Naval Research Laboratory
Washington, DC 20375-5320

NRL/5510/MR—2023/3

Combining Perception with Structured Knowledge for Rich Causal Reasoning in a Computational Cognitive Architecture

PAUL BELLO

Interactive Systems Section
Information Technology Division

October 6, 2023

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
06-10-2023	NRL Memorandum Report	01-10-2019 – 30-09-22

4. TITLE AND SUBTITLE

Combining Perception with Structured Knowledge for Rich Causal Reasoning in a Computational Cognitive Architecture

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61153N

6. AUTHOR(S)

Paul Bello

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
1P80

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/5510/MR--2023/3

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
One Liberty Center
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR / MONITOR'S ACRONYM(S)

ONR

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report describes an effort to build a computational pipeline that accepts video input of interactions between agents with which (moral) norm violations lead to bad outcomes, extracts and reasons about the associated events, and is capable of recalling events in support of making judgments of blame after the video is processed. The architecture has been designed and extended leaning on the findings of psychological studies conducted as a part of this work.

15. SUBJECT TERMS

Causal reasoning
Cognitive architecture
Blame attribution
Attention

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Paul Bello
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	U	16	**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U			(845) 742-9598

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

This final closeout report documents research done from FY20 through FY22 on a 6.1 NRL base program effort entitled "Combining Perception with Structured Knowledge for Rich Causal Reasoning in a Computational Cognitive Architecture." During the course of the project advances were made along several fronts to be described herein. Firstly, the ARCADIA computational cognitive architecture was extended with significant new capabilities to visually extract, encode, and use relational information present in videos, along with having been endowed with an episodic memory capable of storing hierarchical representations of temporally ordered events and episodes. ARCADIA's existing capabilities for visually verifying causal relations was extended to handle the tricker case of omitted events being recognized as causes. Finally, these capabilities were brought together and married with a simplified approach to counterfactual reasoning in which ARCADIA was able to deliberate about what would have happened in a video sequence had an omitted event (or more than one omitted event) actually happened. Secondly, a psychologically plausible model of the blame attribution process was implemented in ARCADIA allowing the system to visually inspect text strings that described interactions between agents where moral violations were described and to output initial and subsequent updated attributions to blame as more text describing the context of the violation was provided. The model was tested on published human data and matched using the same stimuli with a good fit being obtained. While the intention of this project was to combine this new approach to blame attribution with the previously described ARCADIA extensions, this work is still yet to be finished. Finally, several human subject studies were done looking to investigate how humans represent and reason about causes, including omitted causes interact with norms and other situational constraints. These basic findings were used to inform the computational approach wherever possible. In summary, the results of these three project thrusts provide some initial evidence that it may be possible to have an autonomous system be able to perceptually parse up complex social interactions between agents in situations where norms are violated and to initially apportion and update blame judgments accordingly. While not addressing what might be done with these blame judgments after the fact, it seems quite reasonable to suppose that they would make a different to trust and other dimensions of human-machine or machine-machine teaming.

This page intentionally left blank.

COMBINING PERCEPTION WITH STRUCTURED KNOWLEDGE FOR RICH CAUSAL REASONING IN A COMPUTATIONAL COGNITIVE ARCHITECTURE

INTRODUCTION

Technical Objectives

Of the handful of existing computational models of causal cognition, none deal explicitly with the relationships between perception, memory, language, and reasoning. To do so would require an approach to causal reasoning within the framework of a broader computational approach to human cognition. The project we describe herein is a continuation. As one part of a larger effort funded by a FY17 base program project, a loose amalgam of the ARCADIA cognitive architecture and the mental model theory of human causal reasoning was developed that could observe simple causal interactions in video clips, constructing mental models, and making corresponding causal judgments. Capacity limits on perception and in memory (determined by attention) precluded the system being able to keep large amounts of causal information in mind at a single time, leading to incremental, human-like causal inference. The resulting system exquisitely fit human judgment data, including eye movements. The perceptual demands in these clips were relatively low, and no significant background knowledge about the events in the clips was required to make sense of the interactions. Given these initial successes, the goal of this follow-on effort is to create a system that is robust for more realistic tasks where causal reasoning features centrally, i.e., ones that go beyond simple physical interactions to perceptually dense social interactions. We will focus on the dynamics of blame attribution as an example of a complex social interaction that critically involves inferences about causes in the presence of norms and norm violation and use it as a target for our computational modeling activities. Ideally, we will develop our own human subjects experiments that consist in visual/video stimuli in which we can explore the dynamics of human blame judgments using both the usual set of metrics for text-based stimuli along with additional measures to probe the role of attention over the course of individual trials. Evaluation will consist in our extended version of ARCADIA being a virtual human subject in an experiment on the dynamics of blame attribution using stimuli from an existing human subjects study. Success will be determined by the fit of ARCADIA's judgments over time to those of the human subjects.

Prior to modeling blame, we will construct a variety of video examples in which blame features centrally. We have in mind a simplified over-head view of a stretch of roadway, with police officers (both on-duty, and off-duty) interacting with civilian drivers, some of whom are obeying the law, and some not. We are interested in studying how background knowledge, including norms, drive the causal judgments that ultimately bear on how blame is apportioned. Several fundamental extensions must be made to ARCADIA to support the ongoing parsing of visual and textual input into units that can be subsequently analyzed and used to produce blaming behaviors and to support the complex forms of reasoning about how events might have gone differently (i.e., counterfactual reasoning). We do not propose any formal evaluation for this portion of the project since it is primarily concerned with building up computational foundations for the modeling work on blame attribution.

We also seek to build on the base of human subjects studies initiated in the prior base program effort to specifically investigate the role of situational norms in biasing causal judgment. Finally, we plan to follow up on an important discovery from the prior project on whether causal information is

Manuscript approved October 5, 2023.

mentally represented as being discrete (i.e., caused(x,y), ~caused(x,y)) or continuous (e.g., caused(x,y, 0.75)) or whether both are available but used under different circumstances. Evaluation of our theories will follow the usual standards in psychological science and will be subject to peer review in appropriate journals.

TECHNICAL APPROACH AND PROGRESS AGAINST OBJECTIVES

Preliminaries

This effort combines human subject studies along with computational cognitive modeling to support the building of intelligent systems capable of sophisticated social cognition. The human subjects studies reported on here were primarily conducted using Amazon Mechanical Turk and the Qualtrics platform in collaboration with partners and NRL colleagues at Duke University. With one exception, the computational modeling results reported here were generated using the ARCADIA architecture, under continuous development in Code 5512 since FY15. ARCADIA was initially designed as a computational framework for modeling human attention and to explore its role in all aspects of cognition. Early to midterm work has focused on ARCADIA as a platform for modeling the role of attention in perception (Briggs et al. 2017), including multisensory integration between vision and audition. More recently, ARCADIA research has been focused on the role of attention in cognitive control (Bello & Bridewell 2017), including applications to multitasking (Bridewell et al. 2018) and object tracking (Lovett et al.), along with aspects of planning, deliberation, and higher-level reasoning as reported below.

Computational Modeling: ARCADIA Extensions

Briefly, ARCADIA consists of a set of modules called "components" whose computations are performed in parallel and are influenced and organized by a set of attentional priorities. On each cycle of operation, components individually produce their outputs and one of these is ultimately selected with respect to attentional priorities and subsequently broadcast to all components on the subsequent cycle (Bridewell & Bello 2016). At this point, if any of the components are "focus-responsive" and have means to process the broadcast element from the prior cycle, they do so. The rest of the components continue to operate in their default state. In this way, attentional priorities, which are associated with task/goal representations in ARCADIA, bias system operation from the top-down. The selection operation on each cycle should not necessarily be identified with the act of attending, however. We see attention as the overall effect of various mechanisms operating in concert over short periods of time, rather than being one omnibus algorithm or mechanism within ARCADIA (Lovett et al. 2021).

It should be noted that ARCADIA components are designed for distributed heterogeneous computations. They each implement their own proprietary data structures and algorithms internally which are chosen specifically with respect to the tasks that they perform. For example, some components process structured rules, while others use control-theoretic machinery, and others use algorithms for processing raw sensor data. Every component must implement an interface that allows for cross-component communication. This is done by requiring that all components read from and write to ARCADIA's common representational format called "interlingua." ARCADIA components are either domain-general or task-specific. By domain-general we mean that they are used or usable by every ARCADIA model. Some examples might be the image segmentation, feature-binding, and sensory short-term memory

components in ARCADIA's visual pipeline, as well as components involved in maintaining task representations. Task-specific components might perform task-specific skills, such as learned driving behaviors.

Over the course of this project, we sought to build and test new components in ARCADIA to support the visual parsing of videos where complex interactions between agents would be used to drive blame judgments. One notable characteristic of blame judgments is that they are retrospective. An observing agent will have encoded a memory of the events in question and then use these memories in service of arriving at a judgment of blame. Most, if not all forms of substantive social interaction require a ledger of events be kept in memory. Events themselves are complex (see figure 1), often involving agents, outcomes, and various types of relations between them that may change in certain respects over the course of what we might call an "episode." In this project, we built and tested an initial implementation of episodic memory in the ARCADIA system. The core assumption in our implementation is that attention is necessary to populate an episode with events, agents, relations, and so on, as well as being necessary for determining the boundaries of episodes.

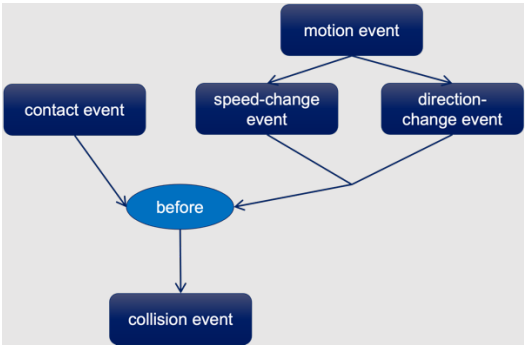

Figure 1: : How a complex event such as an observed collision is represented in episodic memory

Episodes consist of a series of temporally extended moments (see figure 2, shown as black boxes) that are populated by streams of events. Many ongoing events may be happening contemporaneously at a particular moment. Often events are extended in time, motivating our choice of representing events as streams. Streams are populated by the ongoing observation of an event through time and thus required us to equip ARCADIA with functionality for making ongoing judgments about whether an event observed at time t_i was the same event as the currently observed event at time t_j. These equality checks are the basis for extending an event stream. Because all the previously mentioned computations depend on timely observation, there is a danger of dropping an event stream due to attention being occupied by other stimuli. To deal with this issue, we assigned activation to event streams that decays with time but allows episodic memory to smooth out discontinuity due to inattention. Episodes are chunked and put onto an episodic history whenever an event stream is either dropped or added to the set of active event streams for task-related reasons. It should be noted that ARCADIA's current approach to episodic memory was strongly influenced by our colleagues in Code 5515 (Khemlani et al 2015) as part of their broader investigation of human reasoning

about temporal relations. Indeed, the ARCADIA implementation also provides full support for reasoning about the standard set of temporal relations including "before" "after" "during" and "while" through querying episodic history. Notably, ARCADIA's episodic representations do not tag observations with times. Doing so sets up roadblocks for compressing episodic memory representations and leads to inefficient search.

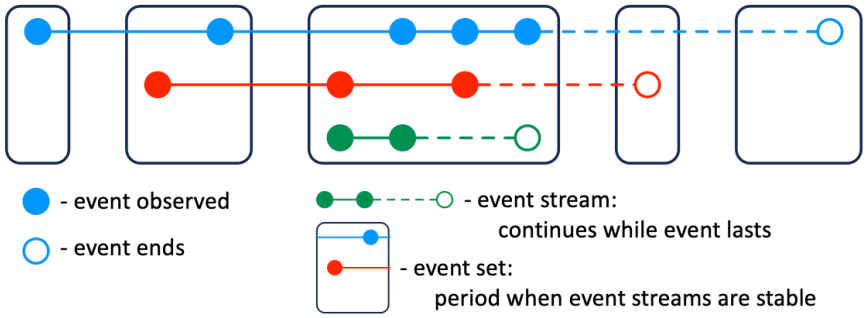

Figure 2: The organization of episodic memory in ARCADIA

The final bit of functionality that was added to our implementation of episodic memory was the ability to query along several dimensions to produce context-specific recall. For the purposes of our investigation, this meant being able to preferentially retrieve episodes containing norm-violation events, violations driven by an overt action (rather than an omission) of an offending agent, and specifically actions that were highly abnormal, statistically speaking. To ground this out a bit, we can think about slightly violating the speed limit while driving as being a relatively normal violation with respect to intentionally running a traffic light in a busy intersection.

In general, the blame process that we seek to model ideally depends on *counterfactual reasoning*, which is reasoning about what could have been the case had some particular fact about the world been different. It is well known that counterfactuals often come to mind as a function of norm violation. With counterfactuals, we attempt to work out in our minds how a more norm-conforming situation might have played out. The connection to blaming behavior is rather obvious here: when we observe an agent violate a norm and cause what looks to be a bad outcome, we might think about how the world might have turned out (better) had that agent done something different. Interrogating the result of this reasoning might well be the basis for assigning degrees of blame. While we did not address counterfactual reasoning during this project given the massive challenge involved, we did incorporate counterfactuals around the edges of the demonstration project that we built for episodic memory.

Our demonstration project involved an overhead traffic scene with five interacting agents. There were two police officers, one civilian car who was speeding, another who was obeying the traffic laws and another who rapidly approaches the latter car on a collision course. We wrote a set of simple rules that applied to agents depending on their status as a civilian or an officer. Civilians were prohibited from speeding, tailgating, and disobeying the instructions of an officer. Officers were obligated to pull over speeders, prevent collisions, respond to any calls for backup, and to complete any

citations in progress before leaving the scene of a crime. ARCADIA, extended with episodic memory, was shown a video sequence in which an officer pulls over a speeder on the bottom portion of the screen, calls a second officer on the top portion of the screen for backup and begins writing a citation. The second officer fails to respond, an omission considering the norms. In the meantime, the law-abiding civilian car is driving along and is approached at high speed by another speeder on a collision course. The second officer remains in place, failing also to pull over this new speeder. The first officer responds to the more urgent collision possibility, leaving the scene of the crime and violating the norm to complete all citations; however, the original speeder drives away in the meantime, escaping a speeding ticket – yet another violation. ARCADIA employed its capability to extract dynamic relations from video data to populate episodic memory with temporally ordered events, including norm violations. This was possible because norms set up expectations for what should or shouldn't happen in particular situations. Comparing the events extracted in perception to these expectations allowed inferences to be made about norm violations and stored episodically along with more pedestrian episodic information about agents, objects, and visible relations.

Once encoded in episodic memory after the video completed, ARCADIA was able to query episodic memory for highly available counterfactuals based on the statistical likelihood of a particular norm being violated (e.g., speeding/high, leaving a crime scene/low), episodes containing norm violations versus those that didn't, and so on. Indeed, episodic memory returned the final part of the sequence in which the first officer and the first speeder respectively left the scene of the crime, with episodes containing the doubly negligent second officer's omissions coming in behind in terms of counterfactuals generated.

Computational Modeling: Blame Judgments

Significant progress was made towards developing a computational cognitive model of norm-guided blaming in the ARCADIA framework. We took Bertram Malle's *Path Model of Blame* (Monroe & Malle 2019) as a guide for our implementation.

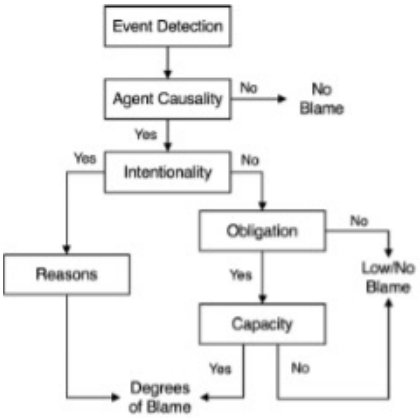

Figure 3: The path model of blame: blame-related concepts and information processing pathways.

The path model is shown above in figure 3 and details empirically validated pathways by which blamers receive and process information in settings where a violation has occurred. There are two major pathways in the process of blame assignment that diverge at a judgment of whether an agent intentionally caused the bad outcome in question. If the action was performed intentionally, blamers will expect reasons to be given, and the quality of these reasons will impact the degree to which blame is assigned. If it is determined that the outcome wasn't intended by the offending agent, a more complex search for information ensues. First, the blaming agent ascertains whether the offending agent had an obligation to prevent the bad outcome in question and if it is determined that no such obligation was in force, little to no blame is assigned. If, however there was an obligation to prevent the bad outcome, the blaming agent attempts to determine whether the offending agent had the capacity (understood broadly) to prevent the bad event from happening. If the offending agent didn't have the capacity to prevent, little to no blame is assigned, but if they did, then an appropriate degree of blame is assigned. We built a set of components and attentional strategies in the ARCADIA system to take perceptual descriptions of vignettes where violations occur. In these vignettes, pertinent information that drive movement down the various blame pathways is presented sequentially to ARCADIA such that system makes initial judgments that are revised as new information about causation, intentionality, reasons, obligation, and capacity becomes available over time. A partial sketch of the overall model is shown below.

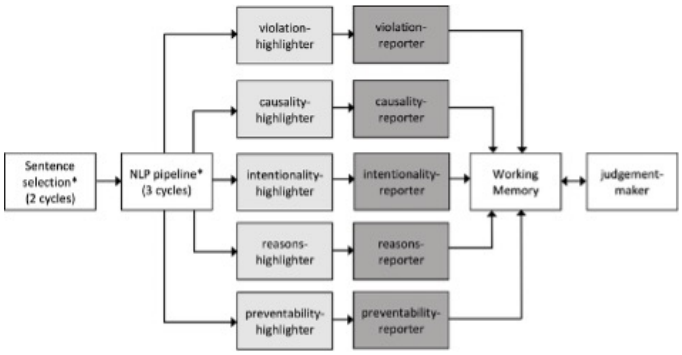

Figure 4: A partial selection of ARCADIA's blame-assignment model.

Missing from the left-hand side of the model-sketch is a substantial number of components for visual perception and the assignment of semantic features to perceived objects. In the implemented model, all stimuli were visual images of sentences that were perceived and subsequently "read" by a natural language processing capability in ARCADIA. In the third layer, several "highlighter" components that request attentional focus be shifted to particular words or pairings of words in the input sentence. For example, if given "He wanted to prevent George from damaging Molly's car" as an input sentence, the reason-highlighter picks up on the fragment "wanted to prevent" and outputs an impression. Notably, an evaluation of the reason in question hasn't occurred yet. The fourth layer in the figure above is fed information from outputs of the highlighter components and perform task-specific operations to produce component judgments that will be stored in working memory and integrated over time to yield a

final blame assignment. In the case of the sentence given above, the reasons-reporter in the fourth layer has access to semantic information in the form of rules about cars being damaged and judges that George has good reasons for doing whatever prima facie bad action he did that subsequently initiated the whole blame attribution process.

Finally, it is worth mentioning that the ordering of steps in the path model of blame is realized in ARCADIA through the dynamic modulation of attentional priorities. We treat the blame attribution process as a form of information foraging. ARCADIA starts at the top, looking for violations or bad outcomes. Once identified, attentional priorities are modified to search for information indicating an agent who is causally connected to the outcome. Once found and stored in working memory, the process of reprioritization continues until a final blame judgment is generated. As an evaluation of the model, we used available stimuli and human judgment data in (Monroe & Malle 2019) to give quantitative shape to the judgments produced by ARCADIA, being only concerned with the stepwise dynamics of the blame process as more information became available to the system within each trial. Our overall results were excellent and can be seen in the graphs below.

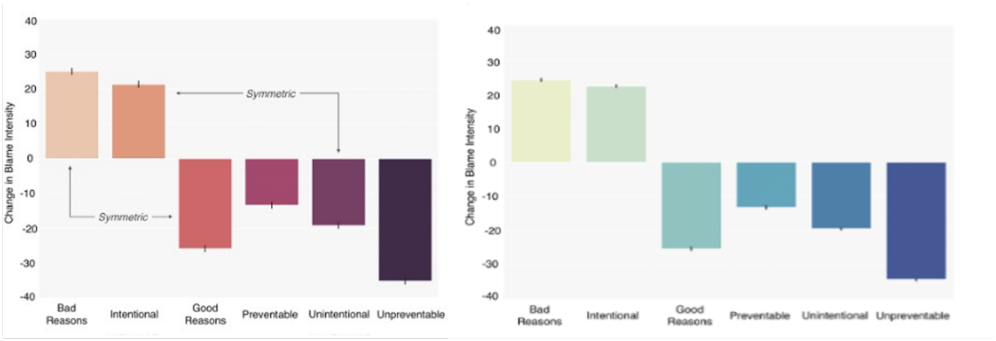

Figure 5: Human blame judgments for stimuli given in (Monroe & Malle 2019) on the left and on the right, blame judgments produced by ARCADIA given identical stimuli.

Further research is needed to advance the model's representation of norms and background knowledge, which, in turn, will support more sophisticated perception of and reasoning about blame-relevant information, especially in the case of reasoning about multimodal stimuli (e.g., audio, video). To this end, we have been exploring multiple open scientific questions including:

1. How do humans internally represent the norms that frame initial judgments in the path model?
2. How do humans detect norm violations in an environment?
3. How do norms guide attention in the process of blaming?
4. How people identify whether sufficient information is available for a particular blame concept (e.g., intentionality, reasons for acting):

- Is there a waiting threshold involved?
- Is partial or low-confidence information sufficient for traversing between parts of the path model?

5. In the case that people determine that there is not sufficient information:
   - Do they make a plan to find the information?
   - Do they move on to the next blame concept, leaving some sort of placeholder concept/object in their mental representation of the scenario?

Notably, all the information needed to make the blame judgments in the Monroe study was given in short bursts and was kept in ARCADIA's working memory. While we made substantial progress on developing functionality for episodic memory in ARCADIA, we didn't have occasion to use it in the context of blame attribution. Partially, this was due to the nature of the Monroe et al stimuli. Our initial plan was to generate a canonical set of episodic stimuli that would have required ARCADIA to visually process a stream of events similar to the traffic scenario mentioned earlier. Unfortunately, running short on remaining time in FY22 and having a major loss of personnel in the final year of the effort made it infeasible. While admittedly very preliminary, we have taken some steps toward addressing an underserved issue in human machine teaming, and one that will undoubtedly arise once autonomous systems become taskable in natural language and routinely engage with the same human confederate over a series of interactions. Much of the recent focus on explainable systems within the DoD research community is laudable, but if those explanations are disconnected from socio-moral practices such as blame attribution, they will be missing vital connections to team-centric learning and the development of trust.

The Psychology of Causal Representation, Reasoning, and Judgment

From a theory perspective, we have conducted a wide-ranging meta-analysis of the literature on so-called "selection effects" in causal judgments which concern how, out of all the possible contributing causal factors for an event, we humans manage to select a singular cause. The objective wasn't to wade into the debates on the various drivers for causal selection, but rather to look at whether those studies might tell us something about whether we mentally represent causes in a binary way or whether our representations are inherently probabilistic. Probabilistic models of causal judgment have risen to prominence in recent years, but the results of our analysis suggest that at the very least, causal judgments appear to be primarily binary, with confidence judgments about candidate causes predicting when they shift from binary to being graded. Our analysis seems to point out methodological flaws in a non-trivial number of published studies that suppose causal judgments to always be graded. The results of this work were submitted to the journal *Cognition* and was accepted (O'Neill et al. 2022).

A second line of empirical work conducted during the period of performance involved extending and refining a psychological theory of how humans represent and reason about so-called omissive causes developed in a prior base program project, and recently published (Khemlani et al. 2021). Omissive causes are challenging for both psychology and for artificial intelligence for a variety of reasons. First, it isn't immediately clear how humans mentally represent absences or non-events and further how we manage to focus in on non-events to represent, since there are potentially

infinitely many choices. Secondly, from the perspective of artificial intelligence, detecting an omission or constructing one from an episodic trace of behavior is similarly challenging. One way to narrow the space of possible omissions to look out for is through their interaction with norms – rules that dictate what agents should do, are permitted to do, and are forbidden from doing (Henne et al. 2021). These normative categories can often mark particular actions and outcomes as important, but they can also mark classes of actions and events as important as well. In either case, norms help to narrow how we look for omissions along with interacting with background knowledge for reasoning about norm violations. This latter feature was especially of concern to the generation and elaboration of counterfactuals, mentioned earlier in the discussion of episodic memory. Under this scheme omitted events can be counterfactually replaced by their non-omitted counterpart, with mental simulation working out the consequences of how things might have turned out better (or worse). This capability is central to complex socio-moral practices such as blaming, and we would also argue is critical for complex human-like one-shot learning that goes beyond the data-driven approaches typical of contemporary machine learning research.

CONCLUSIONS

Substantial progress was made on both empirical and computational fronts toward capturing complex socio-cognitive and particularly socio-moral human practices for eventual application to human-machine teaming. Admittedly we didn't get as far as we would have liked with integrating the three threads of this project, but we believe that enough progress has been made on these fronts to push our efforts forward into the future.

REFERENCES

1. Lovett, A., Bridewell, W., & Bello, P. (2021). Selection, engagement, & enhancement: a framework for modeling visual attention. *Proceedings of 43rd Annual Meeting of the Cognitive Science Society*, 1893–1899. Vienna, Austria.

2. Lovett, A., Bridewell, W., Bello, P. (2019). Selection enables enhancement: an integrated model of object tracking. *Journal of Vision*, 19(14):23.

3. Monroe, A., Malle, B. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, 116(2):215.

4. Bridewell, W., Wasylyshyn, C., Bello, P. (2018). Towards an attention-driven model of task switching. *Advances in Cognitive Systems*, 6, 85–100.

5. Bello, P., Bridewell, W. (2017). There is no agency without attention. *AI Magazine*, 38(4), 27–33.

6. Briggs, G., Bridewell, W., Bello, P. (2017). A computational model of the role of attention in subitizing. *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society*, 1672–1677. London, UK.

7. Bridewell, W., Bello, P. (2016). A theory of attention of cognitive systems. *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*, 3. Evanston, IL.

PUBLICATIONS

Journal/Book

1. O'Neill, K., Henne, P., Bello, P. Pearson, J. & De Brigard, F. (2022). Confidence and gradation in causal judgments. *Cognition*. 223.

2. Bello, P., Malle, B. (in press). Computational approaches to morality. In R. Sun (ed), Cambridge Handbook of Computational Psychology, Cambridge University Press.

3. Henne, P., O'Neill, K., Bello, P., & Khemlani, S. (2021). Norms affect prospective causal judgments. *Cognitive Science*, 44, e12931.

4. Khemlani, S., Bello, P., Briggs, G., Harner, H. & Wasylyshyn, C. (2021). Much ado about nothing: The mental representation of omissive relations. *Frontiers in Psychology*, 11.

Conference

1. LeBlanc, E. (2021). Toward a model of the dynamics of norm-guided blaming. Paper presented at *Qualitative Reasoning 2021*.

2. O'Neill, K., Henne, P., Bello, P., Pearson, J., De Brigard, F. (2021) Confidence effects on causal judgment. Abstract/Poster presented at the 62nd Annual Meeting of the Psychonomics Society.

3. O'Neill, K., Henne, P., Bello, P., Pearson, J., De Brigard, F. (2021) Degrading causation. Abstract/Poster presented at the 47th Annual Meeting of the Society for the Philosophy of Psychology.

4. O'Neill, K., Henne, P., Bello, P., Pearson, J., De Brigard, F. (2021). "Confidence effects on causal judgment". *Psychonomics*.

NAVAL NEED

We have identified several applications of interest to the Navy/NRL that potentially can benefit from work on computationally modeling representations of causality, causal reasoning, and associated capacities for social/moral judgment such as blame attribution:

1. Autonomous systems that must interact with humans and other autonomous agents as teammates.
2. A capability for machine perception guided by high-level knowledge of norms and causal relationships between entities and events in a scene to facilitate decision-making.
3. A framework for dynamically generating attributions of blame to teammates who violate norms and expectations, and a way to use this information to guide future interactions (both positive and negative) with the offender.

These applications share the research challenge of needing to understand and represent the relationships between causes and norms for enhancing the capability of autonomous systems to engage in team behavior.