SANDIA REPORT
SAND2022-4632
Printed April 2022

Sandia National Laboratories

Instantiation of HCML Demonstrating Bayesian Predictive Modeling for Attentional Control

Julie Bugg, Joshua Clifford, Nicole Murchison*, Christina Ting

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods

ABSTRACT

The research team developed models of Attentional Control (AC) that are unique to existing modeling approaches in the literature. The goal was to enable the research team to (1) make predictions about AC and human performance in real-world scenarios and (2) to make predictions about individual characteristics based on human data. First, the team developed a proof-of-concept approach for representing an experimental design and human subjects data in a Bayesian model, then demonstrated an ability to draw inferences about conditions of interest relevant to real-world scenarios. Ultimately, this effort was successful, and we were able to make reasonable (meaning supported by behavioral data) inferences about conditions of interest to develop a risk model for AC (where risk is defined as a mismatch between AC and attentional demand). The team additionally defined a path forward for a human-constrained machine learning (HCML) approach to make predictions about an individual's state based on performance data. The effort represents a successful first step in both modeling efforts and serves as a basis for future work activities. Numerous opportunities for future work have been defined.

Key Words: Bayesian Inferential Modeling, Human Constrained Machine Learning, Attentional Control

CONTENTS

Nomenclature 7

1. Introduction 9
   1.1. A Use Case for Attentional Control 9
   1.2. Instantiation of HCML with Bayesian Modeling 9

2. Attentional Control 11
   2.1. Theoretical Background on Attentional Control 11
   2.2. Background on Pre-Cueing Attentional Control Demands from [10] Experiments 1 and 2 13
   2.3. Relevant Models of Attentional Control 15
      2.3.1. Conflict-Monitoring Model 15
      2.3.2. Applying Conflict-Monitoring Model to Bugg et al. (2015) 15

3. Relevant Models of Attentional Control 16
   3.1. Conflict-Monitoring Model 16
   3.2. Applying Conflict-Monitoring Model to Bugg et al. (2015) 17

4. Data 18
   4.1. Data Preparation 18

5. Bayesian Modeling Approach and Results 18
   5.1. Model Specification 18
   5.2. Analysis Approach 19
   5.3. Model Results 20
   5.4. Reference Prior Model (Experiment One) 20
      5.4.1. Informative Prior Construction 21
      5.4.2. Informative Prior Model (Experiment Two) 21
      5.4.3. Inferential Power of Bayesian Approach 23
   5.5. Discussion 24

6. Machine Learning Model 24
   6.1. Data 24
   6.2. Method 25
      6.2.1. Data preprocessing and feature engineering 25
      6.2.2. Model selection and validation 25
   6.3. Results 26
   6.4. Discussion 26

7. Concluding Remarks 28

8. Future Work 29
   8.1. Future Direction 1: Modifying Botvinick et al. model to account for intentional, goal-directed AC 29

8.2. Future Direction 2: Uncovering individual and contextual determinants of intentional goal-driven AC 31
8.3. Future Direction 3: Modeling stimulus-driven AC 32
8.4. Future Direction 4: Weighting goal-directed and stimulus-driven AC 35
8.5. Future Direction 5: Human Constrained ML Approach to Differentiate Individual Characteristics 35
8.6. Future Direction 6: Building upon Bayesian Inferential Modeling Approach 36

References 37

Appendix A. Appendix: Bugg (2015) Codebook 40
A.1. Experiment 1 40
A.2. Experiment 2 41

LIST OF FIGURES

Figure 1-1. Risk Model for AC. 10
Figure 2-1. Pre-cued lists paradigm [10] 14
Figure 5-1. Posterior distributions for reference prior model beta parameters. 20
Figure 5-2. Posterior distributions for reference prior model beta parameters compared to corresponding informative priors. 21
Figure 5-3. Posterior distributions for reference prior model beta parameters. 22
Figure 5-4. Posterior parameter distribution comparisons for experiment two data models. . 23
Figure 5-5. Posterior distributions for $\beta_1 + \beta_3$, showing expected change in mostly incongruent list Stroop effect when adding a cue. Both models are shown with the posterior mean in red and the 95% probability interval in blue. 24
Figure A-1. Trial names for cued and uncued conditions. 42

LIST OF TABLES

Table 5-1. Parameter estimates for reference prior model (experiment one). 20
Table 5-2. Parameter estimates for informative prior model (experiment two). 22
Table 5-3. Parameter estimates for informative prior model (experiment two). 23
Table 6-1. Accuracy (mean and standard deviation in parentheses), together with feature importance obtained from the weights of the SVM, using the different feature representations. 27
Table 6-2. Confusion matrices for predicting cue condition using response time (with color, right). 27
Table 6-3. Confusion matrices for predicting cue condition using Stroop effect (with color, right). 27

NOMENCLATURE

AC Attentional Control

CSPC Context-Specific Proportion Congruence

HCML Human Constrained Machine Learning

ISPC Item-Specific Proportion Congruence

LWPC List-Wide Proportion Congruence

MC Mostly Congruent

MI Mostly Incongruent

ML Machine Learning

RT Reaction Time

SVM Support Vector Machine

1. INTRODUCTION

1.1. A Use Case for Attentional Control

Attentional control (AC) refers to a set of processes thought to dynamically adjust attention (e.g., direct, correct, and redirect attention) in a goal-oriented and context-sensitive fashion (Braem et al., 2019). Successful engagement of AC refers to attending to goal-relevant information while ignoring or minimizing the influence of goal-irrelevant information; conversely, failure to engage AC means that relevant information was not adequately attended and/or goal-irrelevant information was not successfully ignored. For purposes of this project, the use case of interest concerns how a security guard responds to the attentional demands associated with their position, particularly how they accomplish goals that involve selectively processing some aspects of their environment (i.e., targets; relevant information) while ignoring other aspects (i.e., distractors; irrelevant information).

Consider a security guard who is tasked with checking the paperwork provided by incoming vehicles at a security gate. A high attentional control setting might be one in which the guard is attending to the paperwork (relevant information) and not the information the occupant of the vehicle is communicating (a form of distraction/irrelevant information in this example), which might conflict with the paperwork. As another example, consider a security guard who is monitoring the external environment for security breaches. A high attentional control setting might be one in which the guard is attending to and checking for disturbances in the areas of interest (near the gate or other pre-determined boundaries of importance, i.e., relevant information) and not attending to activity (e.g., a person running; a vehicle driving by) that may capture their attention or divert their attention to areas outside the areas of interest (i.e., irrelevant information). A key goal of this project is to conceptualize a risk model (where risk is represented in Figure 1-1 as mismatches between attentional demand and attentional control) to determine factors that would make a security guard vulnerable to AC failure (or conversely, determine factors that would predict successful engagement of AC by the security guard) and quantify a model of such risk.

In terms of success with attentional control (i.e., attention is not diverted to irrelevant information), of greatest interest is the specific match highlighted by the green shape in the lower right cell (see Figure 1-1). This represents a low-risk scenario in that the AC setting is appropriately matched to AC demand (i.e., the guard is in an attentionally demanding situation, and the guard is engaging a high level of AC). In this case, the risk of AC failure is low and thus a high level of performance should be achieved.

1.2. Instantiation of HCML with Bayesian Modeling

A key goal of this project is to conceptualize a risk model and within this model, identify factors that would heighten risk of a security guard's AC failure (or conversely, identify factors that would predict successful engagement of AC by the security guard). The conceptual risk model is illustrated below. The model considers two factors that influence risk. One is the AC demand of

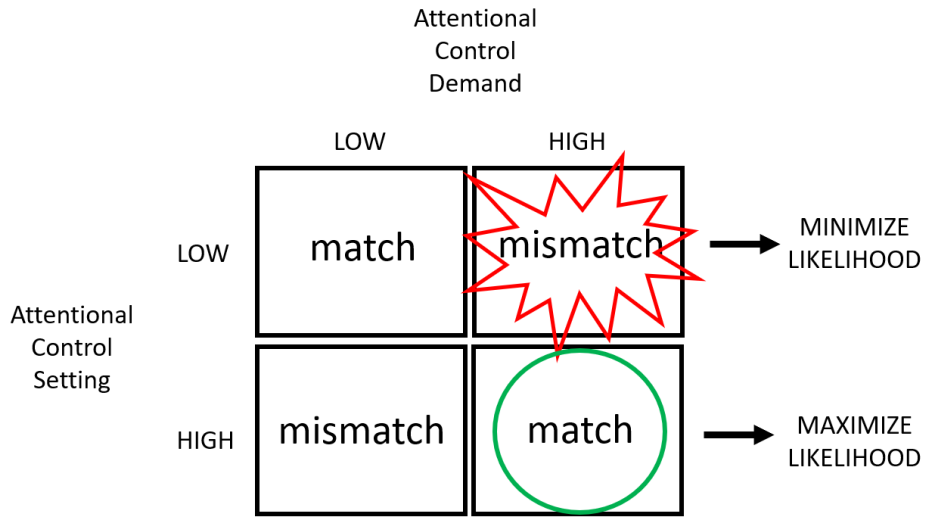

Figure 1-1 Risk Model for AC.

the current situation (e.g., low——high) and the second is the individual's AC setting (e.g., low——high), which refers to the degree of processing selectivity engaged by the individual (i.e., degree to which goal-relevant information is selectively attended and irrelevant information is ignored). The red shape in the upper right cell signals a high-risk scenario where there is a critical mismatch between AC demand and AC setting. Specifically, the AC demand is high, but the AC setting is low and thus there is a high risk that performance will be slow and error prone. The opposite mismatch (low AC demand, high AC setting, as in the lower left cell) is also not optimal, but in real-world situations it is almost certainly less risky with respect to performance outcomes in that the primary risk (cost) is the unnecessary heightening of AC/expenditure of resources when not needed.

From an applied perspective, the goal is to minimize the likelihood that the critical mismatch will occur (resulting in a high risk of AC failure) and maximize the likelihood that the critical match will occur (resulting in a low risk of AC failure). Figure 1-1 provides a risk model for conditions where the critical mismatch will occur. The lower left-hand quadrant represents a scenario where an individual has low attentional demand, but high attentional control. In this case, one may encounter low-demand stimuli but are at risk of not relaxing control when responding to them. In contrast, the upper right-hand quadrant represents a scenario where an individual has high attentional demand, but low attentional control, so may encounter a scenario where they need to heighten attentional control to respond to high-demand stimuli. This represents the highest risk for missed or delayed response in real-world settings.

Two modeling techniques will be used to evaluate data from experimental studies of AC [10]. First, we will use Bayesian predictive modeling to model questions related to our risk matrix to predict the probability of specific scenarios we would be concerned with in real-world settings, for example: (1) What is the probability that the Stroop effect will increase when in a low-demand context compared to a high-demand context? (2) What is the probability that the Stroop effect will increase in a low-demand context if the subject is told to anticipate low-demand stimuli? (3)

What is the probability that the Stroop effect will decrease in a high-demand context if the subject is told to anticipate high-demand stimuli? [Note: Questions 2 and 3 addressed in Section 5].

Second, we will use machine learning to determine if we can predict a subjects' AC state, namely if they are in a cued or uncued state. Knowing that being in a cued state in a low-demand context (when expecting low demand trials), results in a greater Stroop effect (i.e., worse performance) means that the ability to predict a cued vs. uncued state can allow us to anticipate the risk of AC failure for a given subject at a given point in time. Similarly, if being in a cued state in a high-demand context were to result in a smaller Stroop effect (i.e., better performance), then we could anticipate the likelihood of AC success for a given subject at a given point in time based on their state.

2. ATTENTIONAL CONTROL

2.1. Theoretical Background on Attentional Control

Researchers have distinguished between goal-directed AC and stimulus-driven AC, and this distinction is important when considering factors that predict how successful people will be in engaging AC (e.g., [5, 7]. In this section, we will first define and differentiate the two types of AC. Then we will briefly detail several factors that predict AC performance for each type.

Goal-directed AC is control that is mediated by top-down biasing of attention based on one's goals. While not necessarily intentional (i.e., one could theoretically activate goals implicitly and these goals may bias attention toward relevant information), for purposes of this project we are especially interested in goal-directed AC that is intentional. For example, consider again the security guard in the preceding section. Let's assume that through intelligence, Sandia has learned that there is a high risk of a security breach at Gate 10. Assume further that this information is communicated to the security guard. If the security guard uses this information to intentionally heighten AC (i.e., to direct attention to goal-relevant information including any activities occurring near Gate 10 and ignore irrelevant information such as activities occurring outside this target area), then this example would illustrate an intentional form of goal-directed AC.

In contrast, stimulus-driven AC is control that is governed by external cues that have become associated with attentional demands. These cues can later trigger adjustments in attention (e.g., focusing more on relevant information), with such adjustments often occurring outside of people's awareness (e.g., [17]). That is, unlike goal-driven AC, and especially intentional forms of goal-driven AC, stimulus-driven AC does not involve the individual actively holding in mind one's goals or intentionally varying how heightened (or relaxed) their AC is. An example of stimulus-driven AC as applied to the use case of security is as follows. Assume that through experience over time, a security guard mostly encounters security threats in the northeast part of the campus and very infrequently encounters security threats in the northwest part of the campus. Accordingly, the security guard's experience in the northeast part of campus mostly entails higher AC engagement and their experience in the northwest mostly entails lower AC engagement. In other words, over time a high level of AC becomes associated with the northeast location whereas a low level of AC becomes associated with the northwest location. Consequently, when the

security guard encounters a person in the northeast location, their AC is heightened. Theoretically, this is thought to occur because the location serves as a cue that retrieves the associated AC level (in this case, high AC; see e.g., [16]). Some evidence suggests stimulus-driven retrieval of associated levels of AC may become automatized, such that even under concurrently demanding situations (when attention is devoted to another task while performing security related tasks), cues can effectively heighten AC [26].

Returning to the question of how successful a person (e.g., a security guard) will be in situations that demand high AC, different factors are predictive of goal-directed AC and stimulus-driven AC. On the goal-directed AC side, and perhaps especially for the intentional form of goal-directed AC, primary predictors would be activation and/or maintenance (persistent activation) of the requisite goal, concurrent demands (e.g., is the person concurrently performing other cognitive tasks while attempting to engage AC), cognitive effort, and motivation. High levels of activation/maintenance, low concurrent demands, high levels of effort/low levels of effort avoidance, and high levels of motivation predict success in engaging goal-directed AC. Conversely, low levels of activation/maintenance, high concurrent demands, low levels of effort/high levels of effort avoidance, and low levels of motivation failure in engaging goal-directed AC. Put simply, people are more likely to successfully engage goal-directed AC when needed if they a) have activated the requisite goal (e.g., security guard activates the goal to direct attention to goal-relevant information including any activities occurring near Gate 10 and ignore irrelevant information such as activities occurring outside this target area), b) are not engaging in another high demand cognitive task (e.g., the security guard is only focused on attending to relevant information and is not also simultaneously responding to work emails), c) are willing to expend the effort to heighten AC (e.g., the security guard does not find cognitive effort to be aversive), and d) are motivated to accomplish the goal.

On the stimulus-driven AC side, primary predictors are factors such as learning (e.g., having sufficient prior experience associating cues with AC demands), paying attention to cues that are predictive of demand (attention to cues can also be influenced by the quality of the cue independent of attention; for example, if cues are unintelligible/imperceptible/otherwise poor, attention to cues is likely to be low), and cue similarity. High levels of learning, high attention to cues, and high cue similarity predict successful stimulus-driven AC. Low levels of learning, low attention to cues, and low cue similarity predict failed stimulus-driven AC. Put simply, environmental cues are more likely to successfully trigger a heightening of attentional control when needed if a) participants have successfully learned associations between cues and AC demands (e.g., security guard has had plenty of prior experiences in the northeast and northwest areas to learn attentional demands tend to be high in the northeast but low in the northwest; again, note that this learning is likely implicit), b) cues are salient or otherwise readily attended (e.g., security guard readily detects occurrence of activity in a specific location), and c) cues are highly similar to those encountered during the learning process.

From here forward, we will focus mainly on goal-directed AC, including experimental manipulations designed to examine how well people heighten (or relax) AC based on expected demands, the ability of extant computational models of AC to account for patterns of AC heightening and relaxation, and our approach to modeling predictors of goal-directed AC.

2.2. Background on Pre-Cueing Attentional Control Demands from [10] Experiments 1 and 2

The primary experimental manipulation that has been used to examine the intentional modulation of goal-directed AC is a pre-cueing manipulation. Pre-cues refer to information that is explicitly provided in advance of performance. For purposes of this project, we will focus on a pre-cueing manipulation developed by Bugg, Diede, Cohen-Shikora, and Selmeczy [10]. The general purpose of the research of [10] was to examine whether people can use advanced knowledge in the form of valid explicit pre-cues to relax AC (intentionally adopt a low AC setting) and most importantly heighten AC (intentionally adopt a high attentional control setting) when needed (when cues signal those AC levels are appropriate). In other words, they aimed to examine the role of intentions (expectations) in goal-directed AC via a cueing manipulation. As in many studies of AC, [10] utilized a Stroop task. In the Stroop task, participants are instructed to name the ink color of a color word while ignoring the word. The color is thus considered goal-relevant (to-be attended/target dimension of the stimulus) and the word is irrelevant (to-be ignored/distractor dimension). A low demand trial is one that is congruent (e.g., word RED in red ink) whereas a high demand trial is one that is incongruent (e.g., word BLUE in red ink). A low AC level (i.e., relatively relaxed AC) is one with low processing selectivity meaning that the relevant dimension is attended as is the irrelevant dimension. A high AC level (i.e., relatively focused AC) is one with high processing selectivity meaning that the relevant dimension is attended to a greater degree than the irrelevant dimension. The Stroop effect (incongruent RT – congruent RT) can be taken as an indicator of AC with larger Stroop effects indicating less successful AC.

The general procedure used by [10] (Experiments 1 and 2) was as follows. After a small number of practice trials with the Stroop task, participants encountered mini blocks comprised of 10 trials each in a pre-cued lists paradigm (see Figure 2-1). Half of the mini blocks (mini lists) were preceded by a valid pre-cue. Participants were told that the upcoming list would be mostly matching which meant mostly congruent, or mostly conflicting which meant mostly incongruent, and they were encouraged to use the pre-cues. The other half of the mini blocks were uncued. For these lists, participants were not told what type of list it would be; instead, question marks appeared on the cue slide. Half of the uncued lists were mostly congruent, and half were mostly incongruent. On each trial within a list, participants named aloud the color and ignored the word as quickly and accurately as possible. The stimulus (color word in a color) appeared on screen and remained until the participant responded. The experimenter coded the spoken response and the next stimulus appeared. After the 10th trial within a mini-block was completed, a break occurred after which the next mini-block was presented. Experiment 2 was the same as Experiment 1 except two other lists were included: cued 50% congruent list (i.e., half of trials were congruent, and half were incongruent) and an uncued 50% congruent list. In sum, there were three key manipulations in the experiments. One was a trial type (congruency) manipulation such that some trials were congruent (e.g., word RED in red ink) and some were incongruent (e.g., word BLUE in red ink). Second, there was a list-wide proportion congruence (LWPC) manipulation—the mini blocks (lists) were either mostly congruent or mostly incongruent (or 50% congruent in Experiment 2). Finally, there was a cueing manipulation with half of the lists being cued and half being uncued.

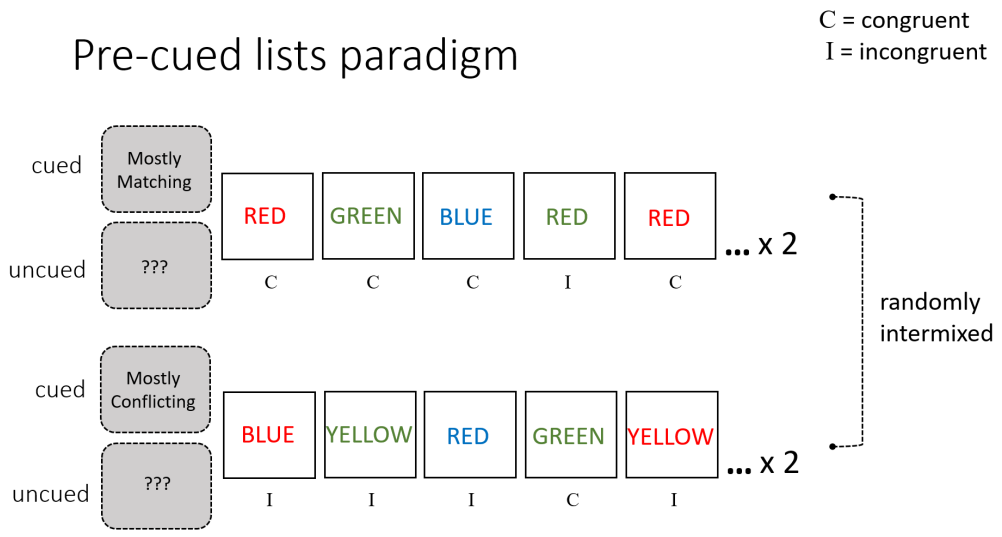

Figure 2-1 Pre-cued lists paradigm [10]

To gauge use of intentional goal-driven AC (i.e., the degree to which participants used the cues to adjust control), [10] compared Stroop effects between cued and uncued lists that shared the same LWPC. For mostly congruent lists, which yield relatively large Stroop effects, the prediction was that the cued lists should yield an even larger Stroop effect than the uncued lists indicating participants relaxed AC. For mostly incongruent lists, which yield relatively small Stroop effects, the prediction was that cued lists should yield an even smaller Stroop effect than the uncued lists indicating participants heightened AC. The key findings were as follows. In both experiments, on average, participants intentionally relaxed AC when cued to do so as indicated by a larger Stroop effect in the cued mostly congruent condition compared with the uncued mostly congruent condition; however, they did not intentionally heighten AC when cued to do so as indicated by a Stroop effect that was equal in size for the cued mostly incongruent condition and uncued mostly incongruent condition. These findings were replicated by [6].

Considering again the use case of security, this experiment informs the question of whether people will heighten AC when they have advance information that advises them to do so, and conversely though less important from an applied perspective, whether they will relax AC when they have advance information that advises them to do so. To the extent one can generalize from the lab to the use case, the findings suggest that supervisors could expect a guard to relax AC if they were told that intelligence suggests a low risk of a security breach at Gate 10, but of concern, it is questionable whether they could rely on the guard to heighten AC if intelligence suggested a high risk of a security breach at Gate 10. This is a key takeaway because it demonstrates that there are limitations to goal-directed AC.

2.3. Relevant Models of Attentional Control

2.3.1. Conflict-Monitoring Model

In addition to studying AC via experiments such as the experiments described in Chapter 1 [10], cognitive psychologists have also studied AC via computational modeling. Computational models enable researchers to study complex systems such as the AC system of the human brain. One model that has been highly successful in simulating key behavioral patterns from the AC literature is Botvinick and colleagues' conflict-monitoring model [3]. The essence of this model is that a conflict monitor (conflict detection module situated in anterior cingulate cortex) tracks information processing conflict (e.g., occurrence of incongruent trials in a Stroop task) and signals to control regions such as lateral prefrontal cortex to increase top-down biasing (goal-directed AC), with adjustments in AC corresponding to the degree of conflict that is detected. This creates a conflict-control loop whereby evaluation of conflict trial-to-trial (technically, an exponentially weighted average of conflict across several preceding trials is calculated) leads to increases (when conflict is relatively high) or decreases (when conflict is relatively low) in AC. The model thus addresses the question of how control processes know when to intervene without the need to refer to a homunculus.

An important aspect of this model is that top-down biasing (the control adjustment) refers to general (pathway-level) adjustments in task representation weights, that is, how much color-naming (attention to the target/relevant color) is weighted relative to word reading (attention to the distracting/irrelevant word). If conflict is detected on trial n – 1 (i.e., an incongruent trial is presented), the idea is that this conflict will lead to a generally greater weighting of the color naming task compared to the word reading task on the next trial such that the response that corresponds to color-naming dominates response activation. By "generally" greater weighting, emphasis is placed on the pathway-level nature of the control adjustment. That is, the adjustment does not occur at the item-level (level of the specific color presented on n – 1) as in other computational models [1].

Botvinick [3] tested the assumptions of their model by simulating behavioral data from several prior studies that demonstrated key AC phenomena. Most relevant for present purposes is their simulation of a LWPC-like effect from [27]. They exposed the model to Stroop trials (inputs of color/word combinations) in three list conditions that varied in their trial type proportions using parameter settings specified in Botvinick et al. Then they determined if the reaction times produced by the model varied in the same way that Tzelgov et al. observed (e.g., faster incongruent trial reaction times in conditions with more incongruent trials overall). The reaction times produced by the model did mirror those of Tzelgov et al., consistent with the notion that the occurrence of incongruent trials elicits a strong conflict signal and leads to a heightening of control (increased color-naming weight in task representation unit).

2.3.2. Applying Conflict-Monitoring Model to Bugg et al. (2015)

Given the results of this simulation, one can infer that the model could successfully simulate patterns of LWPC effects more generally, including performance in the uncued condition of [10].

In other words, in Bugg et al., a LWPC effect was observed whereby the Stroop effect was larger in the uncued mostly congruent condition compared with the uncued mostly incongruent condition, a highly replicable pattern observed widely throughout the literature (see e.g., [6], for review). Consistent with the conflict-monitoring model, the more frequently incongruent trials were experienced (cumulative effects of conflict), the greater AC was heightened based on the cumulative effects of conflict. Conversely, when incongruent trials were rare (as in the mostly congruent condition), control was gradually decreased such that word reading had a larger influence on performance (e.g., slower incongruent RTs). Note that when comparing the cued mostly congruent condition to the cued mostly incongruent condition, a LWPC effect also emerged, and the conflict-monitoring model can accommodate this result via the same mechanism.

A key limitation, however, concerns the most critical behavioral pattern from [10]. Recall that they found an effect of the pre-cue for the mostly congruent condition but not for the mostly incongruent condition. In other words, when comparing the uncued and cued mostly congruent conditions, there was a difference in the magnitude of the Stroop effect with a larger effect observed for the cued condition, consistent with the idea that participants intentionally relaxed goal-directed AC in response to the pre-cue. However, when comparing the uncued and cued mostly incongruent conditions, there was not a difference in the magnitude of the Stroop effect. This suggests participants did not intentionally heighten goal-directed AC in response to the pre-cue. This asymmetry cannot be explained by the conflict-monitoring model (other related models are similarly limited, e.g., [18, 22, 21]. Rather, the conflict-monitoring model predicts equivalent Stroop effects for the cued and uncued mostly congruent conditions, in addition to the cued and uncued mostly incongruent conditions since conflict experience (accumulation of conflict) within cued and uncued lists is equivalent for each of these comparisons.

Indeed, no extant model has attempted to model goal-driven AC in a cueing paradigm where the participant (model) knows in advance the likelihood of conflict in the upcoming list and can thus intentionally adjust AC without having to experience the degree of conflict (monitor for conflict). In the next section, we detail potential approaches to modeling intentional goal-driven AC with the goal of developing an approach that enables us to account for the asymmetry observed by [10].

3. RELEVANT MODELS OF ATTENTIONAL CONTROL

3.1. Conflict-Monitoring Model

In addition to studying AC via experiments such as the experiments described in Chapter 1 [10], cognitive psychologists have also studied AC via computational modeling. Computational models enable researchers to study complex systems such as the AC system of the human brain. One model that has been highly successful in simulating key behavioral patterns from the AC literature is Botvinick and colleagues' conflict-monitoring model [3]. The essence of this model is that a conflict monitor (conflict detection module situated in anterior cingulate cortex) tracks information processing conflict (e.g., occurrence of incongruent trials in a Stroop task) and

signals to control regions such as lateral prefrontal cortex to increase top-down biasing (goal-directed AC), with adjustments in AC corresponding to the degree of conflict that is detected. This creates a conflict-control loop whereby evaluation of conflict trial-to-trial (technically, an exponentially weighted average of conflict across several preceding trials is calculated) leads to increases (when conflict is relatively high) or decreases (when conflict is relatively low) in AC. The model thus addresses the question of how control processes know when to intervene without the need to refer to a homunculus.

An important aspect of this model is that top-down biasing (the control adjustment) refers to general (pathway-level) adjustments in task representation weights, that is, how much color-naming (attention to the target/relevant color) is weighted relative to word reading (attention to the distracting/irrelevant word). If conflict is detected on trial n – 1 (i.e., an incongruent trial is presented), the idea is that this conflict will lead to a generally greater weighting of the color naming task compared to the word reading task on the next trial such that the response that corresponds to color-naming dominates response activation. By "generally" greater weighting, emphasis is placed on the pathway-level nature of the control adjustment. That is, the adjustment does not occur at the item-level (level of the specific color presented on n – 1) as in other computational models [1].

[3] tested the assumptions of their model by simulating behavioral data from several prior studies that demonstrated key AC phenomena. Most relevant for present purposes is their simulation of a LWPC-like effect from [27]. They exposed the model to Stroop trials (inputs of color/word combinations) in three conditions that varied in their trial type proportions using parameter settings specified in Botvinick et al. Then they determined if the reaction times produced by the model varied in the same way that Tzelgov et al. observed (e.g., faster incongruent trial reaction times in conditions with more incongruent trials overall). The reaction times produced by the model did mirror those of Tzelgov et al., consistent with the notion that the occurrence of incongruent trials elicits a strong conflict signal and leads to a heightening of control (increased color-naming weight in task representation unit).

3.2. Applying Conflict-Monitoring Model to Bugg et al. (2015)

Given the results of this simulation, one can infer that the model could successfully simulate patterns of LWPC effects more generally, including performance in the uncued condition of [10]. In other words, in Bugg et al., a LWPC effect was observed whereby the Stroop effect was larger in the uncued mostly congruent condition compared with the uncued mostly incongruent condition, a highly replicable pattern observed widely throughout the literature (see [6] for review). Consistent with the conflict-monitoring model, the more frequently incongruent trials were experienced (cumulative effects of conflict), the greater AC was heightened based on the cumulative effects of conflict. Conversely, when incongruent trials were rare (as in the mostly congruent condition), control was gradually decreased such that word reading had a larger influence on performance (e.g., slower incongruent RTs). Note that when comparing the cued mostly congruent condition to the cued mostly incongruent condition, a LWPC effect also emerged, and the conflict-monitoring model can also accommodate this result. A key limitation, however, concerns the most critical behavioral pattern from [10].

Recall that they found an effect of the pre-cue for the mostly congruent condition but not for the mostly incongruent condition. In other words, when comparing the uncued and cued mostly congruent conditions, there was a difference in the magnitude of the Stroop effect with a larger effect observed for the cued condition, consistent with the idea that participants intentionally relaxed goal-directed AC in response to the pre-cue. However, when comparing the uncued and cued mostly incongruent conditions, there was not a difference in the magnitude of the Stroop effect. This suggests participants did not intentionally heighten goal-directed AC in response to the pre-cue. This asymmetry cannot be explained by the conflict-monitoring model (other related models are similarly limited, e.g., [18, 22, 21]. Rather, the conflict-monitoring model predicts equivalent Stroop effects for the cued and uncued mostly congruent conditions, in addition to the cued and uncued mostly incongruent conditions since conflict experience (accumulation of conflict) within cued and uncued lists is equivalent for each of these comparisons.

Indeed, no extant model has attempted to model goal-driven AC in a cueing paradigm where the participant (model) knows in advance the likelihood of conflict in the upcoming list and can thus intentionally adjust AC without having to experience the degree of conflict (monitor for conflict). In the next section, we detail potential approaches to modeling intentional goal-driven AC with the goal of developing an approach that enables us to account for the asymmetry observed by [10].

4. DATA

4.1. Data Preparation

The models use the data collected in the Bugg et al. (2015) experiments. Trials were filtered as in [10], with practice trials, error trials, and trials with response times faster than 200 ms or slower than 3,000 ms all dropped. In addition, to obtain consistency between the models for the two experiments, 50% congruent list blocks were dropped from the second experiment data. Finally, the data for each experiment was summarized at the list level for each subject, with variables for cue condition (cued/uncued block), list congruence (mostly incongruent/congruent), and Stroop effect for the list (calculated by subtracting the mean RT for congruent trials from the mean RT for incongruent trials).

5. BAYESIAN MODELING APPROACH AND RESULTS

5.1. Model Specification

For this proof of concept, the model is a basic representation of the effects of relevant factors (attention demand level and cue condition) on AC, not necessarily the highest fidelity model of AC (such as the Botvinick model). The intent is to provide a straightforward demonstration of how using psychological research in conjunction with Bayesian methods can inform a predictive model. Linear regression will be used to provide a connection to machine learning methods.

The model predicts the Stroop effect in a block for participants in the experiments using the cueing condition and list congruence. It is worth noting that, since the predictors are categorical variables, the regression model is an alternative framing of an ANOVA model and will provide conclusions similar to those in the original paper, except for block-level Stroop effects rather than individual trial reaction times. An interaction term is included in the model to account for asymmetry in AC change in response to cueing, where participants show evidence of cue use for mostly congruent lists but not mostly incongruent lists. The model takes the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i, \tag{1}$$

where:

- Y_i is the response (Stroop effect on a list)
- X_{i1} is the cue condition (dummy variable, 1 if participant was cued)
- X_{i2} is the list congruence (dummy variable, 1 if the list was mostly incongruent)
- β_0 is the intercept term (predicted response for reference group, uncued mostly congruent).
- β_1 is the parameter for the cue condition (predicted change in response from reference when cue is introduced)
- β_2 is the parameter for the trial type (predicted change in response from reference when list is incongruent)
- β_3 is the interaction term for cue condition and trial type (predicted additional change in response from reference when a cue is present and the list is incongruent)
- ε_i is the error term (assumed independently normally distributed around 0 with unknown variance σ^2)

5.2. Analysis Approach

The general analysis approach will be to use information from the first experiment to inform a model for the second experiment, which is expected to provide increased precision in model parameter estimates (and, as a result, increased predictive power of the model). First, a Bayesian regression model will be estimated using the first experiment data and reference priors that provide no additional information for model parameter estimation. These results will be compared with results from the original ANOVA model to confirm they are consistent. Next, the posterior parameter distributions will be utilized as informative priors for estimating a model on the second experiment data. A sensitivity analysis will be done, checking slight variations of the informative priors to ensure the results are not dramatically changed and, thus, overly dependent on the specific formulation of the informative prior. The informative prior model will be compared to a model using reference priors will be done to understand how using the previous knowledge from experiment one impacted the model for the second experiment. Finally, a demonstration of the additional inferential capabilities allowed by the Bayesian approach will be provided. All models will be estimated using JAGS [24], a Gibbs sampler to estimate Bayesian models using Markov chain Monte Carlo simulation that provides a great deal of flexibility in model specification.

5.3. Model Results

5.4. Reference Prior Model (Experiment One)

The experiment one data reference prior model parameters were estimated with 10,000 sample values from the posterior parameter distributions. Two chains were used, and convergence was assessed through trace plots, Gelman-Rubin diagnostics, and autocorrelation plots (which led to use of a thinning interval of five for the final posterior samples). Additionally, the regression model assumptions were assessed using visual inspection of residual plots (residuals by predicted values for the constant variance and linearity assumptions, residual Q-Q plot for the normality assumption, and residuals by trial number for the independence assumption), with residuals calculated using posterior mean point estimates for model parameters. The Q-Q plot suggested heavy tails, but otherwise the results adequately satisfied model assumptions.

Table 5-1 presents point estimates, standard deviations, and 95% credible or probability intervals for the model parameters, and Figure 5-1 shows the posterior distributions for the β parameters.

Parameter	Estimate	SD	95% PI
β_0	181.6	10.1	(161.8, 200.9)
β_1	41.9	14.3	(14.9, 69.2)
β_2	-100.8	14.2	(-128.7, -73.3)
β_3	-43.1	20.0	(-81.3, -4.1)
σ^2	131.0	3.5	(124.3, 138.0)

Table 5-1 Parameter estimates for reference prior model (experiment one).

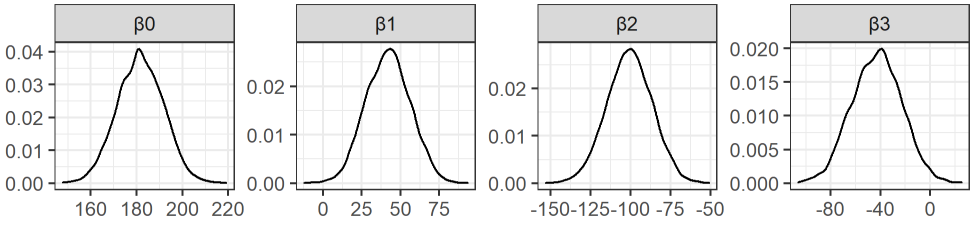

Figure 5-1 Posterior distributions for reference prior model beta parameters.

The Stroop effect estimates for various conditions (cued/uncued, mostly congruent/incongruent) are consistent with expectations from the ANOVA model in the initial paper, so the model is properly specified and these posterior distributions can be used as informative priors.

5.4.1. Informative Prior Construction

The mean and standard deviation for each β model parameter were used to construct informative priors using normal distributions. The specific prior distributions used were:

- $\beta_0 \sim N(181.6, 102)$
- $\beta_1 \sim N(42.1, 204.5)$
- $\beta_2 \sim N(-100.9, 201.6)$
- $\beta_3 \sim N(-43.3, 400)$

This approach technically uses the marginal posteriors to construct informative priors, so it does not take into account potential correlations between parameters. Including the full multivariate posterior distributions from the first model could improve the informative priors and should be explored in future work.

Since the residual variance is treated as unknown in the model, the posterior distributions are technically *t*-distributed. However, the normal distributions are good approximations, as can be seen in Figure 5-2, where the 10,000 posterior distribution samples for the parameters from the first model are compared to 10,000 samples from the corresponding informative priors noted above.

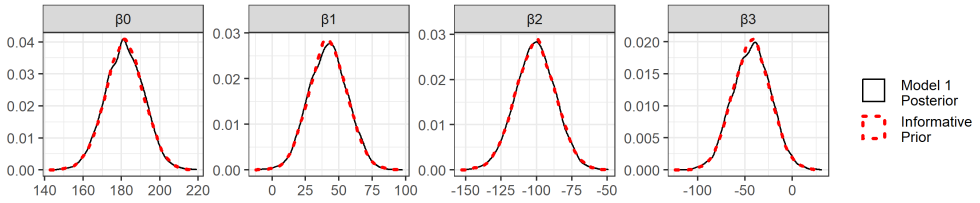

Figure 5-2 Posterior distributions for reference prior model beta parameters compared to corresponding informative priors.

5.4.2. Informative Prior Model (Experiment Two)

The informative prior model using experiment two data was estimated using the same process as the reference prior model for the experiment one data. Additionally, a sensitivity analysis was done with slight variations on the informative priors (changing the mean by ± 10 and the variance by ± 30). Final estimates for β_1, β_2, and β_3 were unaffected, but the β_0 did show slight sensitivity to the prior, with the mean of the posterior distribution changing by 10-20 ms depending on the prior.

Table 5-2 presents point estimates and 95% probability intervals for the model parameters, and Figure 5-3 shows the posterior distributions for the β parameters.

Parameter	Estimate	SD	95% PI
β_0	156.6	6.3	(144.1, 169.0)
β_1	37.5	8.9	(20.1, 55.0)
β_2	-100.7	9.2	(-118.1, -82.4)
β_3	-41.8	12.5	(-66.1, -16.9)
σ^2	134.9	3.9	(127.5, 142.8)

Table 5-2 Parameter estimates for informative prior model (experiment two).

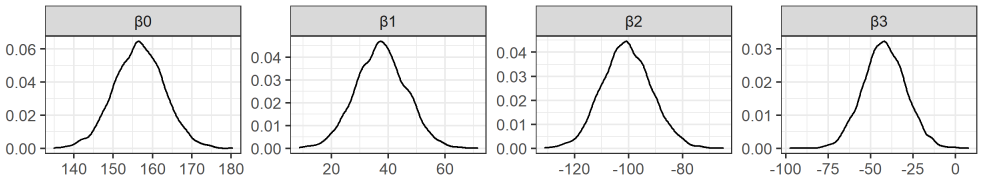

Figure 5-3 Posterior distributions for reference prior model beta parameters.

Again, the Stroop effect estimates for various conditions match expectations from the paper. Also consistent with expectations, the parameter estimates for this informative prior model are more precise (have a smaller probability interval range) than those observed for the experiment one reference prior model.

To more directly observe how informative the priors were for this model, a model using reference priors was also estimated on the experiment two data. Note that the estimates for the reference prior model are similar to what would be obtained by fitting a traditional model using least squares or maximum likelihood estimation. The differences between the final posterior parameter distributions for the two models are shown in Figure 5-4. The posterior sample statistics for the reference prior model can be found in Table 5-3 (for comparison with Table 5-2). It is apparent that the informative priors are driving more precise parameter estimates and differences in the center of the posterior distributions.

It is important to emphasize that more precise estimates do not necessarily mean *better* estimates. The data from experiment one could be biased in some way (for example, having a random sample of people who happen to be good at focusing their attention in demanding scenarios) and, as a result, not a good representation of overall trends in the AC task. If there is some sort of bias present in the data being used to inform the model on experiment two data, then the parameter estimates, while more precise, will also be biased. How informative an "informative" prior truly is depends on the quality of information used to construct the prior. A better understanding of how useful the informative prior model is in a predictive modeling or machine learning context could be obtained by assessing the quality of predictions from the informative prior model versus those from the reference prior model.

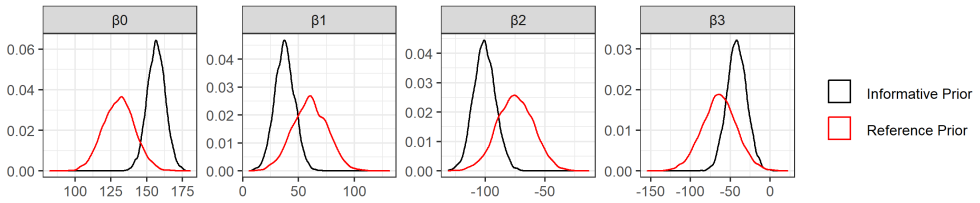

Figure 5-4 Posterior parameter distribution comparisons for experiment two data models.

Parameter	Estimate	SD	95% PI
β_0	130.7	11.0	(109.7, 152.6)
β_1	60.2	15.5	(29.5, 89.9)
β_2	-74.9	15.3	(-105.0, -45.2)
β_3	-63.8	21.7	(-106.5, -21.6)
σ^2	134.4	3.8	(127.3, 142.3)

Table 5-3 Parameter estimates for informative prior model (experiment two).

5.4.3. Inferential Power of Bayesian Approach

The Bayesian approach to modeling does not only provide the benefit of allowing prior knowledge to be incorporated. It also provides a flexible framework for making intuitive inferences about relevant research questions.

For example, using this approach in the context of this data, we can get a direct estimate of the probability that telling a subject to expect incongruent trials will decrease the Stroop effect, getting at the question of how likely a cue is to improve performance in a high AC setting. This question can be operationalized in the model by calculating the posterior distribution of $\beta_1 + \beta_3$. In this case, the estimated probability that a cue leads to a decrease in the Stroop effect for a mostly incongruent list is 64.34%. The reference prior model provides a similar conclusion The expected magnitude of the difference in Stroop effect for both the informative prior and reference prior models can be seen in Figure 5-5. While decreases in Stroop effect are more frequent than not, we do not have strong evidence to suggest the improvement is significantly different from zero. Additionally, improvements most often only represent up to 30 ms improvements (there is only a 1.72% estimated probability of a larger improvement from the informative prior model).

We can also estimate the probability that a cue will increase the Stroop effect in a mostly congruent list, getting at the question of how likely a cue is to reduce performance in a low AC setting. The β_1 posterior distribution (seen in Figure 5-3) can be used directly and suggests that the Stroop effect has a 100% likelihood of increasing (for both the reference prior and informative prior models).

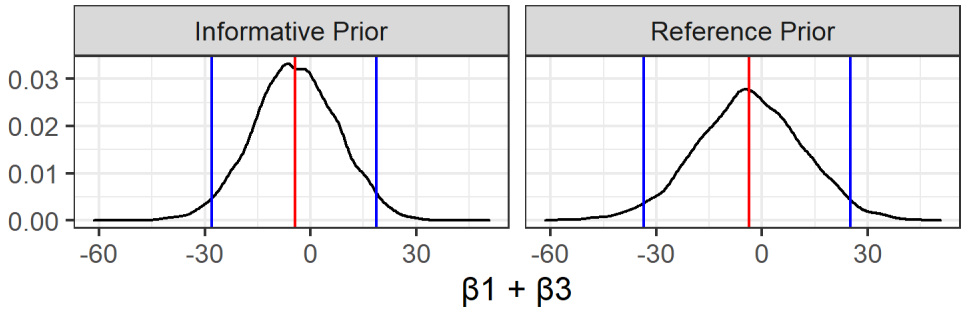

Figure 5-5 Posterior distributions for $\beta_1 + \beta_3$, showing expected change in mostly incongruent list Stroop effect when adding a cue. Both models are shown with the posterior mean in red and the 95% probability interval in blue.

5.5. Discussion

This is interesting regarding how participants use the cues, in comparison to the behavioral data from prior human subjects experiments. Namely, individuals are more likely to relax AC rather than heighten it. There is a behavioral consequence to this finding; overall there is a larger Stroop effect when most of the trials are congruent and participants are cued to that fact. The probability of the Stroop effect change is an interesting way of framing the question that can inform future work regarding an individual's decision to use the cue. Further, the fact that there is a smaller likelihood in the MI for using the cue does not mean that all individuals use cues in the MI case the same. Specifically, there may be a range of cue use across different individuals in that case, that can be explored in future work regarding conditions, or individual characteristics that drive someone to use cues in MI conditions.

6. MACHINE LEARNING MODEL

In this Section we describe the model and results of a machine learning (ML) model. In the Bayesian model, the goal was to model the Stroop effects under different cue (and list) conditions. Here, we turn the paradigm around and predict cue condition (state) of an individual based on the Stroop effect.

6.1. Data

The data used for the ML models is also pre-processed, as described in the Bugg et al. 2015 paper [10] and in Section 4.1.

6.2. Method

6.2.1. Data preprocessing and feature engineering

For a general supervised machine learning prediction task, we need a labeled dataset of N training samples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $y_i \in \{\text{cued}, \text{uncued}\}$ is the participant cue state and $x_i = [t_1, \ldots, t_M]$ is a feature vector of normalized[1] response time information for each of the M trials. In ML terminology, this would allow us to train a model that predicts the label y_i using all the information from the feature vector x_i. That is, the model discovers patterns most useful in x_i for predicting y_i.

The 2015 Bugg et al. experiment used a 2 (cued/uncued) $\times$ 2 (MC/MI) within-subjects design. In total, each participant was shown 32 lists, where the lists were randomly intermixed among the four list types. Because each participant is not shown the *same* ordered list of M trials, it is not possible to obtain the same x_i across all participants.

Instead, we must perform the following additional steps. For *each* participant:

1. We aggregate the trials into two groups: cued and uncued. Note that this process actually produces two data points, one for cued and one for uncued, for each participant.
2. We compute the mean of the response time τ by condition: Congruent trial, MC list (*CMC*), Incongruent trial, MC list (*IMC*), Congruent trial, MI list (*CMI*), and Incongruent trial, MI list (*IMI*). We note that we have explored other summary representations of the observed response times, including: median, min, max, and standard deviation but did not observe improved performance.
3. We subtract the mean response times of the congruent trial from the incongruent trial in (2) above to get the mean Stroop effect σ by list condition: MC list (*MC*) and MI list (*MI*).
4. We additionally consider further dividing the conditions in (2) and (3) above by the *color*.

So, for example, (2) above gives us a new feature vector $x_i = [\tau_{CMC}, \tau_{IMC}, \tau_{CMI}, \tau_{IMI}]$, (3) gives $x_i = [\sigma_{MC}, \sigma_{MI}]$, and (4) expands the feature set by further separating τ and σ by color condition.

6.2.2. Model selection and validation

Given labeled data $\{y_i, x_i\}$, $i = 1, \ldots, 2N$ ($2N$ for two data points for each of N participants) it remains to select a ML algorithm for the underlying prediction model. We use the Linear Support Vector Machine (LinearSVM) available in Python's Scikit-learn 0.23.1 with default parameters [23] for its ease of interpretability, particularly with respect to feature importance[2, 15].

Briefly, in its simplest form, the objective of an SVM is to find a hyperplane that separates the labeled data into the two distinct classes (extensions for multiclass problems exist), while also

[1] We normalize the times so that $t_j \in [0, 1]$ for each particpant

maximizing the distance between the hyperplane and the nearest point from either group (hard-margin). The coordinates of the vector orthogonal to the hyperplane form the weights (coefficients) of the model. From the weights, it is possible to do two things. First, we can determine feature importance according to the relative magnitude of the weights. Second, new data items can be labeled depending on which side of the hyperplane they fall (computed by taking the dot product with the orthogonal vector).

In a deployed setting, we would apply our SVM model that has been trained on all the labeled data to make predictions on new, unlabled data. However, without validating the model first, it is not possible to know how good the new predictions are. Therefore, a cross-validation test is performed first, in which part of the labeled data is withheld during training and used to test (validate) the performance of the model during prediction. We use the k-fold cross validator, which splits the data into k consecutive folds. Each fold is then used once as the test (validation) set, while the remaining $k-1$ folds form the training set. We use $k = 5$ and perform 10 runs of each of the cross-validation experiments. We report back on the mean and standard deviation of the accuracies and confusion matrices.

6.3. Results

Table 6-1 shows the prediction accuracy using the different feature sets. The accuracies using mean response times and mean Stroop effects were the poorest performing, with 0.54 (0.12 SD) and 0.59 (0.13) accuracies, respectively. However, the results of this experiment are revealing in that the SVM weights indicate that the trials in the MC list are most useful for predicting cue condition, whereas the trials in the MI list are less useful. In fact, if we remove the MC list from the response times and Stroop effects, we obtain 0.47 (0.09) and 0.48 (0.06), respectively, which is worse than random guessing. As discussed in Section 6.2.2, we also separate out the response times and Stroop effects by color condition. It can be seen that this improves the performance of the model, where we obtain 0.69 (0.11) and 0.64 (0.12) accuracies using the response times and Stroop effects, respectively.

In addition to the accuracies, we also present the confusion matrices using the response times (Table 6-2) and the Stroop effects (Table 6-3), which show the prediction results at the class level.

6.4. Discussion

In this Section, we primarily discuss the limitations of the current model and provide brief suggestions for future work. Section 8 discusses future work in much more detail.

The primary limitation of this effort was a mismatch between data and computational model. Specifically, the experimental data was collected a priori to evaluate the role of expectation (via cued and uncued lists) at the group, not individual level. As such, the experiments were randomized so that (1) list conditions randomly alternated between MC and MI lists, (2) cue conditions randomly alternated between cued and uncued, and (3) trials within lists were presented randomly. After randomizing the experiments, mean response times under different

Feature representation	Accuracy	Feature importance
Response time	0.54 (0.11)	*: $w_{CMC}, w_{IMC}, w_{CMI}, w_{IMI}$ = 0.92, 1.12, 0.20, 0.45
Response time (color)	0.69 (0.11)	B: $w_{CMC}, w_{IMC}, w_{CMI}, w_{IMI}$ = 0.69, 0.75, 0.53, 0.56 G: $w_{CMC}, w_{IMC}, w_{CMI}, w_{IMI}$ = 0.94, 1.06, 0.28, 0.18 R: $w_{CMC}, w_{IMC}, w_{CMI}, w_{IMI}$ = 0.16, 1.65, 0.90, 0.63 Y: $w_{CMC}, w_{IMC}, w_{CMI}, w_{IMI}$ = 0.33, 0.76, 0.22, 0.36
Stroop	0.59 (0.13)	*: w_{MC}, w_{MI} = 1.43, 0.49
Stroop (color)	0.64 (0.12)	B: w_{MC}, w_{MI} = 0.55, 1.18 G: w_{MC}, w_{MI} = 1.18, 0.66 R: w_{MC}, w_{MI} = 1.22, 0.25 Y: w_{MC}, w_{MI} = 0.59, 0.29

Table 6-1 Accuracy (mean and standard deviation in parentheses), together with feature importance obtained from the weights of the SVM, using the different feature representations.

		Predicted Class			
		Response time		Response time (color)	
		cued	uncued	cued	uncued
Actual Class	cued	0.54 (0.28)	0.46 (0.28)	0.67 (0.16)	0.33 (0.16)
	uncued	0.45 (0.28)	0.55 (0.28)	0.30 (0.21)	0.70 (0.21)

Table 6-2 Confusion matrices for predicting cue condition using response time (with color, right).

		Predicted Class			
		Stroop		Stroop (color)	
		cued	uncued	cued	uncued
Actual Class	cued	0.58 (0.19)	0.42 (0.19)	0.60 (0.16)	0.40 (0.16)
	uncued	0.39 (0.20)	0.61 (0.20)	0.32 (0.20)	0.68 (0.20)

Table 6-3 Confusion matrices for predicting cue condition using Stroop effect (with color, right).

conditions were obtained and the significance in differences of Stroop effects were evaluated in the context of expectation.

In future work, we would design an experiment such that every participant receives the same order of trials in each of the four quadrants of the AC settings versus demand risk matrix. This would allow us to develop a ML model that compares trial-by-trial response times across participants, rather than collapsing all the data into summaries of the response time observations (here, we only looked at mean, median, min, max, and standard deviation). Finally, the order of the presented lists and trials could be designed with HCML in mind to capture trait versus state level cue use and effects of learning and fatigue, among others; see Section 8 for further discussion.

7. CONCLUDING REMARKS

Attempts to model AC were pursued to enable the research team to (1) draw inferences about AC and human performance in real-world scenarios (i.e., our guard example) and (2) to make predictions about individual characteristics based on human data. Both approaches are unique as compared to what has been done in the literature to date measuring and modeling AC.

The first approach was to develop a Bayesian model that represents the experimental conditions. The goal was to be able to query the model regarding scenarios of interest that represent a risk model for AC mismatches. Recall, there are two experimental conditions that represent an AC mismatch - one in which an individual is in a state of low AC but encounters a high attentional demand scenario. This can lead to missed, inaccurate, or delayed responses. The second mismatch occurs when an individual is in a state of high AC but encounters a low attentional demand scenario. In this case, unnecessary attentional resources may be dedicated to providing an accurate or timely response. Possible consequences of this include a limited capacity to respond to subsequent additional information (such as a guard dividing attention between competing tasks) or fatigue.

The research team posed questions, and leveraged the Bayesian model to draw inferences about those questions, which included:

- What is the probability that telling a subject to expect incongruent trials will decrease the Stroop effect? This question provides insight into whether, and how likely it is that a cue will improve performance in a high AC demand context. It was found that a cue lead to a decrease in the Stroop effect 64% of the time. This can be interpreted as participants intentionally using the cue to heighten attentional control (some of the time, or some of the participants heighten attentional control).
- What is the probability that a cue will increase the Stroop effect in a mostly congruent list? This question provides insight regarding how likely a cue is to reduce performance (increase the Stroop effect) in a low AC demand context. In this case, it is likely that 100% of the time, the Stroop effect will increase, which suggests that participants are relaxing AC in an intentional fashion to cues that signal low AC demand.

Framing questions regarding AC in this way is unique as compared to literature on modeling AC, and represents a fruitful approach for examining real-world-related questions that are informed by carefully controlled laboratory data. This enables us to study conditions or individual characteristics in real-world settings.

In the ML modeling approach, the goal was to predict the state an individual was in - either cued or uncued. Results in this case demonstrated that with 69% accuracy, we could differentiate a cued from uncued state, which does not represent significantly-greater-than-chance performance. This is likely a symptom of the experimental design, which was carefully controlled to allow for causal interpretations of the data but limited the ability to develop a model, as the experiments were not designed for use in ML modeling to differentiate cued from uncued conditions. However, this represents a promising avenue for future work. Human subjects studies can be designed to allow for better predictions in ML models, either by (1) developing an experimental

design that more closely resembles real-world scenarios, or (2) by designing an experiment using matched lists in a fixed order for all subjects, thus allowing for a direct mapping of trial types and conditions across participants.

Based on our success implementing a Bayesian model and arriving at reasonable (supported by human behavioral data) inferences to questions that represent real-world scenarios of interest, this effort shows great potential for future work. This work should include developing higher-fidelity models, as well as hierarchical models. Further, while the ML models did not achieve significantly-greater-than-chance performance differentiating cued from uncued states, predicting an individual's AC state is a worthwhile endeavor, for which the research team has developed avenues to pursue that are well-suited to an HCML approach. Six avenues to pursue for future research follow.

8. FUTURE WORK

8.1. Future Direction 1: Modifying Botvinick et al. model to account for intentional, goal-directed AC

The data from [10] showing an asymmetric influence of the pre-cues cast doubt on individuals' ability to heighten AC when advance information signals it is valuable to do so (a high-risk situation), but show individuals readily relax AC when they are informed that attentional demands will be low (a low-risk situation). The Botvinick [3] model, as noted, is unable to account for this pattern. To our knowledge, there is no other extant model that can account for this pattern since extant models have ignored the role of intention (expectations) in adjustments to AC. Yet, there is both basic and applied value in accounting for this pattern, and more generally anticipating via a model a) the conditions under which an individual will intentionally heighten AC in response to advance information signaling they should do so, and/or b) which individuals will be most likely to do so. Therefore, implementing a higher-fidelity model of AC (such as Botvincik's model) in the Bayesian framework would provide a better overall representation of all relevant factors that affect AC and could be used for accounting and making predictions on individuals' intentions.

Another approach is to adapt the Botvinick [3] model to account for intention, captured via modifications to Equation 2 in [3], which specifies the degree of control adjustment in response to the degree of conflict detected over multiple preceding trials. Equation 2 might consider not only conflict detected on "multiple preceding trials" but also the conflict anticipated by the pre-cue, with weights being assigned to each. Assuming cues are valid, these two pieces of information converge but they might be weighted differently for the mostly incongruent and mostly congruent condition. For example, the weight assigned to the cue might be low in the mostly incongruent condition but high in the mostly congruent condition while the conflict experienced (on multiple preceding trials) might be weighted equally for the two conditions. A high cue weight in the mostly congruent condition might lead to some multiplicative effect on the degree of adjustment (e.g., without cues the adjustment might be X units of control but with the cue it may be X times 1.5 units of control). These weights could be model-learned based on the group-level behavioral data from [10] by estimating how much the Stroop effect increases (or decreases) on average in

the cued compared to uncued conditions. More sophisticated modeling could attempt to get at how long individuals like a security guard can sustain an intentional heightening of AC (an increased control adjustment based on the cue) by incorporating trial level data within each list. For example, Suh and Bugg (in press) showed that pre-cues are used initially but cue use decreases rapidly across a 10-trial list in the mostly incongruent condition, whereas in the mostly congruent condition, cue use is high initially and remains high throughout the list. This suggests that the intentional heightening of control may be more difficult to sustain than the relaxation of control.

Another potential way to adapt the [3] model is to incorporate a decision module (module that captures individuals' decisions to use or to not use external information such as pre-cues). Incorporating a decision module seems potentially valuable because the primary pattern (asymmetrical influence of the pre-cues) in [10] might most parsimoniously be explained by a decision module whose outcome is "no" (not to use the pre-cue) in mostly incongruent lists but is "yes" (use the pre-cue) in mostly congruent lists. A Bayesian analysis on existing datasets could inform the prior probabilities of heightening or relaxing AC when pre-cues are provided in each condition (mostly incongruent and mostly congruent).

For example, one could get Bayesian estimates of choice to use the cue based on the numbers of lists where participants show a smaller Stroop effect in the cued mostly incongruent condition compared to the group average of the uncued mostly incongruent condition (and similarly in the cued mostly congruent condition compared to the group average of the uncued mostly congruent condition. Based on the group level behavioral data of [10], the probability should be low for heightening of control (cue use in mostly incongruent condition) and high for relaxation of control (cue use in mostly congruent condition). The standard Botvinick et al. model would again apply to the uncued lists (since there is no decision to use or not use cues) but whether the cue-based control adjustment occurs in the cued lists would be informed by the Bayesian analysis. One could then determine if this model can account for behavior in a new dataset that was not used for the Bayesian analysis.

Additional strengths of the Bayesian approach not covered in this paper could be brought to bear for this future direction and are worth demonstrating. First, Bayesian methods readily extend to hierarchical or multilevel models, which could be used to understand differences in performance related to individual research subjects. Second, while normal distributions were used as informative priors in this paper, the MCMC sampling approach allows great flexibility in types of distributions one can use. Finally, though only one step was demonstrated in this paper, Bayesian methods allow easy updating of a model sequentially (using the outcome of one model to inform an updated model on new data), lending itself well to a sequential learning framework that could be useful as part of a HCML capability.

Thus, to represent an individual-level model using our Bayesian modeling approach, we would develop a hierarchical model, where the first level would represent the individual with factors and a random effects term; the next level would represent effects associated with lists; an additional block-level could be included as well. In this case, we could make queries regarding if particular individuals tend to show more shifts toward negative when they are in a cued versus an uncued situation. This would also permit us to look at individualized parameters to compare distributions of expected increases in intention between participants, based on individual-level factors.

8.2. Future Direction 2: Uncovering individual and contextual determinants of intentional goal-driven AC

A related but distinct future direction will aim to uncover the individual and contextual determinants of intentional goal-driven AC, that is to determine the attributes of individuals or contexts that are associated with cue use and potentially therefore decreased risk of AC failures. This has future applied value because it informs the question of who is most likely to intentionally heighten control when instructed to do so (which persons would be most effective as a security guard in situations that demand such heightening of control?) and under what conditions is cue use most probable.

For the question of who is most likely to intentionally heighten control, for each participant (rather than for the group-level data) one could compare the mean Stroop effect in each cued list (e.g., each mostly incongruent cued list) to the average Stroop effect across all uncued lists of the same type (e.g., average of all mostly incongruent uncued lists). For each participant, two values should be derived—(1) the number of lists that AC was intentionally heightened for the mostly incongruent cued condition as indicated by a smaller Stroop effect relative to the uncued condition, and (2) the number of lists that AC was intentionally relaxed for the mostly congruent cued condition as indicated by a larger Stroop effect relative to the uncued condition. Presumably, across participants one will observe a range for each value with some participants having high numbers (indicating high frequency of intentional adjustments in AC) and some having low numbers. Ultimately, one could conduct a new study in which they administer the pre-cued lists paradigm in conjunction with individual differences measures that are of value to Sandia's mission (e.g., measures that might typically be used for purposes of assessment, interviewing, as well as other psychological indicators anticipated to be associated with the willingness/ability to intentionally adjust AC). Since there is greatest interest in the high attentional demand mostly incongruent condition, one could focus on trying to identify predictors of cue use in this condition (i.e., what measures correlate with the number of cued lists on which AC was heightened in the mostly incongruent condition?

For the question regarding the conditions under which cue use is most probable, one could return to the group-level analytic approach and compare the same groups' cue use across two different contexts. For example, in Bugg et al. (2015) Experiment 4, there were two contextual conditions—a low incentive condition in which participants earned a small incentive for good performance and a high incentive condition in which they instead earned a large incentive. In this experiment, it was found that the high incentive condition promoted pre-cue use in the critical high AC demand, mostly incongruent condition. This fits with the notion that motivational factors are important in goal-driven AC. Future experiments could explore other contextual factors that have applied value such as: various training approaches (how is information communicated to the guard about the need for intentional heightening of AC at certain high risk times), commitment strategies (whether the guards are asked to engage in commitment strategies like implementation intentions that can promote intention fulfillment), or factors like time-on-task (how long after the start of one's guard shift [or after the imperative information is communicated] does the need for AC arise) and time-of-day.

Combining these two levels of analysis may afford even greater predictive power. That is, one

could simultaneously consider individual and contextual differences. As a starting point, one could use the data from Bugg et al. (2015) Experiment 4 and derive 4 values: number of lists each participant intentionally engages AC in the a) low incentive, mostly incongruent, b) high incentive, mostly incongruent, c) low incentive, mostly congruent, and d) high incentive, mostly congruent cued lists (relative to the average for that participant in each of those four conditions). Possibly, some individuals may be more likely to heighten control when needed compared to others and they may do so regardless of the incentive structure. In contrast, other individuals may only be more likely to heighten control when needed compared to others when they are in a high incentive context (rewarded for their performance). Like the concept of individually-tailored-medicine or therapy, one could use results like these to determine how to maximize the likelihood of AC success (and therefore decrease AC risk) for different employees. Some guards may need contextual boosts to achieve high levels of AC (to encourage them to use advance information to intentionally heighten AC when there is a threat) while others may not, some may be motivated by some contextual boosters but not others, and still others may be neither individually inclined to heighten AC nor driven by contextual changes.

8.3. Future Direction 3: Modeling stimulus-driven AC

As described earlier, stimulus-driven AC is control that is governed by external cues that have become associated with attentional demands. These associations are learned via experience and enable subsequent retrieval of the associated AC setting when encountering the critical cues. A key paradigm for studying stimulus-driven AC in the lab is the item-specific proportion congruence (ISPC) paradigm (Braem et al., 2019). In this paradigm, participants encounter 50% high demand (incongruent, e.g., the word GREEN in red the word DOG paired with a bird picture) and 50% low demand (congruent, e.g., the word GREEN the word BIRD paired with a bird picture) trials, which are randomly intermixed during the experiment. In the picture-word version that is commonly used ([8, 12, 25]), participants name the animal in a picture and ignore the word (e.g., incongruent trials such as the word DOG paired with a bird picture, congruent trials such as the word BIRD paired with a bird picture). Most importantly, unbeknownst to participants, the proportion of incongruent and congruent trials varies across items. This manipulation is called the ISPC manipulation.

For example, for some participants, birds and cats are mostly low demand (referred to as the mostly congruent condition) whereas dogs and fish are mostly high demand (referred to as the mostly incongruent condition). The key findings from these experiments is the ISPC effect–the pattern whereby the Stroop effect is smaller for the mostly incongruent (i.e., mostly high demand trials) condition compared with the mostly congruent (i.e., mostly low demand trials) condition, indicating that a higher AC setting was engaged post-stimulus onset for the mostly incongruent condition compared with the mostly congruent condition.

This is important because the post-stimulus nature of the attentional adjustments helps rule out a goal-directed AC account of the ISPC effect. (Additionally ruling out this account is the fact that all items [birds, cats, dogs, fish] are presented equally often and overall, half of the trials are congruent, and half of the trials are incongruent such that participants can neither predict which

animal is going to appear on the next trial or whether the trial will be congruent or incongruent, information that would be critical for goal-directed AC to operate.).

A related effect is the context-specific proportion congruence (CSPC) effect—the pattern whereby the Stroop effect is smaller for a context in which AC demands are mostly high compared to a context in which AC demands are mostly low. In the lab, the typical manipulation used to produce CSPC effects is a location manipulation (the location on screen where stimuli appear is either mostly incongruent or mostly congruent). From an applied perspective, CSPC effects may better capture stimulus-driven AC in real-world situations considered in our use case example in Chapters 1 and 2 (e.g., when security guard's experience in the northeast part of campus mostly entails higher AC demand/engagement and their experience in the northwest mostly entails lower AC demand/engagement such that over time a high level of AC becomes associated with the northeast location whereas a low level of AC becomes associated with the northwest location).

In the context of our conceptual risk model, the ISPC and CSPC effects imply that a higher risk of AC failure accompanies stimulus-driven AC that is triggered in response to cues that have historically been associated with low demand in the past, but which in the moment co-occur with a high demand situation (a security threat presents itself in the northwest location). Conversely, a low risk of AC failure accompanies stimulus-driven AC that is triggered in response to cues that have historically been associated with high demand in the past, and which in the moment co-occur with a high demand situation (a security threat presents itself in the northeast location).

With respect to modeling, there are some models that have been developed to account for ISPC/CSPC effects [1, 28]. Blais [1] presented an adaptation of the Botvinick [3] model and the primary change is that in this model, the task representation weights are adjusted in an item-specific fashion. In other words, control takes an item-specific form whereby conflict on a given trial changes the connection weight between the task demand unit and the color on that specific trial with the weight being increased to reflect increased attention to color (color naming) when conflict is high (on incongruent trials). On congruent trials, the color-naming weight for the presented color is selectively lowered. To test the model, the authors generated 1000 random lists of 192 trials wherein an ISPC manipulation was implemented, and they submitted these trials to the model resulting in 1000 simulations. They found an ISPC effect whereby the average size of the Stroop effect (here measured in cycles) was larger for the mostly congruent items compared to the mostly incongruent items.

Two limitations of this model (see also Verguts & Notebaert [28]) are that it is unclear if these models can simulate a) ISPC effects in ISPC paradigms that have isolated AC from other mechanisms that can produce ISPC effects (see [8, 12, 25]) such as learning of simple stimulus-response associations, and b) CSPC effects. No study to date has attempted to simulate ISPC effects from such paradigms or CSPC effects using these models. Modeling the conditions under which CSPC effects are observed may be particularly valuable since, compared to ISPC effects, CSPC effects are (as noted above) potentially more applicable to Sandia's mission but additionally are less stable (see [13], for discussion). A model might be especially useful for predicting when stimulus-driven AC will succeed versus fail in situations where a contextual cue like location is responsible for cueing AC.

Such a model might benefit from considering the two "sides" to stimulus-driven AC, which are the learning side (the accumulation of experiences in for example, the northeast and northwest locations that allow the guards to learn the associations between an area of space and AC demand) and the retrieval side (after learning, the retrieval of those associations when encountering stimuli in either of those locations in the future). From a measurement perspective, however, these two are difficult to tease apart using most current experimental designs because the CSPC effect (or ISPC effect) that is measured in such designs reflects the contributions of learning and retrieval. One way to get estimates of both that are perhaps more independent than current designs afford, is to have a training (learning) phase be followed by a subsequent transfer phase that is separated in time from learning and involves stimuli that differ from those that were encountered in training (i.e., transfer stimuli). The CSPC effect in the training phase could then be used as the indicator of learning whereas the CSPC effect in the transfer phase could be used as the indicator of retrieval. Models that include a learning and retrieval component could then be evaluated by determining whether they can successfully simulate data from such an experiment.

We noted in Chapter 1 that, in addition to learning, stimulus-driven AC is affected by attention to predictive cues and the similarity of cues to those encountered during learning. More specifically, we suggested that stimulus-driven AC will be more effective to the extent that cues are salient or otherwise readily attended (e.g., security guard readily detects occurrence of activity in a specific location), and cues are similar to those encountered during the learning process. There is evidence from experimental studies showing that stimulus-driven AC transfers to similar but novel cues that were not encountered during training (e.g., new pictures of dogs, birds, etc. in the ISPC studies; [8, 12], new locations on screen that are nearby trained locations in the CSPC studies [29, 30]. Yet, it remains unknown if such patterns would be observed in a more realistic environment where there is greater competition for an individual's attention (i.e., more cues available that could be attended) and cue similarity is more variable/less controlled than in the lab.

To examine CSPC effects in such an environment, one could design a virtual world that mimics the environment and security challenges at Sandia and embed predictive cues for AC into the virtual environment. Using the location example again, one could make one gate predictive of high AC demands and a different nearby gate predictive of low AC demands. Participants would be asked to respond to stimuli in the different locations. Critically, the environments would be designed to include competing stimuli such as those that a guard might encounter in everyday life like a plane flying nearby, a jogger running past, or a honking horn. An experiment like this would offer important insights into the question of whether participants still learn the relationships between cues and attentional control demands in this type of environment, and if they do, whether stimulus-driven AC continues to operate in the face of competing attentional demands and shows transfer to novel cues (e.g., presentation of a stimulus in a location nearby one of the gates that is not precisely the location encountered during training). Such data would be highly important for informing future environmental modifications that may be made with the goal in mind of promoting use of stimulus-driven AC.

8.4. Future Direction 4: Weighting goal-directed and stimulus-driven AC

Throughout this paper, we have considered goal-directed AC and stimulus-driven AC rather independently. In so doing, one might be left with the impression that one or the other operates at any given time. That may be true under some conditions. For example, [19] found that pre-cues that were presented on each trial (rather than before an entire list of trials as in the pre-cueing paradigm described earlier) reduced or eliminated ISPC effects. This suggests that when attention was intentionally heightened in response to pre-cues signaling that conflict (high AC demand) was likely on the next trial, adjustments in attention based on stimulus-driven AC did not occur. However, in most real-world contexts, advance information about an upcoming high AC demand is likely not repeated over and over prior to each possible occurrence of a stimulus (e.g., threat) but rather is conveyed to individuals before a work shift for example, or at the start of a week (if a threat is expected sometime during that week).

Thus, it remains an open question whether use of intentional goal-directed AC would interfere with stimulus-driven AC in such circumstances given that goal-directed AC is difficult to sustain over time. Future studies might therefore examine this more directly using experimental designs that better mimic the real-world scenario just described. Additionally, the flip question can be asked—when stimulus-driven AC is being used, can goal-directed AC intervene if the outcome of stimulus-driven AC is suboptimal. This question relates back to the conceptual risk model and particularly the upper right cell. In a case where stimulus-driven AC is used, the upper right cell refers to an individual retrieving a low AC setting but encountering a high AC demand resulting in sub-optimal performance. Here, it may be especially important to understand if the individual could intentionally override that retrieval or counteract that retrieval by heightening AC in that moment.

8.5. Future Direction 5: Human Constrained ML Approach to Differentiate Individual Characteristics

Under Future Direction 2, we considered the potential to examine individual differences in intentional goal-driven AC as indicated by cue use. From the analysis proposed in that section, one could identify “high cue users” and “low cue users”. Machine learning could then be applied to determine if there are patterns in the data that accurately differentiate these two groups of individuals. This might provide an individual differences signature of cue use, and more specifically how inclined or able an individual is to take external information about expected AC demands and use it to adjust attention in line with those demands. Theoretically, such individual differences might represent a trait or state level variable. This could be assessed by examining cue use in a different task (or experimental paradigm) and determining if the same individuals who were high cue users in one task are also high cue users in a second task (and those who were low were again low).

For example, this could be evaluated by determining if the signature of cue use identified by machine learning again categorizes the same individuals as high cue users and the other individuals as low cue users. If consistency is observed across tasks, that would provide evidence for a trait level variable. If, by contrast, cue use is a state level variable, then one might instead

observe that cue use is more strongly determined by one's concurrent state—e.g., Is the individual motivated? Fatigued? Interested in the task? etc. From an applied perspective, this approach might begin to identify what individuals may be most well-suited for an AC-demanding job (e.g., a job that requires one to reliably use advance information [like a pre-cue] to intentionally adjust control across situations/contexts).

Another way in which ML could be applied to addressing a question from Future Direction 2 relates to the question of contextual influences on AC. However, rather than consider goal-directed and AC, here the application would be to stimulus-driven AC. As described in Chapter 1, some evidence suggests stimulus-driven retrieval of associated levels of AC may become automatized, such that even under concurrently demanding situations (when attention is devoted to another task while performing security related tasks), cues can effectively heighten AC [25]. In [25], participants completed a Stroop task with an ISPC manipulation embedded while under no load, low load, or high load. Load referred to a concurrent working memory task they performed while doing the Stroop task. The key finding was that the ISPC effect was robust regardless of load. These data provide an opportunity to determine whether machine learning can accurately differentiate load conditions. That is, are there characteristic patterns of responding in these conditions that can inform whether individuals are devoting effort to a secondary task while performing a primary task.

Finally, under Future Direction 3, we considered ways in which stimulus-driven AC might be modeled and noted the importance of being able to differentiate the learning of stimulus-driven AC from the retrieval of stimulus-driven AC settings. ML might also offer a means to distinguish these stages by determining whether there are patterns of data that change across time during an experiment in ways that may theoretically map onto stages of learning and retrieval (post-learning). For example, it may be that RT variability decreases and asymptotes as learning reaches its maximum (ceiling) and one can infer that performance (e.g., CSPC effects) post that point in time, correspond to the retrieval side of stimulus-driven AC. This would hypothetically provide a way to delineate those stages for purposes of modeling the processes involved in each stage.

8.6. Future Direction 6: Building upon Bayesian Inferential Modeling Approach

An essential next step for the Bayesian Inferential Modeling approach appove would be validating whether the increased predictive precision afforded by the informative prior model provides better predictions. This would be done by comparing the predictions made on a new dataset by the reference prior model built on experiment two data in [10] to the predictions made by the informative prior model built on the same data.

Another method worth exploring would be the use of power priors [14, 20], which allows for weighting how much informative priors based on historical data affect the final model. Especially relevant research questions for power priors in this context include (1) how weighting affects predictive accuracy of the model and (2) how appropriate weighting changes based on the amount of historical data available to inform a model of AC.

REFERENCES

[1] Chris Blais, Serje Robidoux, Evan F Risko, and Derek Besner. Item-specific adaptation and the conflict-monitoring hypothesis: a computational model. 2007.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[3] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

[4] Senne Braem, Julie M Bugg, James R Schmidt, Matthew JC Crump, Daniel H Weissman, Wim Notebaert, and Tobias Egner. Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences*, 23(9):769–783, 2019.

[5] Todd S Braver, Jeremy R Gray, and Gregory C Burgess. Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. *Variation in working memory*, 75:106, 2007.

[6] Julie M Bugg. Context, conflict, and control. In Tobias Egner, editor, *The Wiley handbook of cognitive control*, pages 79–96. John Wiley & Sons, 2017.

[7] Julie M Bugg and Matthew JC Crump. In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in psychology*, 3:367, 2012.

[8] Julie M Bugg and Abhishek Dey. When stimulus-driven control settings compete: On the dominance of categories as cues for control. *Journal of Experimental Psychology: Human Perception and Performance*, 44(12):1905, 2018.

[9] Julie M Bugg and Nathaniel T Diede. The effects of awareness and secondary task demands on stroop performance in the pre-cued lists paradigm. *Acta psychologica*, 189:26–35, 2018.

[10] Julie M Bugg, Nathaniel T Diede, Emily R Cohen-Shikora, and Diana Selmeczy. Expectations and experience: Dissociable bases for cognitive control? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(5):1349, 2015.

[11] Julie M Bugg, Nathaniel T Diede, Emily R Cohen-Shikora, and Diana Selmeczy. Expectations and experience: Dissociable bases for cognitive control? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(5):1349, 2015.

[12] Julie M Bugg, Larry L Jacoby, and Swati Chanani. Why it is too early to lose control in accounts of item-specific proportion congruency effects. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3):844, 2011.

[13] Julie M Bugg, Jihyun Suh, Jackson S Colvett, and Spencer G Lehmann. What can be learned in a context-specific proportion congruence paradigm? implications for reproducibility. *Journal of Experimental Psychology: Human Perception and Performance*, 2020.

[14] Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1):46 – 60, 2000.

[15] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[16] Matthew JC Crump and Bruce Milliken. Short article: The flexibility of context-specific control: Evidence for context-driven generalization of item-specific control settings. *Quarterly Journal of Experimental Psychology*, 62(8):1523–1532, 2009.

[17] Matthew JC Crump, Joaquín MM Vaquero, and Bruce Milliken. Context-specific learning and control: The roles of awareness, task relevance, and relative salience. *Consciousness and cognition*, 17(1):22–36, 2008.

[18] Nicola De Pisapia and Todd S Braver. A model of dual control mechanisms through anterior cingulate and prefrontal cortex interactions. *Neurocomputing*, 69(10-12):1322–1326, 2006.

[19] Keith A Hutchison, Julie M Bugg, You Bin Lim, and Mariana R Olsen. Congruency precues moderate item-specific proportion congruency effects. *Attention, Perception, & Psychophysics*, 78(4):1087–1103, 2016.

[20] J. Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and F. Chen. The power prior: theory and applications. *Statistics in medicine*, 34 28:3724–49, 2015.

[21] Jiefeng Jiang, Jeffrey Beck, Katherine Heller, and Tobias Egner. An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nature Communications*, 6(1):1–11, 2015.

[22] Jiefeng Jiang, Katherine Heller, and Tobias Egner. Bayesian modeling of flexible cognitive control. *Neuroscience & Biobehavioral Reviews*, 46:30–43, 2014.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[24] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

[25] Jihyun Suh and Julie M Bugg. On the automaticity of reactive item-specific control as evidenced by its efficiency under load. *Journal of Experimental Psychology: Human Perception and Performance*, 47(7):908, 2021.

[26] Jihyun Suh and Julie M Bugg. The shaping of cognitive control based on the adaptive weighting of expectations and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, in press.

[27] Joseph Tzelgov, Avishai Henik, and Jacqueline Berger. Controlling stroop effects by manipulating expectations for color words. *Memory & cognition*, 20(6):727–735, 1992.

[28] Tom Verguts and Wim Notebaert. Hebbian learning of cognitive control: dealing with specific and nonspecific adaptation. *Psychological review*, 115(2):518, 2008.

[29] Blaire J Weidler and Julie M Bugg. Transfer of location-specific control to untrained locations. *Quarterly Journal of Experimental Psychology*, 69(11):2202–2217, 2016.

[30] Blaire J Weidler, Abhishek Dey, and Julie M Bugg. Attentional control transfers beyond the reference frame. *Psychological research*, 84(1):217–230, 2020.

APPENDIX A. APPENDIX: BUGG (2015) CODEBOOK

In Figure A-1, presents a mapping between the specific trials across cued and uncued conditions.

A.1. Experiment 1

Block = block # with numbers 1 – 8 representing practice (should be excluded from analysis) and blocks 9 and higher representing test trials. Note that each number (beginning with 9) appears in 10 consecutive rows because the experiment was comprised of mini-blocks (lists of 10 trials) and each list has a unique block number in this column

Trial = trial position within the block (1 – 10 with 1 representing the first trial, 2 the second trial, and so on in each block)

Color[Trial] = color of the stimulus presented on that trial

Experimenter.ACC = 0 or 1 (codes whether the Stroop response was incorrect or correct, respectively); ** note that this is the accuracy for congruent trials only

Experimenter1.ACC = 0 or 1 (codes whether the Stroop response was incorrect or correct, respectively); ** note that this is the accuracy for incongruent trials only

Procedure[Trial] = CongProc or IncongProc (codes whether the trial was congruent or incongruent in the Stroop task, respectively)

Running[Trial] = NMCList, NMIList, MCList, or MIList (codes whether the mini-block [list] is an uncued mostly congruent list, uncued mostly incongruent list, cued mostly congruent list, or cued mostly incongruent list, respectively); like the Block variable, the same code will appear in 10 consecutive rows because a mini-block was comprised of 10 consecutive trials of the same list condition

Stim.RT = vocal response time on Stroop task in ms; ** note that this is the RT for congruent trials only

Stim1.RT = vocal response time on Stroop task in ms; ** note that this is the RT for incongruent trials only

Word = color word that was presented on the trial

A.2. Experiment 2

Block = block # with numbers 1 – 8 representing practice (should be excluded from analysis) and blocks 9 and higher representing test trials. Note that each number (beginning with 9) appears in 10 consecutive rows because the experiment was comprised of mini-blocks (lists of 10 trials) and each list has a unique block number in this column

Trial = trial position within the block (1 – 10 with 1 representing the first trial, 2 the second trial, and so on in each block)

Color[Trial] = color of the stimulus presented on that trial

Experimenter.ACC = 0 or 1 (codes whether the Stroop response was incorrect or correct, respectively); ** note that this is the accuracy for congruent trials only

Experimenter1.ACC = 0 or 1 (codes whether the Stroop response was incorrect or correct, respectively); ** note that this is the accuracy for incongruent trials only

Procedure[Trial] = CongProc or IncongProc (codes whether the trial was congruent or incongruent in the Stroop task, respectively)

Running[Trial] = NMCList, NMIList, NFiftyList, MCList, MIList, or FiftyList (codes whether the mini-block [list] is an uncued mostly congruent list, uncued mostly incongruent list, uncued 50% congruent list, cued mostly congruent list, cued mostly incongruent list, or cued 50% congruent list, respectively); like the Block variable, the same code will appear in 10 consecutive rows because a mini-block was comprised of 10 consecutive trials of the same list condition

Stim.RT = vocal response time on Stroop task in ms; ** note that this is the RT for congruent trials only

Stim1.RT = vocal response time on Stroop task in ms; ** note that this is the RT for incongruent trials only

Word = color word that was presented on the trial

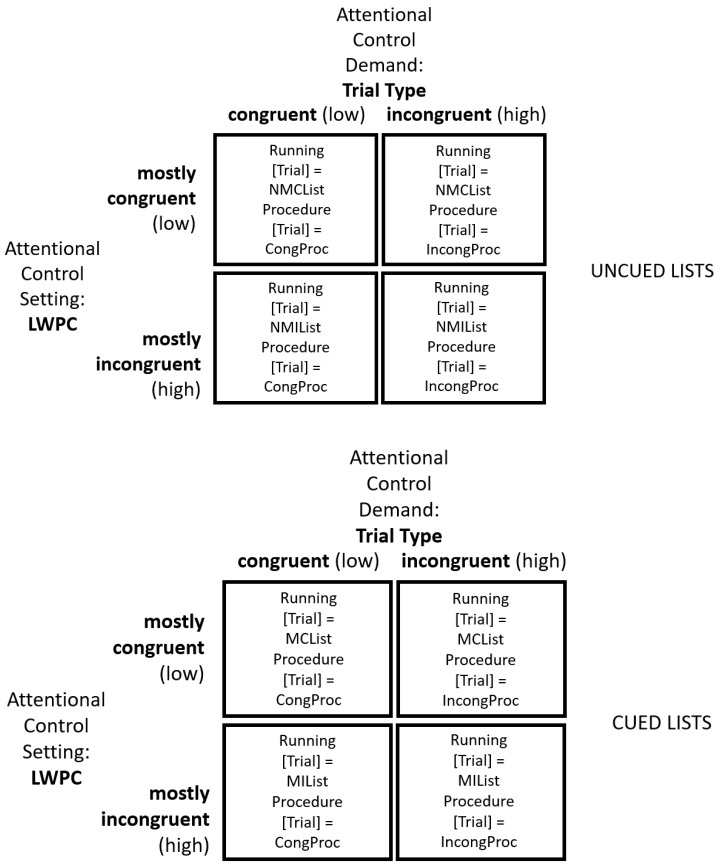

Figure A-1 Trial names for cued and uncued conditions.

DISTRIBUTION

Hardcopy—Internal

Number of Copies	Name	Org.	Mailstop
1	Courtney Dornburg	5954	1027
1	Susan Adams	6672	0152

Email—Internal (encrypt for OUO)

Name	Org.	Sandia Email Address
Melissa Finley	6824	mfinley@sandia.gov
Technical Library	1911	sanddocs@sandia.gov
Technical Library	01177	libref@sandia.gov

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.