes. The charts show the performance of TAD and the three SDS. More results are available in the Appendix D.

A. Results

The plots in Figs. 1, 2, 3, and 4 depict, respectively, the results for the alignment scores with dataset splits *50-50*, *context*, *target*, and *probe*, produced by TAD and the three SDSs.

The hypothesis is confirmed in the three first charts, *50-50*, *context* and *target* splits but not in the *probe* split. In the 50-50 split, TAD shows better alignment with substantial difference. In the context and target splits, TAD shows superior performance with greater ratio of alignment scores in the good alignment region than the SDSs. In the probe split, TAD is too close or behind the *Average* SDS. TAD also has more probes in the low alignment region. These results suggest that the non-i.i.d. splits may not have enough training data to learn alignment when probes are removed from the training. This is a high demanding task, and further studies should investigate the minimum amount of data for different types of splits where different components are removed. Another aspect to investigate is whether only four alternative decisions represents the dataset is too simple and at least a fifth alternative decision should be added. Even if an added decision would probably not impact TAD's performance, it would certainly be more difficult for the simplistic SDS heuristics. These results only include probes where the most influential attribute is risk tolerance. Results for the probes with stronger influence from the choice attribute, where TAD performs better in the probe split are available in the Appendix Section D for examination.

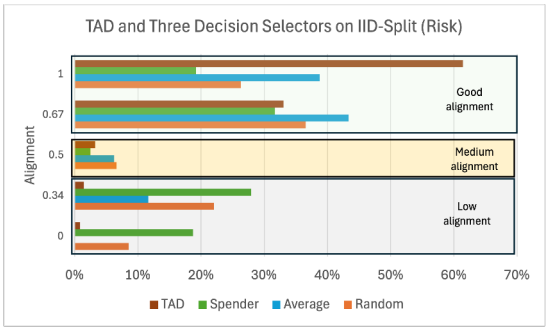


Fig. 1. Alignment comparison of TAD and three SDSs on *50-50* split.

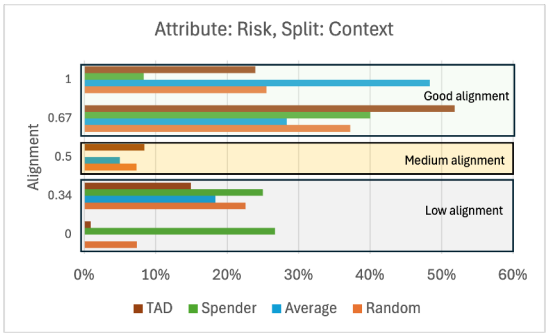


Fig. 2. Alignment comparison of TAD and three SDSs on *Context* split.

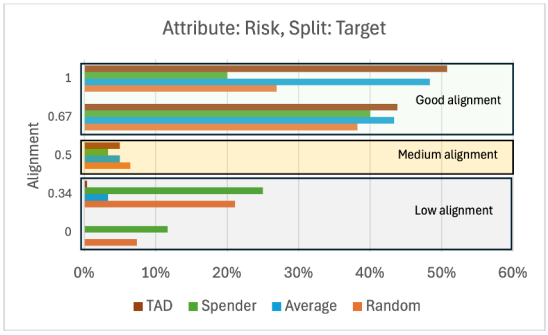


Fig. 3. Alignment comparison of TAD and three SDSs on *Target* split.

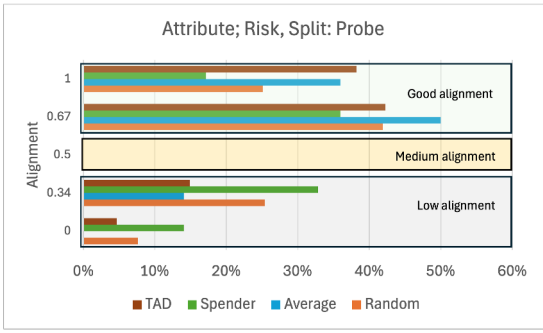


Fig. 4. Alignment comparison of TAD and three SDSs on *Probe* split.

V. Concluding Remarks

This paper describes an approach to building dataset environments to investigate DMA and how to use the resulting data to train and evaluate how well an algorithmic decision maker aligns with different target decision makers. We illustrate the methodology with the creation of one environment and one dataset with four different splits. This approach is an alternative to expensive approaches that rely on human studies and can only produce limited sized data. The core of the approach relies on identifying the main components of a DMA problem and build relations between those components. One of the components, the contextual categories summarize various features representing each category. The features in each category are populated based on probability distributions with weights and ranges that fit the category as a concept class. Every time the feature values are generated, another dataset is created but the patterns of each target decision maker remains the same. For changing the target decision maker patterns, it is necessary to create new relations between the components as per Section III-D.

The environment described in this paper uses plausible probes inspired by health insurance plans. The environment and the datasets that can be built are simple and uses only four alternative decisions, four decision maker characterizations, and four contextual categories. The first dataset we created has 32,000 samples. The methodology can be replicated from realistic probes from different domains and the number of parameters can be increased to produce even larger number of samples in the data.

The results of our study suggests interesting open questions. For example, data is needed for an algorithmic decision maker to learn to align to target decision makers when any of the DMA components, context, target, or probes, are previously unseen.

References

[1] A. Pan, K. Bhatia, and J. Steinhardt, "The effects of reward misspecification: Mapping and mitigating misaligned models," *arXiv preprint arXiv:2201.03544*, 2022.

[2] J. Stray, "Aligning ai optimization to community well-being," *International Journal of Community Well-Being*, vol. 3, no. 4, pp. 443–463, 2020.

[3] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *Scientific reports*, vol. 12, no. 1, pp. 7166–7166, 2022.

[4] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 25–34, 2021.

[5] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ: Pearson, 2016.

[6] P. Slovic, B. Fischhoff, and S. Lichtenstein, "Behavioral decision theory," *Annual Review of Psychology*, vol. 28, no. 1, p. 1–39, 1977.

[7] A. Edland and O. Svenson, *Judgment and Decision Making Under Time Pressure*. Boston, MA: Springer US, 1993, pp. 27–40.

[8] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[9] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the rationality debate?" *Behavioral and Brain Sciences*, vol. 23, no. 5, p. 645–665, 2000.

[10] ——, "Individual differences in rational thought," *Journal of experimental psychology. General*, vol. 127, no. 2, pp. 161–188, 1998.

[11] R. Frey, A. Pedroni, R. Mata, J. Rieskamp, and R. Hertwig, "Risk preference shares the psychometric structure of major psychological traits," *Science Advances*, vol. 3, no. 10, p. e1701381, 2017.

[12] M. Lauriola, I. P. Levin, and S. S. Hart, "Common and distinct factors in decision making under ambiguity and risk: A psychometric study of individual differences," *Organizational Behavior and Human Decision Processes*, vol. 104, no. 2, pp. 130–149, 2007.

[13] J. H. Kirchner, L. Smith, J. Thibodeau, K. McDonnell, and L. Reynolds, "Understanding ai alignment research: A systematic analysis," *arXiv preprint arXiv:2022.4338861*, 2022. [Online]. Available: https://arxiv.org/abs/2022.4338861

[14] Z. Wang, Y. Dong, O. Delalleau, J. Zeng, G. Shen, D. Egert, J. J. Zhang, M. N. Sreedhar, and O. Kuchaiev, "Helpsteer2: Open-source dataset for training top-performing reward models," *arXiv preprint arXiv:2406.08673*, 2024. [Online]. Available: https://arxiv.org/abs/2406.08673

[15] J. Ji, D. Hong, B. Zhang, B. Chen, J. Dai, B. Zheng, T. Qiu, B. Li, and Y. Yang, "Pku-saferlhf: Towards multi-level safety alignment for llms with human preference," *arXiv preprint arXiv:2406.15513*, 2024. [Online]. Available: https://arxiv.org/abs/2406.15513

[16] G. A. Miller, "The cognitive revolution: a historical perspective," *Trends in Cognitive Sciences*, vol. 7, no. 3, pp. 141–144, 2003.

[17] H. A. Simon, "A behavioral model of rational choice," *The Quarterly Journal of Economics*, vol. 69, no. 1, pp. 99–118, 1955. [Online]. Available: http://www.jstor.org/stable/1884852

[18] P. R. Smoliński and H. Brycz, "Individual differences in inaccurate versus accurate economic judgment and decision making. metacognitive approach," *Personality and Individual Differences*, vol. 219, p. 112500, 2024.

[19] "In the moment (itm)," https://www.darpa.mil/research/programs/in-the-moment, accessed: Mar. 7, 2025.

[20] N. D. Shortland, L. J. Alison, and J. M. Moran, *Conflict: How soldiers make impossible decisions*. Oxford University Press, 2019.

[21] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the rationality debate?" *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 645–665, 2000.

[22] D. Kahneman, *Thinking, fast and slow*, 1st ed. New York: Farrar, Straus and Giroux, 2011 - 2011.

[23] C. Gonzalez, "Decision-making: A cognitive science perspective," in *The Oxford Handbook of Cognitive Science*. Oxford University Press, 10 2017.

[24] M. Molineaux, R. O. Weber, M. W. Floyd, D. Menager, O. Larue, U. Addison, R. Kulhanek, N. Reifsnyder, C. Rauch, M. Mainali, A. Sen, P. Goel, J. Karneeb, J. Turner, and J. Meyer, "Aligning to human decision-makers in military medical triage," in *Case-Based Reasoning Research and Development*, J. A. Recio-Garcia, M. G. Orozco-del

Castillo, and D. Bridge, Eds. Cham: Springer Nature Switzerland, 2024, pp. 371–387.
[25] C. B. Rauch, U. Addison, M. Floyd, P. Goel, J. Karneeb, R. Kulhanek, O. Larue, D. Ménager, M. Mainali, M. Molineaux, A. Pease, A. Sen, J. Turner, and R. Weber, "Algorithmic decision-making in difficult scenarios," *Proceedings of the AAAI Symposium Series*, vol. 3, no. 1, pp. 583–585, 2024.
[26] A. Summerville, L. Marti, I. Juvina, L. Welborn, C. Widmer, and A. Leung, "A proof-of-concept validation of alignment in decision-making attributes for trustworthy AI," in *IEEE CAI Workshop on Human Alignment in AI Decision-Making Systems*, 2025.
[27] P. Shafto and O. Nasraoui, "Human-recommender systems: From benchmark data to benchmark cognitive models," in *Proceedings of the 10th ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2016, pp. 127–130.
[28] M. C. Joseph Borders, Alice Leung, "A framework for identifying key decision-maker attributes in uncertain and complex environments," in *Proceedings of the Human Alignment in AI Decision-Making Systems:An Inter-disciplinary Approach towards Trustworthy AI*. Santa Clara, California, USA. May 5-7, 2025: IEEE CAI 2025 Workshop, 2025.
[29] M. Mainali and R. O. Weber, "Exploring cognitive attributes in financial decision-making," in *METACOG-25: 2nd Workshop on Metacognitive Prediction of AI Behavior, SIAM International Conference on Data Mining (SDM25)*, April 2025.
[30] J. E. Grable and J. J. Xiao, "Risk tolerance," in *Handbook of Consumer Finance Research*. New York, NY: Springer New York, 2008, pp. 3–19.

APPENDIX

A. Dataset Specification

TABLE IV
ATTRIBUTE SPECIFICATION OF SYNTHETIC HEALTH INSURANCE DATASET.

Feature Type	Feature Name	Data Type
Contextual Features	no of children under 4	Integer
	no of children under 12	
	no of children under 18	
	no of children under 26	
	distance from decision-maker's home to employer headquarter	
	no of medical visits previous year	
	percentage of family members with chronic conditions	
	percentage of family members playing sports	
	employment_type	String
	owns or rents a house	
	travel location known	Boolean
Probe Features	probe_id	Integer
	network_status	String
	expense_type	String
Alternative Decisions	option one	Integer
	option two	
	option three	
	option four	
Cognitive Attributes	attribute - two possible values: RISK, CHOICE	String
	attribute value - two possible values: HIGH, LOW	String
Ground Truth	Action (a value from the four options)	Integer

B. Alignment Score Mapping

TABLE V
ALIGNMENT SCORE CALCULATION TABLE WHEN THERE ARE FOUR DISTINCT OPTIONS TO ANSWER A PROBE

		Ground Truth			
		highest	2nd highest	2nd lowest	lowest
Prediction	highest	1	0.67	0.34	0
	2nd highest	0.67	1	0.67	0.34
	2nd lowest	0.34	0.67	1	0.67
	lowest	0	0.34	0.67	1

TABLE VI
ALIGNMENT SCORE CALCULATION TABLE WHEN THERE ARE THREE DISTINCT OPTIONS TO ANSWER A PROBE.

		Ground Truth		
		highest	mid	lowest
Prediction	highest	1	0.5	0
	mid	0.5	1	0.5
	lowest	0	0.5	1

C. Alignment Score of Synthetic Decision Selector (SDS)

1) Alignment Score for Random: The Random SDS randomly selects a value from either three or four distinct values from a given list of options. First, a value is randomly chosen from the list and designated as the "aligned" value. The process is shown in Algorithm 1.

Algorithm 1 Random SDS for computing alignment scores

Input: $options$ (list of values)
Output: Three alignment scores
$aligned \leftarrow$ RandomChoice($options$)
if Length(DISTINCT(($options$)) == 3 **then**
    Compare ground truth decision with $aligned$ to compute alignment score using Table VI of Appendix B.
else
    Compare ground truth decision with $aligned$ to compute alignment score using Table V of Appendix B.
end if

2) Alignment Score for Spender: The Spender SDS for computing alignment scores algorithm evaluates and computes alignment scores based on a set of options and an expense type. The algorithm begins by sorting the options based on the specified expense type. If $expense\_type$ is either 'MAXIMUM COST' or 'PERCENT PLAN PAYS', the options are sorted in ascending order; otherwise, they are sorted in descending order. After sorting, the first value in the sorted list is assigned as the "aligned" value. Following this selection, the algorithm checks the number of distinct values in the options list. If there are exactly three distinct values, it compares the ground truth decision with the aligned selection, computing the respective alignment scores using a predefined table (Table VI). If there are more than three distinct values, a different table (Table V) is used to compute the alignment score.

Algorithm 2 Spender SDS for computing alignment scores

Input: $options$ (list of values), $expense\_type$
Output: Three alignment scores
if expense_type = 'MAXIMUM COST' or expense_type = 'PERCENT PLAN PAYS' **then**
    $options \leftarrow$ sort $options$ in ascending order
else
    $options \leftarrow$ sort $options$ in descending order
end if
$aligned \leftarrow options[0]$
if Length(DISTINCT($options$)) == 3 **then**
    Compare ground truth decision with $aligned$ to compute alignment score using Table VI of Appendix B.
else
    Compare ground truth decision with $aligned$ to compute alignment score using Table V of Appendix B.
end if

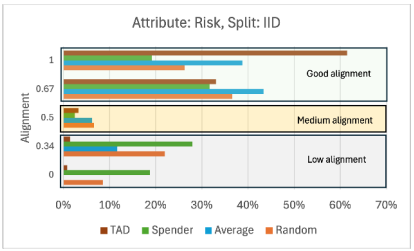
3) Alignment Score for Average: The Average SDS dynamically selects alignment scores based on the distinctness of the available input options. It first sorts the options in

descending order and checks the number of distinct values. If there are three distinct values, it selects the second value as the ”aligned” value. The algorithm then compares the ”aligned” value with a ground truth decision to compute alignment scores using a predefined similarity table (Table VI) for three values. If there are four distinct values, the algorithm calculates the average value and selects the closest one as the ”aligned” value. The working process is outlined in Algorithm 3. For the four available options, the alignment score is calculated using Table V by comparing the ground truth with the decision.

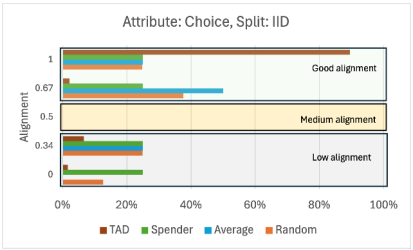
Algorithm 3 Average SDS for computing alignment scores

Input: $options$ (list of values)
Output: Three alignment scores
$options \leftarrow$ sort $options$ in descending order
$unique\_values \leftarrow$ set($options$)
if $length(unique\_values) == 3$ **then**
    $options \leftarrow$ sorted($unique\_values$, reverse=True)
    $aligned \leftarrow options[1]$
else
    $avg \leftarrow$ sum($options$) / length($options$)
    if abs($avg - options[1]$) < abs($avg - options[2]$) **then**
        $aligned \leftarrow options[1]$
    else
        $aligned \leftarrow options[2]$
    end if
end if
Score calculation for three or four distinct options is similar to Algorithm 1 or Algorithm 2.

D. Alignment Score Results Comparison



(a) Alignment score on i.i.d. split based on *Risk*.



(b) Alignment score on i.i.d. split based on *Choice*.

Fig. 5. Alignment score of TAD and three SDSs based on *Risk* and *Choice* for the *i.i.d.* split of dataset.

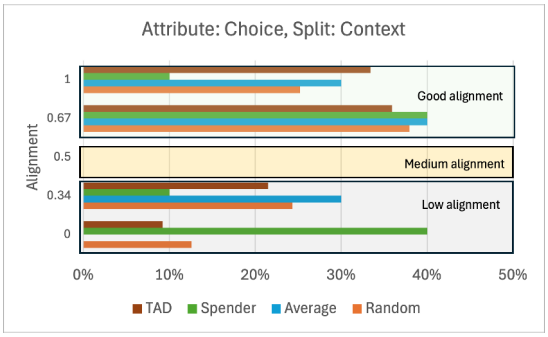


Fig. 6. Alignment score of TAD and three decision selectors on *context* split based on *Choice*.

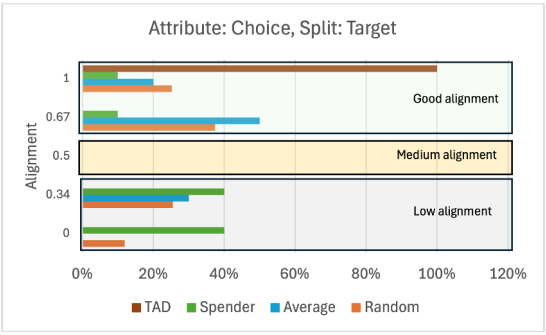


Fig. 7. Alignment score of TAD and three decision selectors on *target* split based on *Choice*.

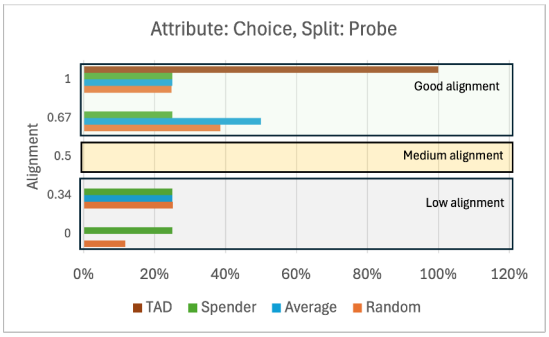


Fig. 8. Alignment score of TAD and three decision selectors on *probe* split based on *Choice*.