Learning to Infer Causal Structure over Time

Randall C. O'Reilly[1†], co-authors..

[1] Department of Psychology and Neuroscience
University of Colorado Boulder

[†]To whom correspondence should be addressed; email: randy.oreilly@colorado.edu

A cornerstone of human intelligence is the ability to make complex inferences about novel situations involving multiple interacting actors, based on prior knowledge about relevant relationships or proclivities. Here, we explore the neural basis of this ability through biologically-based neural network models that learn about causal relationships by observing (abstracted) event sequences over time, and can then generalize this knowledge in sophisticated ways to novel test cases. We show that bidirectional activation dynamics and the ability to integrate information over time are critical to these abilities, along with a learning mechanism based on learning from errors in predictions. In addition to providing a novel account of a wide range of central data in human causal reasoning, our model establishes the importance of several core biological mechanisms underlying advanced human cognitive capabilities.

How do the neural networks in our brains learn to encode the relationships among different entities, and apply this *relational structure* to making relevant inferences? This capacity is central to human intelligence, as manifest in everyday situations, or more challenging tests. For example, if I tell you that *"my car is in the shop"*, you can infer many further things that are likely to be true, e.g.,: I'll be even later to meetings; I'll be taking phone calls during these meetings (which I normally wouldn't do); And I might be kind of grumpy (which I am normally not). And if it turns out that these predictions are false, you might start to wonder if I have a second car. All of these inferences are very different than if I told you my car was in the *lot*, instead of the *shop*. Or that it was *not* in the shop. Critically, this inferential power is not well captured by existing pattern recognition mechanisms, which end up being dominated by overall similarity structure (where *"My car is in the shop"* and *"My car is not in the shop"* are highly similar), and not *relational* structure, which is the essential information that drives inferences.

Advocates of the Bayesian modeling framework have emphasized the importance of relational structure,

and made claims that brain-based neural network models are not capable of encoding such information in a way that then supports flexible inferences (*1*). Consistent with this perspective, much of the focus in machine learning has been on the problem of pattern recognition, and recently "deep" neural networks have proven to be the most successful at challenging tasks such as invariant object recognition (*2–5*). Thus, although these neural network models capture the synergy between certain features of the brain and a functional understanding of what enables our own powerful perceptual skills (*6, 7*), they do little to address the fundamental challenge posed by the Bayesian modelers.

The central goal of this paper is to show how neural networks can learn, through a biologically and psychologically plausible learning mechanism, to encode relational structure, and apply it to making relevant inferences. To move beyond the limitations of the pattern recognition framework, we introduce two essential ingredients: *the ability to integrate information over time*, and *bidirectional (top-down and bottom-up) activation dynamics*, which allow information from different channels to be integrated and processed in flexible ways. Specifically, interactions among entities unfold over time in ways that reveal their relationships, and likewise, inferences also need to unfold over time. This is evident in the most basic form of relationships: *causal* relationships, which unfold over the arrow of time: event A must precede event B for A to have caused B. In addition to time, we argue that *recurrence* or *bidirectional processing* is essential for encoding and inferring relational structure: one's knowledge about two entities in a relationship must be *mutually constraining* on the representations of these entities: if you say that *Jane loves Bill*, this has immediate, bidirectional implications for each individual (though they remain quite underspecified with just that one statement).

To test our hypothesis that time and bidirectional dynamics are the essential additional ingredients needed to support structured inferential reasoning in a neural network, we employed a recently-developed neural algorithm known as *LeabraTI*, which contains both of these elements (*8*). This algorithm builds upon the recurrent processing in the longstanding *Leabra* algorithm (*9, 10*) by adding a temporal integration (TI) mechanism, which enables it to process temporally-extended sequences of inputs in a powerful way. LeabraTI represents a synthesis between biologically-motivated ideas about the different contributions of the deep versus superficial layers of neocortex (and their interconnections with the thalamus), and the computational power of the simple recurrent network (SRN) (*11–13*). Specifically, we associate the role of a context layer in the SRN with the deep layers of the neocortex, such that every cortical area has its own temporal context representation, and, unlike in the standard SRN, these context representations can interconnect directly and indirectly in powerful ways. See the supplemental data and (*8*) for more information.

We applied the LeabraTI model to the classic *Blicket* detector task of causal reasoning, where a blicket

is defined as a particular type of object that the blicket detector detects, and you only know when something is a blicket when the detector goes off (*14–17*). One particularly important case of causal reasoning in this task is known as *screening off*, where you initially put two objects on the blicket detector, and the light goes off. Then, you put one of the two objects alone on the detector, and the light does *not* go off. Immediately, you can infer that the *other* object must have been a blicket. This inferential leap is important because it involves updating your understanding of an object which is not within the current focus of attention — simple "associationist" models would seem to falter because of this. Interestingly, children as young as 4 years old can make this inference. But on the other hand, four years represents a very long time available to acquire this kind of causal knowledge through experience with a wide variety of situations. Thus, we hypothesize that the reason these children are capable of the inference is not that they have amazing powers of deduction, but rather that they have *amazing powers of generalization:* they can represent their causal knowledge acquired across a large number of prior experiences in a sufficiently abstract manner, that it then generalizes to novel situations such as the blicket detector task. This is essentially the standard answer that neural network models always give to these types of questions, but here we demonstrate that it gains particular power when combined with temporally-extended processing and recurrent activation dynamics.

In addition to the screening off and closely related *backwards blocking* cases, we also address a set of other challenging phenomena, including the sensitivity of the model to deterministic vs. stochastic behavior, and to the apparent base rate probabilities (*15, 16*). Taken together, our results demonstrate that networks can perform relatively sophisticated forms of inference, based on simple learning mechanisms, operating over time, and with the advantages of bidirectional processing. Our account differs in many ways from an earlier neural-network based model (*17*), which required a number of specialized mechanisms and systems orchestrated in a specific way by the modeler to solve these kinds of problems, in contrast to the unified, single-system approach used here. Nevertheless, both models emphasize the importance of bidirectional activation dynamics in this domain.

The LeabraTI causal inference model (Figure 1) is designed to provide a simple, minimalist framework for learning these causal tasks (see supplemental online material for full details). There are three different input channels, so the network can encode up to three different objects present at a time, and each channel has high-level semantic features that distinguish between the following classes of inputs: strong agent (e.g., a person, who typically is the causal agent); weak agent (e.g., requiring multiple such agents to achieve a causal outcome); object (never a causal agent); probabilistic strong agent (a strong agent with noisy causal behavior); probabilistic object (an object that is also noisy – sometimes appears to be causal); unknown (a novel input – e.g., a blicket); and blank, which indicates the absence of an input in this channel. These inputs

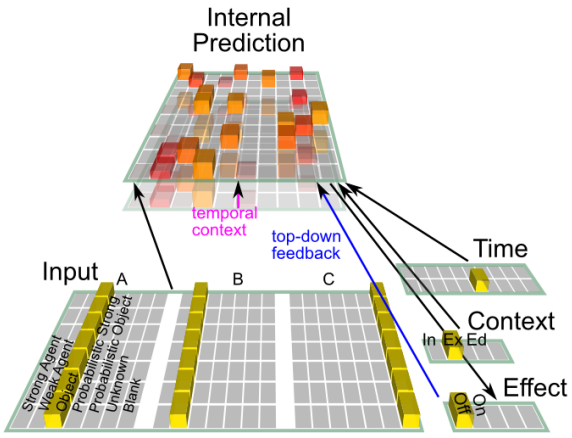

Figure 1: The LeabraTI causal inference network. Entities in the environment are encoded through 3 input channels, which represent high-level semantic representations that distinguish between inputs that are likely to be causal agents (e.g., people), weak agents that require multiple such agents to effect an outcome, and objects that are unlikely to be causal agents. Noisy versions of agents and objects are also encoded. For testing in the blicket-like tasks, the inputs activate the *unknown* category, and their causal status must be inferred as a result of the pattern of *Effect* activations over time, which feed back via bidirectional connections to the internal prediction layer to shape its representations. This internal layer also has the deep cortical temporal context layer, which enables this information to be integrated over time. The Time and Context inputs reflect the structure of the task, and are useful for visualization and decoding, but do not affect network behavior significantly.

all feed into a common internal processing and prediction layer, which has an associated temporal context layer that corresponds to the deep cortical layers in the LeabraTI framework. The primary output of the model is a prediction, generated by this internal prediction layer, of whether there will be a causal outcome of some kind (e.g., the blicket detector going off), as a result of the inputs. There are also context and time inputs that define the structure of an input sequence (these do not affect network performance significantly, but are useful for visualization and decoding). A typical input sequence involves presentation of the full set of objects present in a given scene, with no effects (to give the network the relevant scene context), followed by a sequence of object / effect trials, which were repeated twice. For example, during training, a strong agent and an object could be present, and the strong agent alone or in combination with the object causes the effect, while the object alone does not. All possible combinations of inputs and sequences were presented during training, simulating the background experience someone would have coming into a blicket experiment. This includes 20% of trials that included an unknown object (which was randomly assigned with uniform probability to a causal status), so that the network gained familiarity with the meaning of these inputs. A total of 16,000 different causal input scenes were presented during training (although network performance was already quite good after 4,000).

During testing, which followed training, only unknown inputs were presented, and the network had to infer their causal properties based on the pattern of effects across trials. The screening off and backwards

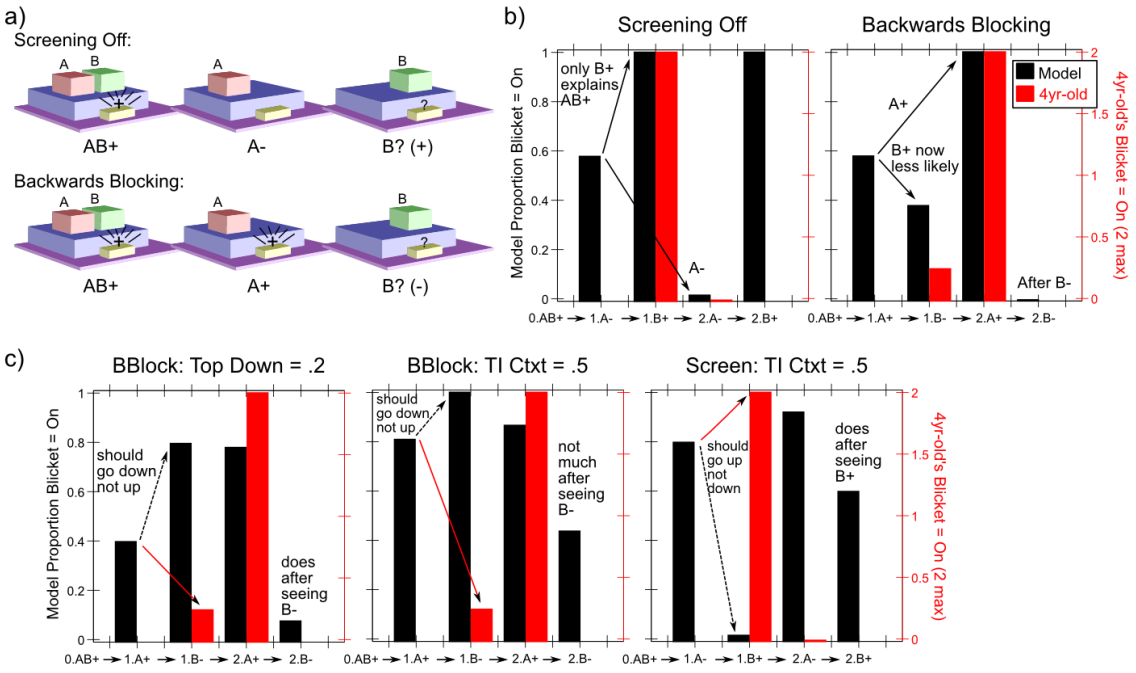

Figure 2: The screening off and backwards blocking tests, and results (N=25 batches, SEM error bars too small to see). **a)** shows the two test sequences of interest. **b)** shows the average output of the Effect On unit for the two sequences, compared to 4 year-old's performance from Sobel et al., (2004). These Effect outputs were measured in the minus phase of each trial, when the network takes its guess as to what will happen, and then it experiences the actual effect as indicated (+ = On, – = Off). The network is initially uncertain about the causal status of the A input, but after experiencing either the + (in blocking) or the – (in screening off), this information feeds back and shapes the internal representations, such that it can now better expect what will happen when B alone is presented. **c)** These updated predictions are not made accurately when either the top-down weight scale is reduced, or the strength of the temporal context is reduced, thus demonstrating the importance of these parameters. Network performance is unaffected by similar increases in these parameters, as a control condition (not shown).

blocking conditions are illustrated in Figure 2, along with the results from the network in these cases, in comparison to those from 4-year-olds (*15*). These results show that the network was able to generalize its learning to these novel blicket test cases, in the same way that people do. Specifically, when AB+ (+ indicates blicket detector going off) is followed by A– (– is no blicket activation), then the model correctly predicts that B alone will activate it (screening off). Conversely, the backwards blocking case of AB+, A+, leads to a somewhat weaker inference that B alone is less likely to activate the detector. To test for the importance of both bidirectional processing and integration of information over time, we modified these parameters in our model. As shown in Figure 2c, decreasing the top-down connections from effect back to internal representation, or reducing the strength of the TI context inputs, significantly impaired performance on both the screening off and backwards blocking conditions. Importantly, these results held whether the parameters were changed at the time of initial learning, or only at test, and only reductions, not comparable increases, led to these effects.

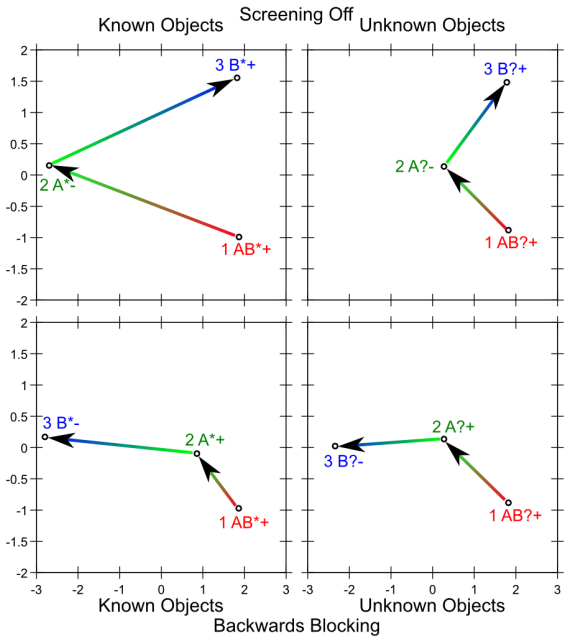

Figure 3: Trajectories of the minus (prediction) state of the internal representation layer in the model over the screening off and backwards blocking sequences, using PCA to reduce to a 2 dimensional projection along the principal eigenvectors. The pattern for known objects (where the input explicitly represents their causal status, indicated by a * in the labels) is shown on the left, for comparison against the same conditions with unknown objects on the right (where causal status must be inferred through feedback from Effect outcomes, indicated by a ?). All cases start off in the same part of representational space for the AB inputs, but for the screening off case (top row), the network represents the unknown A? (top right) in the same way it does for the known A*+ (bottom left). After it experiences the – outcome for A, the representation of B is systematically shifted to coincide with the B*+ known case. This is due to top-down feedback from the Effect output shaping the internal representation layer – here we can see the network effectively "thinking": *A didn't activate it, so it must be B*. A similar dynamic occurs in the backwards blocking case, and it is clear that the network assigns very different representations to B based on the prior Effect outcome for A.

We can examine the representations learned in the network to discover how the network accomplishes this systematic form of inference. Figure 3 plots the internal representation activations projected down to a 2 dimensional space through principal components analysis (PCA) on the activation patterns over time. This captures the primary dimensions of variability in these patterns, and shows how the initially ambiguous input patterns come to be represented similarly to causal agents or non-causal objects depending on the pattern of causal outcomes over time. These causal outcomes feed back into the internal representation layer, and thereby shape the evolution of the representation patterns. This is how the recurrent or bidirectional processing feature of our models (and by hypothesis, the brain) plays such a central role in the flexible generalization of causal inferences to novel situations.

Finally, Figure 4 shows the results from our simulation of (*16*), which systematically manipulates the pattern of causal evidence presented to subjects prior to giving them a common ambiguous test probe. They observed systematic, sensible effects of these manipulations, indicating that adult subjects were able to adapt

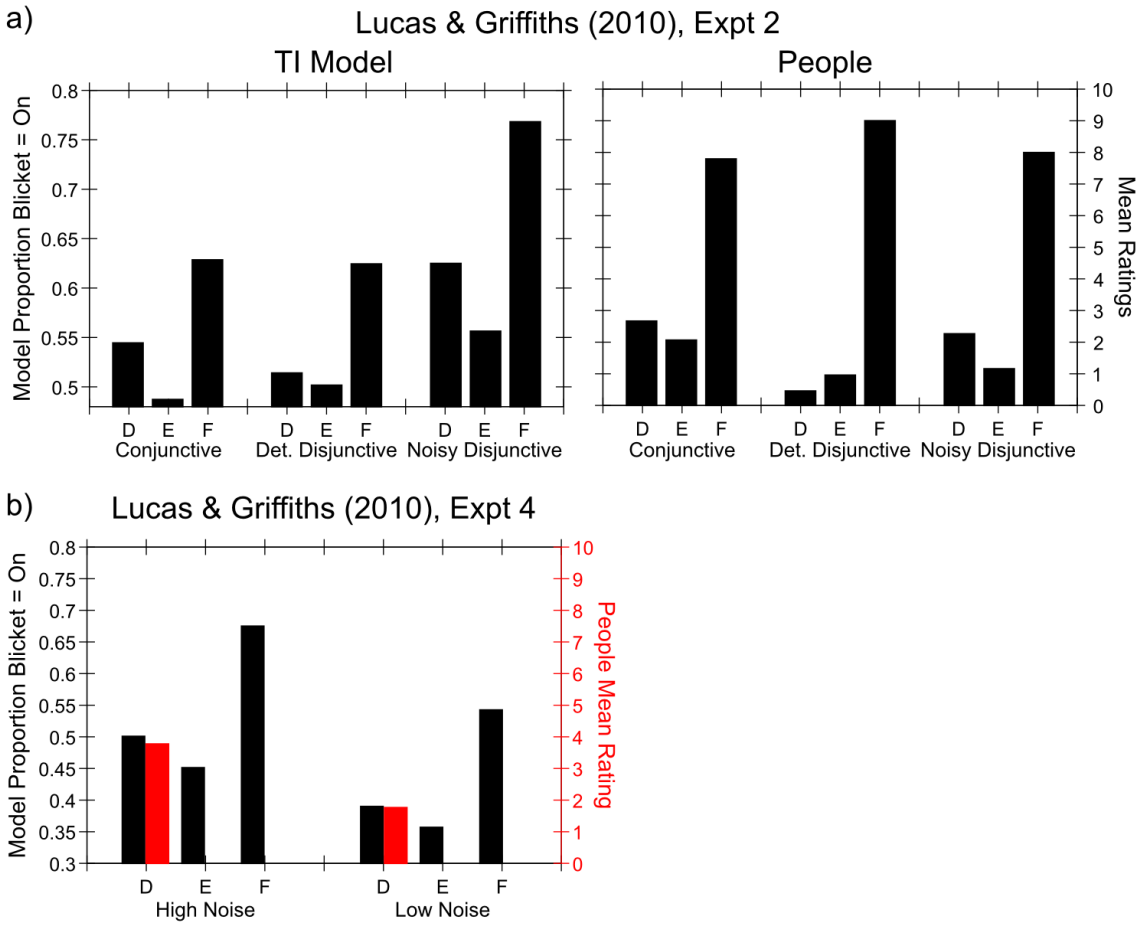

Figure 4: Simulation of Lucas & Griffiths (2010) experiments 2 and 4, showing how different prior training conditions affect responses to a common ambiguous test probe. **a)** Conjunctive: A–, B–, C–, AB–, AC+, BC– (infer that A and C must work together); Deterministic Disjunctive: A+, B–, C–, AB+, AC+, BC– (infer that A acts alone); Noisy Disjunctive: A+, B–, C–, AB–, AC+, BC– (infer that A acts alone, most of the time); Test probe is: D–, D–, D–, E–, DF+, DF+ (F more likely than others, D could be conjunctive or not depending on bias). The model captures the qualitative patterns accurately: F always higher than others, and D higher for conjunctive than disjunctive, only for deterministic case. E is not very relevant, and the noisy disjunctive case is a bit strange and not very diagnostic in people or the model. **b)** Low noise: A+, A+, A+, A–, A+, A+, B–, C–, AB–, AC+, BC–; High noise: A–, A–, A–, A+, A–, A–, B–, C–, AB–, AC+, BC–. Again, model shows that these noise manipulations have appropriate effects on inferences.

their causal expectations rapidly. The very same model we used to simulate the blicket task was also capable of exhibiting this rapid adaptation, based on the same dynamic updating of the internal representation in response to a temporal pattern of observed Effect outcomes. This model was trained on randomly generated versions of the kinds of event sequences used in the experiment, and all tests were done on unknown inputs. Thus, the model was able to extract the patterns associated with weak (conjunctive) and strong (disjunctive) agents, and when the prior exposure biased it into one or the other of these patterns, it generalized this pattern to subsequent ambiguous inputs. Similarly, it was capable of generalizing prior levels of noise to the

novel test case as well.

These impressive forms of generalization can be understood in more familiar terms, if we think in terms of spatiotemporally defined categories learned over temporally extended patterns — just as abstracted categorical representations over static patterns support generalization to novel such input exemplars, so do these spatiotemporal categories. Our model was able to induce these categorical representations from specific patterns over experience, instead of having them defined in advance by the modeler, representing an advance over existing Bayesian models, and suggesting that people could be relying on similar learning mechanisms (see (*8*) for extensive additional biological and behavioral data). Furthermore, as emphasized, this model points to the importance of bidirectional activation dynamics, which are largely missing in most current machine learning and deep network models, for supporting more complex forms of structured relational processing and inference. In future work, we plan to explore how such learning mechanisms could also potentially explain the well-documented failures of human causal reasoning, such as our pervasive bias to infer causality from correlation, and neglect of base rates.

References

1. T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, J. B. Tenenbaum, *Trends in cognitive sciences* **14**, 357 (2010).

2. D. C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber, *Neural Computation* **22**, 3207 (2010).

3. D. Ciresan, U. Meier, J. Schmidhuber, *IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012* pp. 3642–3649 (2012).

4. A. Krizhevsky, I. Sutskever, G. E. Hinton, *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.

5. Y. Bengio, A. Courville, P. Vincent, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1798 (2013).

6. M. Riesenhuber, T. Poggio, *Current Opinion in Neurobiology* **12**, 162 (2002).

7. R. C. O'Reilly, D. Wyatte, S. Herd, B. Mingus, D. J. Jilk, *Frontiers in Psychology* **4** (2013).

8. R. C. O'Reilly, D. Wyatte, J. Rohrlich, *Preprint at: http://arxiv.org/abs/1407.3432* (submitted).

9. R. C. O'Reilly, Y. Munakata, M. J. Frank, T. E. Hazy, Contributors, *Computational Cognitive Neuroscience* (Wiki Book, 1st Edition, URL: `http://ccnbook.colorado.edu`, 2012).

10. R. C. O'Reilly, Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (The MIT Press, Cambridge, MA, 2000).

11. J. L. Elman, *Cognitive Science* **14**, 179 (1990).

12. J. L. Elman, *Machine Learning* **7**, 195 (1991).

13. M. I. Jordan, *Advances in Connectionist Theory: Speech*, J. L. Elman, D. E. Rumelhart, eds. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1989).

14. A. Gopnik, D. M. Sobel, *Child development* **71**, 1205 (2000).

15. D. M. Sobel, J. B. Tenenbaum, A. Gopnik, *Cognitive Science* **28**, 303 (2004).

16. C. G. Lucas, T. L. Griffiths, *Cognitive Science* **34**, 113 (2010).

17. J. L. McClelland, R. M. Thompson, *Developmental Science* **10**, 333 (2007).

18. Supported by: ONR grant N00014-13-1-0067, ONR D00014-12-C-0638, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.