Second Workshop on Metacognitive Prediction of AI Behavior

Relevance Scoring as a Feature of Metacognitive Artificial Intelligence

Alexander M. Berenbeim, *United States Military Academy, West Point, NY, 10996, USA*
Nathaniel D. Bastian, *United States Military Academy, West Point, NY, 10996, USA*

Abstract—Conventional artificial intelligence (AI) lacks a level of adaptability required for the dynamic environments and threats faced by safety-critical tasks. Metacognitive AI, which is AI technology capable of introspection, self-monitoring, and self-adaptation, would be a development that can improve the explainability and trustworthiness of systems for human operators, and improve model development and resilience when encountering new domains. In this paper, we provide a novel framework for AI improvement in operational and cross-platform domain settings in the form of a well-defined formal foundational framework relating AI metacognition, training, and past performance that extends the certainty and competence framework for discriminative models through the introduction of relevance structures, and related scores. We introduce the theory of relevance structures, provide a preliminary accounting of meta-objective functions relative to formal task specification based on relevance, and discuss how such a framework accounts for Knightian uncertainty as encountered by AI models in open-world environments. The intent of this paper is to outline the potential of this system and demonstrate that the notion of relevance is a necessary prerequisite for AI-enabled systems capable of metacognition.

Along with the rapidly increasing development and integration of artificial intelligence (AI) into everyday life, including safety-critical systems, there is a growing need for robust and resilient systems for assurance, monitoring, and regulation in order to improve reliability, error detection and correction, and efficient resource utilization. Metacognitive AI is a promising research area aiming to improve AI-enabled systems by developing capabilities for self-monitoring and self-regulation beyond straightforward expected utility maximization.[1] We propose that the certainty and competence framework developed by Berenbeim et al. can be suitably generalized for developing metacognitive AI by providing a means of furnishing AI systems an internal representation of their capabilities for the purposes of self-monitoring and self-regulation.[2] The certainty and competence framework treats certainty as an intrinsic property of models indicating the relative confidence between decisions, while competencies are scores formed from certainties and the empirical distribution of data.[2] When competence is high, high certainty signals reliable assignments over lowers certainty, which would signal misassignment or possible novelty; this can be a catalyst for introspection.[2] We expand on the certainty and competence framework, proffering the notion of *relevance*, to apply to general classes of AI models where the output of f can be placed into a proximate relation with a target object. The target object can be a singleton set with one label as in the classical case, multisets, bounding boxes, trees, or even hierarchical data structures with metric.

RELEVANCE STRUCTURES

Our definition for *relevance* is guided by our intuition that the output s of an AI model f should overlap with a target t, so that whenever t is properly contained in s, the extraneous information provided by s is deemed *irrelevant*, and whenever s is strictly contained in t, we may only have *partial satisfaction*. These intuitions remain independent of whether the

Digital Object Identifier 10.1109/XXX.0000.0000000

model f is *discriminative* with target t or *generative* with prompt identifying a target t. We first define *relevance structures* before defining notions of *relevance* relative to said structures.

A *relevance* structure $\mathcal{R}$ is defined by a signature $\langle \mathcal{U}, 0, \preceq, \frown, \setminus, |\cdot| \rangle$ satisfying the following:

1) $\langle \mathcal{U}, \preceq \rangle$ is a partial order;
2) for all $x \in \mathcal{U}$, $0 \preceq x$;
3) $\frown$: $\mathcal{U} \times \mathcal{U} \to \mathcal{U}$ such that for all $x, y, z \in \mathcal{U}$, $x \frown y \preceq x, y$ and if $z \preceq x$ and $z \preceq y$, then $z \preceq x \frown y$;
4) $\frown$ is commutative, associative, and idempotent;
5) $\setminus : \mathcal{U} \times \mathcal{U} \to \mathcal{U}$ such that for all $x, y, z \in \mathcal{U}$,
   - $(x \setminus y) \frown (y \setminus x) = 0$;
   - if $y \frown z = 0$, then $x \frown z \preceq x \setminus y$;
6) $|\cdot| : \mathcal{U} \to \mathbf{R}_{\geq 0}$ such that whenever $x \preceq y$, then $|x| \leq |y|$, and $|x| = 0$ if and only if $x = 0$.

A special subclass of relevance structures can be identified with posets which are closed under relative pseudo-complements. If the poset $\mathcal{U}$ of a relevance structure is closed under relative pseudo-complements, that is, there is a relation $\to$ such that $x \to y$ denotes the maximal element such that $x \frown (x \to y) \preceq y$, then we may not only define $\neg x :\equiv x \to 0$, but define $x \setminus y :\equiv x \frown \neg y :\equiv x \frown (y \to 0)$.

We say $f : \mathcal{R} \to \mathcal{R}'$ is a *relevance morphism* between relevance structures if:

i. f is a *poset-morphism*, $x \preceq^{\mathcal{R}} y \Rightarrow f(x) \preceq^{\mathcal{R}'} f(y)$;
ii. f is *meet-preserving*, $f(x \frown^{\mathcal{R}} y) = f(x) \frown^{\mathcal{R}'} f(y)$;
iii. $f(0) = 0$;
iv. induces an order-preserving map $\bar{f} : \mathcal{R} \to \mathcal{R}'$ such that $\bar{f} \circ |\cdot|_{\mathcal{U}} = |\cdot|_{\mathcal{U}'} \circ f$.

If $\bar{f}$ in (iv.) induces to a field morphism between $\mathbf{R}$ and $\mathbf{R}'$, we say f is an *algebraic relevance morphism*. We let `Relv` denote the category of relevance structures, and `ARelv` the category of relevance structures whose morphisms consist solely of algebraic morphisms. For all practical considerations, we may fix $\mathbf{R}$ to be the field of real-numbers, $\mathbb{R}$.

Relevance structures can be found across many mathematical topics and for structures of interest. Of general interest to descriptive set theory, probability theory, and the study of dynamical systems, we find:

Theorem 1:

Both $\mathcal{C}$ and $\mathcal{N}$, Cantor and Baire space respectively, can be given relevance structures.

$\mathcal{C}$ can straightforwardly be given a canonical relevance structure. Since we require that $|0|_{\mathcal{N}} = 0$, in general we cannot straightforwardly define a relevance structure on $\mathcal{N}$ with respect to a given embedding of $\mathcal{N}$ into $\mathbb{R}$.

A *symmetric relevance structure* is any such $\mathcal{R}$ that is further closed with respect to a join operation $\smile$. Consequently,

Theorem 2:

If $\mathcal{L}$ is a relatively complemented lattice, then $\mathcal{L}$ can be endowed with a $\setminus$ function and value map such that $\mathcal{L}$ is a relevance structure.

Many of our target use cases can be given an underlying Heyting algebra structure, and since $\mathcal{L}$ above applies to bounded lattices as well, any practical logic thus can be endowed with a relevance structure.

We note that in any given symmetric relevance structure we can define the symmetric difference operation Δ point-wise as $x \Delta y := (x \setminus y) \smile (y \setminus x)$ for arbitrary x, y as $\mathcal{U}$ will be closed under $\smile$, whereas in arbitrary relevance structures we cannot guarantee the existence of joins.

In general relevance structures, for any non-bottom element $s, t \in \mathcal{U}$ we can define two *directed relevance* scores where t denotes a *target* and s denotes a *response* that provide penalized response scores emphasizing *exploration* and *parsimony* respectively.

$$\text{[Exploratory Relevance]} \qquad \delta_t(s) := \frac{|s \frown t| - |t \setminus s|}{|t|} \tag{1}$$

$$\text{[Parsimonious Relevance]} \qquad \rho_t(s) := \frac{|s \frown t| - |s \setminus t|}{|t|} \tag{2}$$

where $\delta_t(s)$ is defined to penalize responses s for what is missing in the target. On the other hand, the parsimonious relevance function $\rho_t(s)$ will penalize responses s that exceed the scope of t. Further, parsimonious directed relevance can be shown to be unbounded below, i.e. $\rho_t(s) \in (-\infty, 1]$ depending on choice of $|\cdot|$ and s. When recording either or both relevance scores, we report them as the *relevance of s to t* (or *with respect to t*).

For any symmetric relevance structure, we can further define the *mutual relevance of s and t* by:

$$\text{[Symmetric Relevance]} \qquad \rho(s,t) := \frac{|s \frown t| - |s \Delta t|}{|s \smile t|} \tag{3}$$

In particular, we have defined mutual relevance so that $\rho(s,t)$ is a real-valued function whose image lies within $[-1,1]$. Further, we have defined ρ so that when s agrees with t relative to the underlying poset structure, we attain a value of 1, and when s maximally disagrees with t, we attain a value of -1. Recall the definition of (classical) empirical competence:[2]

$$\mathcal{E}(f;D_N) = \frac{1}{|D_N|}\sum_{(x_i,y_i)} \varsigma(f(x_i))[\mathbf{1}_{\equiv y_i} - \mathbf{1}_{\not\equiv y_i}] \tag{4}$$

where $\varsigma(f(x_i))$ is the certainty score of f on sample input x_i. Recognizing that the term $[\mathbf{1}_{\equiv y_i} - \mathbf{1}_{\not\equiv y_i}]$ is the symmetric relevance $\rho(f(x_i), y_i)$ in a discriminative single-label model, we define general empirical competence with respect to ρ, so that *empirical competence of f at task T with relevance structure* $\mathcal{R}$ is a sample average of the product of the certainty score with the relevance of the prediction relative to the target:

$$\mathcal{E}(f;\rho;D_N) = \frac{1}{|D_N|}\sum_{(x_i,y_i)} \varsigma(f(x_i))\rho(f(x_i);y_i) \tag{5}$$

Similarly, we may define the *empirical exploratory competence* and *empirical parsimonious competence* with $\delta_t(s)$ and $\rho_t(s)$.

EXAMPLES OF RELEVANCE STRUCTURES

Relevance structures have been identified for single-label, multi-label, bounding boxes, multiset, poset identification, and multi-poset related tasks. The latter two tasks include those involving formal concept analysis and general hierarchical data structures. Through the use of multi-posets, multi-domain relevance scores can be composed from individual task relevant scores. Initially, relevance scores should be computed for models trained on labeled datasets, although in principle suitably specified relevance structures allow for training involving unlabeled data types. Further, the application of certainty and competence scores with the competence framework have both theoretical and empirical support for the identification of out-of-distribution data, and providing human agents real-time information of model performance.[2,3,4] The empirical work of Berenbeim et. al[2] examining the certainty and competence framework for single-label classification tasks demonstrated that certainty-based features can be used to develop out-of-distribution detection tests that outperform state-of-the-art Energy-Based detection tests, where test performance strongly correlated with model competence, improving upon the baseline Monte Carlo Dropout AUPR-OUT performance on average by 14.4% and 16.5%, and reducing the FPR95TPR by 54.2% and 37.6% across network traffic classification and image classification tasks. Relatedly, neurosymbolic AI systems implemented with logic tensor networks, where neural network classifiers were augmented by axioms guiding classification, had demonstrable improvement in empirical competence within categories over baseline neural network models - for some categories moving from -.994 and -.976 empirical competence to .574 and .694 respectively.[3] These improvements occurred independent of whether data was balanced or unbalanced indicates, and accompanied similar shifts in the F1 score, corroborating the intuition that model competence can indicate a balance between precision and recall, with the added benefit of providing signals that False Positives occur with lower certainty than in category True Positives.[3]

Similar empirical results are expected to follow for multi-label and multiple image recognition tasks. To give a concrete example of a relevance structure for multi-label cases, consider $\langle \mathcal{U}, \preceq \rangle$ to be the power-set over some finite set of labels with ordinary inclusion, and $|\cdot|$ the order map counting the number of objects in the subset. From here, relevance scores can be directly computed with respect to counting the order of the sets, while the certainty score would require computing the differences of the relative implied probabilities of the proposed subset against the next alternative subset. Importantly, the certainty score is not a distance between the proposed solution and the true solution, but a confidence score in the proposed solution over the next alternative.

One can similarly treat the task of bounding box recognition by considering an underlying topology for the unit square whose subbase is generated by *bounding boxes B*, whose coordinates correspond to the rational representation of the corner pixel positions. We can denote efficiently each bounding box as an ordered pair of pairs, so that $B = \langle (x_l, y_l), (x_h, y_h) \rangle$. So for instance, if we normalize a 256x220 pixel image to the unit square format, if $B = \langle (\frac{40}{256}, \frac{2}{220}), (\frac{100}{256}, \frac{56}{220}) \rangle$, would be a bounding box in the lower part of the third quadrant of the unit square from the center. From here, ordinary set operations apply up to and including relative complements, where the natural choice for the

order function $|B|$ is the area of the figure generated by individual bounding boxes. In this setup, we can handle instances where our target consists of one bounding box, many, or otherwise general rectilinear shapes. This setup can be further amended to allow for general polygonal recognition within images by also allowing the generating set to account for triangles formed by taking the upper or lower half of each generating bounding box.

FORMAL TASK SPECIFICATION AND CONTROL

There is a vast and growing literature relating category theory to formal specification,[5,6] control theory,[7,8] and machine learning.[9] There is a deep connection between computation and control theory that can be properly leveraged through Galois connections and adjunctions, viz institutions, dependent lenses, Bayesian open games, and their composition.[5,8] Suitably identified complete partial orders (CPO) can be endowed with relevance structures, which contain inherent fixed points that can inform control structure; every computation has a dual form in a control structure - relevant to machine learning, every forward pass through a neural network is adjoint to back-propagation through training.[9] Moreover, whenever relevance CPOs are identified with formally specified tasks, future metacognitive AI architectures may be capable of implementing formal concept analysis between tasks identified through relevance scores; this would provide a means of AI-enabled systems to assess whenever their outputs are in relevant alignment to given tasks.

METACOGNITION AND KNIGHTIAN UNCERTAINTY

Evidence is being gathered that AI models as presently trained become less corrigible over time, and become resistant to belief updates.[10] If models become less corrigible with greater training, they become increasingly vulnerable to Knightian uncertainty, especially as they encounter data outside of previously formally well-defined tasks beyond their symbolic vocabulary and prior probability distributions. Townsend et al. identify that the benefits of practical AI-enabled systems are contingent upon four interrelated problems where situational uncertainty cannot be quantified or measured with respect to probability weight assignments: actor ignorance, practical indeterminism, agentic novelty, and competitive recursion.[11]

Because relevance scores are defined as intrinsic point estimates of models themselves relative to their empirical performance, we propose their use to address metacognitive AI challenges, particularly those concerning Knightian uncertainty and the problem of incorrigibility. Specifically, relevance can be used to allow AI-enabled systems to recognize the boundaries of their own competence (introspection) determined by their training and previous deployment in field. Relevance structures immediately address the problem of actor ignorance, which refers to the difficulty faced in specifying whether different possibilities are relevant to a decision outcome, as well as the problem of practical indeterminism faced when multiple unknown future possibilities generated by the decision environment have yet to be specified, including those of *agentic novelty*, which may be introduced by users or other AI agents themselves.[11] Incorrigibility may be remedied by more heavily penalizing models which overestimate their competencies, and especially those which fail to identify the human agents consensus around relevant responses.

This use of relevance structures and competence as a means to address actor ignorance and indeterminism finds support in Berenbeim et al.[2], specifically experiments involving out-of-context image recognition. Berenbeim et al. demonstrated that using competence and certainty guided detection methods with the outputs of pre-trained ResNet model yielded better performance than comparable EnergyBased out-of-distribution detection methods, particularly indicating suitability for actor ignorance and agentic novelty with minimal overhead training.[3] Further, Berenbeim et al. introduce cost-functions for competence optimization, which would improve corrigibility during training.[2] These cost-functions can be generalized using the relevance structure framework introduced here.

CONCLUSION

The use of relevance scoring to provide immediate internal displays for metacognitive AI is in keeping with the suggestion of Tankelevitch et al. that it is necessary for AI-enabled systems to have the ability to self-monitor and self-recognize when its own cognitive processes are inadequate and require adaptation and additional input from the user.[12] Moreover, implementations of relevance scoring for control helps address four of the AI failures that can be addressed by metacognitive AI: transparency, reasoning, adaptation, and perception;[12] and mediating conflicts that arise from actor ignorance, practical indeterminism and agentic novelty.

ACKNOWLEDGMENTS

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Military Academy, Department of the Army, Department of Defense, or U.S. Government.

References

1. P. Shakarian, G. I. Simari, and N. D. Bastian, *Probabilistic foundations for metacognition via hybrid-ai*, 2025. arXiv: 2502.05398 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2502.05398.
2. A. M. Berenbeim, A. D. Cobb, A. Roy, S. Jha, and N. D. Bastian, "Applications of certainty scoring for machine learning classification and out-of-distribution detection," *ACM Transactions on Probabilistic Machine Learning*,
3. A. Berenbeim, R. Kaur, A. D. Cobb, A. Roy, S. Jha, and N. D. Bastian, "Post-hoc uncertainty quantification for neurosymbolic artificial intelligence," in *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*, IEEE, 2024, pp. 1–6.
4. A. M. Berenbeim, A. V. Wei, A. Cobb, A. Roy, S. Jha, and N. D. Bastian, "Bayesian graph representation learning for adversarial patch detection," in *Assurance and Security for AI-enabled Systems*, SPIE, vol. 13054, 2024, pp. 54–70.
5. J. A. Goguen and R. M. Burstall, "Introducing institutions," in *Workshop on Logic of Programs*, Springer, 1983, pp. 221–256.
6. R. Diaconescu, *Institution-independent model theory*. Springer Science & Business Media, 2008.
7. J. Soto-Andrade and F. J. Varela, "Self-reference and fixed points: A discussion and an extension of lawvere's theorem," *Acta Applicandae Mathematica*, vol. 2, pp. 1–19, 1984.
8. J. Bolt, J. Hedges, and P. Zahn, "Bayesian open games," *Compositionality*, vol. 5, 2023.
9. V. Abbott, T. Xu, and Y. Maruyama, "Category theory for artificial general intelligence," in *International Conference on Artificial General Intelligence*, Springer, 2024, pp. 119–129.
10. M. Mazeika, X. Yin, R. Tamirisa, *et al.*, *Utility engineering: Analyzing and controlling emergent value systems in ais*, 2025. arXiv: 2502.08640 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2502.08640.
11. D. M. Townsend, R. A. Hunt, J. Rady, P. Manocha, and J. h. Jin, "Are the futures computable? knightian uncertainty and artificial intelligence," *Academy of Management Review*, no. ja, amr–2022, 2023.
12. L. Tankelevitch, V. Kewenig, A. Simkute, *et al.*, "The metacognitive demands and opportunities of generative ai," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, May 2024, 1–24. DOI: 10.1145/3613904.3642902. [Online]. Available: http://dx.doi.org/10.1145/3613904.3642902.

Alexander M. Berenbeim is currently a Senior Artificial Intelligence Researcher at the Army Cyber Institute within the United States Military Academy at West Point. He received his Ph.D from the University of Illinois at Chicago. His primary research interests are the research and development of neuro-symbolic artificial intelligence frameworks for open-world reasoning. Contact him at alexander.berenbeim@westpoint.edu.

Nathaniel D. Bastian is currently Chief Scientist and Director, Office of Science & Engineering at the Army Cyber Institute within the United States Military Academy at West Point. He received his Ph.D. from the Pennsylvania State University. His primary research interests are artificial intelligence security, assurance and robustness, including uncertainty quantification, with defense and cybersecurity applications. Contact him at nathaniel.bastian@westpoint.edu.