A "Rational" Framework for Self-Control

Adan Gomez (gomeza2@rpi.edu)
Cognitive Science Department, Rensselaer Polytechnic Institute, 110 Eighth Street, NY 12180 USA
Ron Sun (dr.ron.sun@gmail.com)
Cognitive Science Department, Rensselaer Polytechnic Institute, 110 Eighth Street, NY 12180 USA

Abstract

While a number of disciplines have empirically investigated self-control (e.g., psychology, cognitive science, and sociology), along with philosophy, they have offered differing (although sometimes overlapping) perspectives. A process-based, mechanistic theory explaining empirical self-control data can help integrate these perspectives. A mechanistic (computational) approach through a computational cognitive architecture where simulations can be performed may unify the interpretations of empirical studies based on various (e.g., implicit-explicit) conflicts as well as utility calculation (e.g., from motivational considerations). Such a framework facilitates simulations that account for human data and capture notions of self-control capacity and control fatigue/reduction, facilitating detailed explanations.

Keywords: Self-control; Clarion; Cognitive Architecture; Conflict; Implicit; Explicit

Introduction

Self-control has been considered an important topic in a number of different fields of research on human behavior, such as psychology, philosophy, cognitive science, cognitive neuroscience, and sociology (Bertelsen et al., 2009; De Ridder et al., 2012; Gillebaart, 2018; Inzlicht et al., 2021; Kotabe & Hofmann, 2015; Scholz et al., 2022). The conceptualization of self-control, the theoretical models that explain it, and its relations to similar or related concepts such as cognitive control and self-regulation are being discussed in the literature (De Ridder et al., 2012; Eisenberg et al., 2019; Enkavi et al., 2019; Gillebaart, 2018; Inzlicht et al., 2021; Milyavskaya et al., 2019).

First, some brief discussion of very definitions of self-control is in order. Various theoretical perspectives offer overlapping insights into the notion of self-control despite differences in emphasis: some focus on overriding impulses for greater long-term rewards (Ainslie, 1975; Hoch & Loewenstein, 1991; Kirby & Herrnstein, 1995; Mischel, 1973; Rachlin & Green, 1972; Strotz, 1973), some others on altering prepotent responses (i.e., the colloquial notion of "willpower"; Friese & Hofmann, 2009; Fujita, 2011; Hofmann et al., 2009; Metcalfe & Mischel, 1999; Myrseth & Fishbach, 2009), and still others on motivational conflicts between desires and goals or between knowledge types (Fujita, 2008; Hoch & Loewenstein, 1991; Inzlicht et al., 2014; Mischel, 1973; Rachlin, 2004; Wehrt et al., 2020).

While nuanced differences do exist, in our opinion, these views may jointly illuminate facets of self-control rather than fundamentally opposing one another. While conceptual disagreements may hinder progress, a rigorous framework for understanding self-control that generates precise, detailed explanations that account for empirical data and phenomena of self-control can lead to a theory that embodies a deeper understanding of the underlying mechanisms of self-control. The present work is a preliminary step in that direction. It uses computational modeling to interpret experimental findings to address the lack of mechanistic theories in self-control research. Leveraging ideas of computational psychology, underlying processes are simulated to (hopefully and eventually) systematically explain self-control outcomes and effects. The present work conceptualizes self-control as a (often motivationally relevant) conflict that enables overriding or altering predominant, pre-potent, or automatic response tendencies. This work views such a conflict as competing mental representations that are simultaneously activated yet incompatible (Botvinick et al., 2004; Carter & Van Veen, 2007; Kotabe & Hofmann, 2015). Although Kotabe and Hofmann (2015) acknowledged that motivation-related conflicts underlay self-control phenomena, such as conflicts between desires and goals, other (e.g., symmetric) cases were excluded, and their potential relation to the desire-goal conflict is unclear. Across all various conflict types that we examined, the mutual incompatibility of simultaneously activated mental representations creates the circumstances underlying self-control. Resolving a conflict may sometimes require reducing or increasing the activation of one of the competing representations (Hofmann et al., 2009).

In some recent rational choice theories (Berkman et al., 2017; C. Chen et al., 2022; Z. Chen et al., 2020), the decision criterion used involves cost-benefit analysis weighing competing motivations. In these utility-maximization models, behaviors emerge from rational judgments about maximizing overall well-being. Classic decision theory (Neumann & Morgenstern, 1947) assumes perfect rationality in maximizing subjective value, while bounded rationality research recognizes human limitations in terms of deviation from perfect rationality. Some (Luce, 2014; Train, 2009) have modeled bounded rationality using stochastic choice (e.g., by a distribution with a "temperature" parameter balancing exploitation and exploration). Such stochastic choice models sample actions through utility values (Fudenberg & Kreps, 1993; Luce, 2012). Sun et al. (2022) proposed a comprehensive framework for explaining human performance using stochastic choice (with a Boltzmann distribution) based on utility calculation from intrinsic human motives and needs. Due to their theoretical appeals, the present work adopts utility calculation and stochastic selection (with a Boltzmann distribution), as specified by a generic computational model (i.e., the Clarion cognitive architecture), to synthesize several studies of self-control in a

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

"rational" (e.g., utility maximization) framework, both clarifying and unifying them.

This paper is structured as follows: First, the mechanisms in the Clarion cognitive architecture used in modeling self-control are described. Next, simulations and interpretations of two psychological experiments involving self-control are presented using the framework. Finally, the results of the simulations are discussed, and the framework that led to the simulations is considered.

The Clarion Cognitive Architecture

Clarion is a comprehensive, domain-general cognitive architecture that represents the essential structures, processes, and mechanisms of the mind (Sun et al., 2016). It contains four central subsystems: the action-centered subsystem (ACS), the non-action centered subsystem (NACS), the motivational subsystem (MS), and the metacognitive subsystem (MCS). Within each subsystem, there are explicit (controlled) modules at the top level and implicit (automatic) modules at the bottom level (Sun, 2002, 2016). This leads to the distinction between explicit deliberate symbolic and implicit reactive subsymbolic processing at the two levels, respectively. Due to the dual-representational characteristics, Clarion can model relevant psychological dynamics such as the explicit and implicit attitudes. This has been demonstrated through extensive empirical validation across various psychological domains (e.g., Bretz & Sun, 2018; Hélie & Sun, 2010; Sun et al., 2005).

The present work focuses on the dynamics between three subsystems within Clarion: the MS, the ACS, and the MCS. Below, aspects of Clarion relevant to self-control will be described, omitting details not essential to this topic (see Sun et al., 2001, 2005, 2016; Sun & Mathews, 2012 for these other aspects).

Action-Centered Subsystem (ACS)

The ACS handles action selection in interactions with the world (Sun et al., 2022).

The ACS's perception-action cycle operates as follows (Sun, 2002, 2016): it perceives the current state, then at each level computes possible actions. A utility calculation weighing the costs and benefits of applying the implicit and the explicit level to various extents determines the chosen level of explicitness (such as 0.5 or 0.9, with 1 indicating explicit processes determining behavior and 0 denoting entirely implicit processes). (Explicitness may also be determined by other circumstances as detailed later.) Then the cycle starts again.

Specifically, the utility calculation weighs the costs and benefits of different levels of explicitness (analogous to Sun et al., 2022). Benefits evaluate the likelihood of need satisfaction from outcomes, while costs increase with greater explicitness. Linking utility to motivation enables modeling aspects missing from existing models (cf. Braver et al., 2014).

Utility U_j is calculated (by the MCS; more later) as follows:

$$U_j = benefit_j - \nu \times cost_j \quad (1)$$

where υ is a scaling parameter that balances the benefit and cost values (Sun, 2016). After some algebraic derivation (Sun et al., 2022), the utility becomes:

$$U_j = explicitnesslevel_j \times (\alpha \times value(g) - c) \quad (2)$$

where U_j represents the utility of $explicitnesslevel_j$, α is the benefit coefficient, and c is the cost coefficient. $value(g)$ is determined by:

$$value(g) = \sum_d ds_d \times satisfaction_d(g) \quad (3)$$

where $satisfaction_d(g)$ indicates how well attaining goal g satisfies drive d (more on drives, i.e., motives, later). The summation is over all currently active drives. Thus $value(g)$ can be seen as the overall *satisfaction* of current needs from reaching goal g (Sun, 2016).

A Boltzmann distribution converts the utilities into selection probabilities (Sun et al., 2022).

$$p(j) = \frac{e^{U_j/\tau}}{\sum_i e^{U_i/\tau}} \quad (4)$$

This equation calculates $p(j)$, the probability of choosing $explicitnesslevel_j$. U_i is the utility of $explicitnesslevel_i$, τ represents the "temperature" (stochasticity in selection), and the summation is over all possible explicitness levels. Based on these probabilities, one explicitness level is stochastically chosen for the ACS (by the MCS; more later). Note that, alternatively, level of explicitness may also be dictated by externally or internally induced circumstances (more later; see Bretz & Sun, 2018; Wilson & Sun, 2021).

The bottom level of the ACS contains reactive routines with implicit procedural knowledge, often through trial-and-error learning in neural networks (with subsymbolic representations). It uses a backpropagation neural network to evaluate potential actions and select an action. When an input (including the current state and the goal) is received, the network computes how desirable each action choice is expected to be in the situation (i.e., computes a Q-value). These Q-values are then converted into a Boltzmann distribution (similar to equation 4), so that actions with higher Q-values have a higher chance of being selected. An action is then sampled from this probability distribution and chosen as the output of the bottom level (for more details, see Sun 2016).

Meanwhile, the top level of the ACS contains explicit knowledge in the form of symbolic rules, acquired, for example, through extracting patterns from the bottom level or through external instructions (Sun, 2002). Rules are in a "state, goal → action" form. When an input (the state and the goal) is received, the rules that are applicable in that situation become active candidates for selection. To choose one of these competing rules probabilistically, a Boltzmann distribution is formed based on a support value (default=1)

assigned to each rule (Sun, 2016), similar to how actions are selected stochastically at the bottom level based on Q-values (i.e., similar to equation 4). Once a rule is sampled from the distribution, the action recommended by the rule is set as the top level's action choice.

The overall action of the ACS is determined based on the chosen explicitness level: When the output actions of both the bottom and the top level of the ACS are available, one of them is chosen through a stochastic process (via a probability distribution by the MCS).

Motivational Subsystem (MS)

The MS handles motivation: that is, drives and goals (at the bottom and the top level, respectively), which guide action selection by the ACS toward satisfying internal needs (Sun, 2009; Sun, 2016). Drives represent fundamental motivational forces (motives), including both physiological and social motives (Sun, 2009). The strength ds_d of a drive d is calculated upon receiving inputs from the current state (Sun et al., 2022):

$$ds_d = stimuluslevel_d \times deficit_d \quad (5)$$

where $stimuluslevel_d$ indicates the current state's relevance in activating drive d (state-specific), and $deficit_d$ reflects individual/cultural internal predisposition or tendencies toward activating drive d (agent-specific).

Higher *stimuluslevel* values increase drive activation, goal outcome value, and thus overall utility. *Stimuluslevel* can be manipulated through external means. In contrast, *deficit* parameters capture individual propensities (Sun & Wilson, 2014). With higher *deficit* values, the corresponding drives and associated goals have greater strengths. As detailed later, we view self-control reduction/fatigue (e.g., Wehrt et al., 2020; Inzlitch et al., 2014) as a kind of motivational shift toward intrinsic, automatic, or leisurely choices due to the recalibration of some drive *deficits* after effortful, extrinsic, deliberate tasks. That is, performing demanding self-control tasks can impact subsequent performance and/or control capacity, which is related to the idea of Inzlicht et al. (2014).

Metacognitive Subsystem (MCS)

The MCS regulates action selection (within the ACS) based on drives and other contexts (Sun, 2009, 2016).

First, the MCS maps the drives to goals. Goal strength gs_g for a goal g is calculated as:

$$gs_g = \sum_d relevance_{s,d\to g} \times ds_d \quad (7)$$

Where $relevance_{s,d\to g}$ measures the relevance of drive d to selecting goal g; ds_d is the strength of drive d. The summation is over all drives. A goal, which directs action selection, is then selected based on a Boltzmann distribution of gs_g's (similar to equation 4).

Second, within the MCS, drive strengths can affect the explicitness level in the ACS (as discussed in detail earlier). However, a drive (via changes in its *deficit*) may weaken over time, due to, for example, repeated satisfaction of the drive, thus potentially reducing utility and control capacity (more on this later). We can model an individual's drive *deficit* reduction on a task over time using a simple linear function: we assume drive *deficit* reduction at each time step is a linear function of the total time the individual has spent on that task, when the task involves satisfying the drive (the linear relationship is the simplest possible; cf. Vancouver et al., 2010).

$$dr_d = -a_d \times time \quad (6)$$

where dr_d represents the deficit reduction for drive d, $time$ is a number measuring time spent on the task, a_d is a scaling parameter (which may be a function of how much satisfaction that one received), and the drive strength at a time step is the initial drive strength minus dr_d.

So, in the present work, self-control is viewed largely as the outcome of a metacognitive decision about how explicit or implicit the ACS processes should be (e.g., using utility calculation) or as dictated by other situational factors (mentioned earlier).

Two Simulations

In this section, as examples, we simulate, in Clarion, two cases of self-control involving explicit-implicit conflicts based on two published human experiments (Friese et al., 2008; Schmeichel, 2007) from social psychology. The models presented are coarse-grained, relying on general mechanisms and emphasizing generality over fine details. Moreover, the simulations focus on explaining self-control rather than reproducing every experimental nuance. Furthermore, we focus on capturing performance differences across groups/conditions, especially statistically significant ones (although not all are statistically significant). It does not involve quantitative model fit and does not aim for more fine-grained matching. In addition, parameter adjustments are minimal and justified based on prior work (Brooks et al., 2012; Sun et al., 2022; Sun & Wilson, 2014; Wilson & Sun, 2021).

Self-control and Memory

Human Experiment Schmeichel (2007) tested whether control capacity decreased after self-control exertion. 141 participants, divided into two groups, performed response inhibition (Activity 1) and then memory updating (Activity 2). For Activity 1, one group of participants had to ignore distractor words in a video (the attention control condition) while the other group did not (the no-control condition). Two working memory tasks were carried out in Activity 2 --- an operation span task (Test 1) and a sentence span task (Test 2). In Test 1, 41 participants from the attention control group and 38 from the no-control group evaluated 48 math equations intermixed with target words that needed to be recalled later. These equation-word pairs were arranged into 15 sets of 2-5 items per set. In Test 2, there were 31 participants from each group. Each participant heard 50 sentences, answered comprehension questions about them, and had to recall the

final word of each sentence. The sentences were organized into 15 variable-sized sets containing 2-5 items each. Both tests assessed cognitive control (working memory).

The performance data revealed that attention control participants performed worse on memory tasks (Test 1 M=32.88, Test 2 M=34.87; compared to the no-control participants: Test 1 M=34.68; Test 2 M=38.32). Although the differences were not statistically significant, the no-control group scored higher on both tests, suggesting control capacity reduction in the attention control group. Prior self-control exertion seemed to impair later control (although not statistically significant).

Conceptual Description Schmeichel (2007) showed that self-control exertion impaired later control (working memory in this case). In Clarion, it involves a conflict between explicit and implicit action selection, which is resolved via the utility calculation of levels of explicitness (Berkman et al., 2017; C. Chen et al., 2022; Z. Chen et al., 2020) based on internal motivation (i.e., drive activations; Sun, 2016; Sun et al., 2022), which determines the control capacity. Analogous to motivational shift theory (Inzlicht et al., 2014), the *deficits* of the drives underlying self-control may decrease through repeated satisfaction of the drives during exertion of self-control, thus reducing control subsequently. This drive recalibration captures the control capacity change in Schmeichel's experiment.

Simulation Setup As in the human experiment, 141 simulated participants (each involving the MS, the MCS, and the ACS of Clarion), split into two groups, performed the experimental tasks.

Drive strength was determined by *stimuluslevel* and *deficit* (equation 5). The MCS selected the goal (to participate in the experiment; by equation 7) and stochastically chose the level of explicitness based on utility calculation (equation 2).

The initial *deficit* of the *achievement* drive was 1. Equation 6 determined the *deficit* decrease (for the attention control group a_d= 0.54; for the no-control group a_d= 0.36), with a more significant decrease for the attention control group, leading to more motivational shift. The *stimuluslevel* parameter was 0.7 across conditions since this represented external factors that were constant for both groups. There were, of course, other drives, but they were not crucial to the present simulation. Exact parameter values were not important; the model was robust.

Due to self-control exertion, the attention control group had more *achievement* drive satisfaction in Activity 1 (0.9 for the attention control group and 0.6 for the no-control group). The utility of levels of explicitness was calculated (equation 3; parameter a= 1 and $c = 0.16$) using drive strengths and satisfaction values (the explicitness level was coded so that a higher number represented higher explicitness). A level of explicitness was selected via the Boltzmann distribution (equation 4; temperature=0.8). During Activity 1, explicitness decreased due to motivational shift caused by the *deficit* decreasing. The attention control group became progressively more implicit, ending with mostly bottom-level processes. The no-control group declined less, retaining more explicitness. In Activity 2, the attention control group stayed largely implicit, relying primarily on bottom-level processing (corresponding to diminished control capacity).

Activity 1 was modeled as follows: the bottom level of the ACS used a backpropagation neural network for the video task, with four input nodes (which included experiment instructions as inputs): "watch video freely," "view video disregarding on-screen distractions," "video intensity," and "distraction level of words on screen"), eight hidden nodes and four output nodes ("do nothing," "read words without watching the video," "watch the video and read words," and "watch the video and do not read words"). The network was trained using supervised learning from a synthetic dataset. The network learned to mimic participants' actions using the backpropagation algorithm. Its performance was validated on a separate portion of the synthetic dataset to assess its accuracy.

To simulate the subsequent memory tasks (Activity 2), a single-layer LSTM network (Hochreiter & Schmidhuber, 1997) in the NACS, with 64 LSTM units, connected to a standard dense output layer of four nodes, constituted memory for number sets. The network was trained on many random number sequences. In this way, the network built internal memory representations with a sense of ordinality. When shown a new number sequence to memorize, the model leveraged this learned sense of ordinality to recall the sequence. Noise was included within neural circuitry via a "temperature" parameter of the LSTM, representing a certain degree of randomness in memory (unrelated to explicitness). Due to space, details are omitted.

The top level of the ACS contained rules from task instructions: three for response inhibition and two for memory updating in LSTM. Activation of a rule was based on the current situational inputs and the goal selected. The top level exclusively provided control of LSTM. Following the self-control exertion of the attention control group, due to deficit changes, the ACS shifted towards implicit actions, reducing explicit control to the LSTM. With curtailed control from the explicit top level of the attention control group, the LSTM exhibited deteriorated performance.

In the MCS, a stochastic selection based on utility decided how implicitly or explicitly an action should be decided, which was determined by the strength of the drives and, thus, their *deficits*. Self-control, in this case, could be seen as the effect of motivation (drive activation); therefore, self-control reduction (fatigue) is the product of motivational shift (caused by *deficit* decreasing).

Simulation Results While no statistically significant differences were found, the simulation showed that the attention control group performed worse on memory tasks, suggesting the possibility of a self-control reduction effect in simulation (Test 1 M=31.8; Test 2 M=33.5; compared to the no-control group: Test 1 M=35.1; Test 2 M=38.0), as in the human data (likewise with no statistically significance).

Explicit and Implicit Attitudes in Food Choices

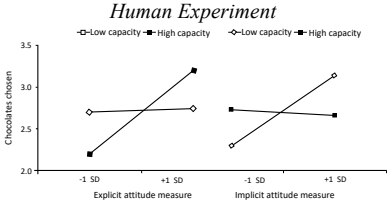
Human Experiment Friese et al. (2008) tested if implicit attitudes guided behavior more under high cognitive load, while explicit attitudes dominated with low load. 85 participants evaluated fruits/chocolates on explicit preference ratings and then performed implicit associations (assessing reaction times when sorting positive or negative stimuli paired with either fruit or chocolate images.) Participants were split into two groups: they memorized either one digit (the high cognitive capacity group, with 42 participants) or eight digits (the low cognitive capacity group, with 43 participants) to manipulate their cognitive capacity. Then, in a choice task, participants selected five between twenty fruits and chocolates. Manipulating cognitive load enabled the modification of the self-control capacity (as shown by the experimental results).

Table 1: Correlations between explicit attitude measure, implicit attitude measure, and the number of chocolates chosen as a function of experimental condition in the human study and the simulation.

	Human Experiment Results			*Simulation Results*		
	1	*2*	*3*	*1*	*2*	*3*
	High cognitive capacity (N = 42)			***High cognitive capacity (N = 42)***		
1.Explicit measure	-	.20	.60*	-	.20	.60*
2.Implicit measure		-	.12		-	.11
3.Chocolates chosen			-			-
	Low cognitive capacity (N = 43)			***Low cognitive capacity (N = 43)***		
1.Explicit measure	-	.37**	.24	-	.36**	.23
2.Implicit measure		-	.45*		-	.45*
3.Chocolates chosen			-			-

Note. *p < .05; **p < .01

In the results of the human experiment, correlation analysis showed a higher correlation of the explicit measure with the dependent variable (number of chocolates chosen) under the high-capacity condition and a higher correlation of the implicit measure under low capacity (see Table 1). Slope test revealed that the explicit measure predicted the high-capacity group's choices. In contrast, the low-capacity group's choices were predicted by the implicit measure (see Figure 1).



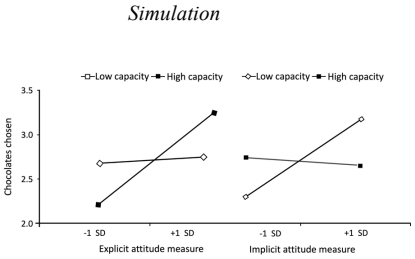


Figure 1: Estimated slopes for the number of chocolates selected as a function of attitude measure (explicit vs. implicit) and cognitive capacity manipulation (low vs. high). The first Figure was taken from Friese et al. (2008) and shows the human experiment results. The second figure represents the simulation results.

Conceptual Description The human experiment directly associated cognitive capacity condition with the level of explicitness. Therefore, the simulation determined explicitness levels through cognitive capacity conditions (rather than through drive deficit reduction as in the previous simulation, due to different experimental settings), because, generally speaking, higher capacity leads to greater explicit processing while lower capacity favors reactive implicit processes (similar to effects of other dual-task distractors, as empirically shown by, e.g., Friese et al., 2008; Lewis & Linder, 1997; Sun et al., 2001; see also Wilson & Sun, 2021). This translated capacity manipulations into explicit versus implicit processing within the ACS.

Simulation Setup As in the human experiment, two groups of 42 and 43 simulated participants represented the two conditions, respectively. Each simulated participant involved the ACS, MS, and MCS of Clarion. Details of drive activation and goal setting were identical to the previous simulation and thus omitted. Because cognitive capacity directly determines explicitness (as shown by Friese et al., 2008; Sun et al., 2001; etc.), the simulation mapped capacity differences to explicitness levels: High-capacity participants were assigned high levels of explicitness (0.7-1.0), engaging more top-level processes; low-capacity participants were assigned low levels of explicitness (0.4-0.6), engaging more bottom-level processes. The specific values were not crucial --- the meaningful aspect was the relative ordering between conditions.

The implicit attitude was simulated through a backpropagation network at the bottom level of the ACS with two input nodes representing the presence of fruits and chocolates, respectively, and two output nodes indicating degrees of preference (e.g., the preference is 20% for fruits and 80% for chocolates). Three hidden nodes propagated input activations to the output nodes. The explicit rating was done using rules at the top level of the ACS for evaluating and responding to situations.

For the choice task, at the top level, rules were generated from the task instructions. Those rules followed the instructions to pick items, with a preference for healthier items (from prior beliefs/knowledge). The choice task also

involved a backpropagation network at the bottom level, with 20 input nodes representing the 20 items (fruits and chocolates), two hidden layers, and an output layer with five output nodes representing the number of selections for each item type. The network had a preference for chocolates.

At the top level of the NACS, rules also coded instructions to memorize one-digit numbers (the high cognitive capacity group) or eight-digit numbers (the low cognitive capacity group). As in the previous simulation, the recall was carried out by an LSTM network. When a top-level rule for memorizing digits was activated, it triggered the LSTM for memorizing sequences. Other simulation details were similar to the previous simulation.

Simulation Results As shown in Table 1, the simulation data corresponded well to the human data. Analysis of the simulated data demonstrated that the explicit measure had a higher correlation with the dependent variable in the high-capacity condition, as in the human data. In contrast, the implicit measure showed a higher correlation in the low-capacity condition, as in the human data. Slope tests also showed results similar to those of the human experiment: the explicit measure strongly predicted choices under the high-capacity condition. The implicit measure strongly predicted low-capacity choices.

Discussion

The model presented in this paper can account for the explicit-implicit conflict within the self-control phenomena. The present paper has examined two tasks: that is, the model has been tested through the simulations of Schmeichel (2007) and Friese et al. (2008). This work shows that the same framework can underlie different phenomena and, in that way, tries to integrate them within that framework. It represents a novel approach to the computational modeling of self-control (in some accordance with some existing theoretical views).

Regarding the first simulation, as Schmeichel (2007) and others (e.g., Inzlicht et al., 2014; Muraven & Baumeister, 2000) demonstrated, exertion of self-control can undermine subsequent control efforts. This aligns with the concept of motivational shift (Inzlicht et al., 2014). In our model, the *deficits* of drives underlying control diminish, thus favoring implicit reaction over effortful explicit thinking, modeled as a shift from explicit to implicit processing after control exertion. Maintaining working memory requires control (Shimamura, 2002; Smith & Jonides, 1999) and thus falters after motivational shift. The Clarion model that underlies the simulation facilitated a unified and mechanistic perspective on empirical findings, connecting the utility calculation of the level of explicitness and the motivational aspects, which mechanistically interpreted such self-control phenomena as an explicit-implicit conflict. This study offers a different explanation of Schmeichel's findings, but it should be considered a preliminary exploration, as it presents a single simulation and serves only as an initial attempt.

Some theories of self-control, such as the ego-depletion theory (e.g., Baumeister & Heatherton, 1996) that views self-control as a limited resource, are not relevant to our simulations, because we did not rely on such theories and instead attempted to provide an alternative, mechanistic interpretation of the empirical data. There are also doubts and skepticism in the literature about the effectiveness of the resource concept in explaining self-control. Research on implicit and explicit attitudes has gained attention during the past decades (Asendorpf et al., 2002; Fazio & Towles-Schwen, 1999; Friese et al., 2006; Strack & Deutsch, 2004).

The second simulation shows that cognitive capacity manipulations may shift reliance on implicit reactivity versus explicit thinking (as documented in the literature and in the cognitive architecture used), which matches the human data of Friese et al. (2008) (see also Bargh, 2002).

Our work is aimed at eventually providing unified explanations and simulations across a wide (if not full) range of self-control phenomena. This focus extends far beyond these two experiments (which will be addressed elsewhere due to space limitations). As Bretz and Sun (2018) confirmed, Clarion can also formally contrast and compare accounts. We are currently working on a much broader range of empirical phenomena and comparisons of different accounts. The model links inputs to outputs through various subsystems. This model connects them into a cohesive framework that accounts for a wide range of empirical data on self-control. Capturing complex phenomena requires precise mechanisms within a broad framework.

This computational model complements empirical research, offering a way of accounting for conflicts between explicit and implicit processes. It represents an initial step toward the needed integration. Future work can further address nuances with regard to effortful "have-to" motivation versus "want-to" motivation or "want-to" actions. Self-control, of course, also involves proactive strategies such as regulation of temptation availability, implementation intentions, and reconstruction that bypasses impulse triggering without requiring inhibition (Baumeister & Heatherton, 1996). By positing a framework of mechanisms, the present model provides a step towards a broad self-control theory.

Acknowledgments

We thank Selmer Bringsjord, Tomek Strzalkowski, and Paul Bello for their helpful comments. The work was conducted while the authors were supported in part by ARI grant W911NF-17-1-0236 and IARPA HIATUS contract 2022-22072200002. Adan Gomez was also supported by a scholarship funded by the Fulbright Program.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological bulletin*, *82*(4), 463.

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of personality and*

social psychology, *83*(2), 380.

Bargh, J. A. (2002). Losing consciousness: Automatic influences on consumer judgment, behavior, and motivation. *Journal of consumer research*, *29*(2), 280-285.

Baumeister, R. F., & Heatherton, T. F. (1996). Self-regulation failure: An overview. *Psychological inquiry*, *7*(1), 1-15.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017). Self-control as value-based choice. *Current directions in psychological science*, *26*(5), 422-428.

Bertelsen, P., Høgh-Olesen, H., & Tønnesvang, J. (2009). *Human characteristics: Evolutionary perspectives on human mind and kind*. Cambridge Scholars Publishing.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in cognitive sciences*, *8*(12), Article 12.

Braver, T. S., Krug, M. K., Chiew, K. S., Kool, W., Westbrook, J. A., Clement, N. J., Adcock, R. A., Barch, D. M., Botvinick, M. M., & Carver, C. S. (2014). Mechanisms of motivation–cognition interaction: Challenges and opportunities. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 443-472.

Bretz, S., & Sun, R. (2018). Two models of moral judgment. *Cognitive science*, *42*, 4-37.

Brooks, J. D., Wilson, N., & Sun, R. (2012). *The effects of performance motivation: A computational exploration of a dynamic decision making task*. 7-14.

Carter, C. S., & Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: An update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), Article 4.

Chen, C., Lin, W., & He, G. (2022). The effect of decision strategy on self-control choice. *Current Psychology*, 1-13.

Chen, Z., Liu, P., Zhang, C., & Feng, T. (2020). Brain morphological dynamics of procrastination: The crucial role of the self-control, emotional, and episodic prospection network. *Cerebral Cortex*, *30*(5), 2834-2853.

De Ridder, D. T., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*(1), 76-99.

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-10301-1

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472-5477. https://doi.org/10.1073/pnas.1818430116

Fazio, R. H., & Towles-Schwen, T. (1999). *The MODE model of attitude-behavior processes.*

Friese, M., & Hofmann, W. (2009). Control me or I will control you: Impulses, trait self-control, and the guidance of behavior. *Journal of Research in Personality*, *43*(5), 795-805.

Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology*, *47*(3), 397-419.

Friese, M., Wänke, M., & Plessner, H. (2006). Implicit consumer preferences and their influence on product choice. *Psychology & Marketing*, *23*(9), 727-740.

Fudenberg, D., & Kreps, D. M. (1993). Learning mixed equilibria. *Games and economic behavior*, *5*(3), 320-367.

Fujita, K. (2008). Seeing the forest beyond the trees: A construal-level approach to self-control. *Social and Personality Psychology Compass*, *2*(3), 1475-1496.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and social psychology review*, *15*(4), Article 4.

Gillebaart, M. (2018). The 'Operational' Definition of Self-Control. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01231

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological review*, *117*(3), 994.

Hoch, S. J., & Loewenstein, G. F. (1991). Time-inconsistent preferences and consumer self-control. *Journal of consumer research*, *17*(4), Article 4.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and self-control from a dual-systems perspective. *Perspectives on psychological science*, *4*(2), 162-176.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in cognitive sciences*, *18*(3), 127-133.

Inzlicht, M., Werner, K. M., Briskin, J. L., & Roberts, B. W. (2021). Integrating Models of Self-Regulation. *Annual Review of Psychology*, *72*(1), 319-345. https://doi.org/10.1146/annurev-psych-061020-105721

Kirby, K. N., & Herrnstein, R. J. (1995). Preference reversals due to myopic discounting of delayed reward. *Psychological science*, *6*(2), 83-89.

Kotabe, H. P., & Hofmann, W. (2015). On integrating the components of self-control. *Perspectives on Psychological Science*, *10*(5), 618-638.

Lewis, B. P., & Linder, D. E. (1997). Thinking about choking? Attentional processes and paradoxical performance. *Personality and social psychology bulletin*, *23*(9), 937-944.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance.* Prentice-Hall, Inc.

Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Luce, R. D. (2014). *Utility of gains and losses: Measurement-theoretical and experimental approaches*.

Psychology Press.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3-19. https://doi.org/10.1037/0033-295X.106.1.3

Milyavskaya, M., Berkman, E. T., & De Ridder, D. T. D. (2019). The many faces of self-control: Tacit assumptions and recommendations to deal with them. *Motivation Science*, *5*(1), 79-85. https://doi.org/10.1037/mot0000108

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological review*, *80*(4), Article 4.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological bulletin*, *126*(2), 247.

Myrseth, K. O. R., & Fishbach, A. (2009). Self-control: A function of knowing when and how to exercise restraint. *Current Directions in Psychological Science*, *18*(4), 247-252.

Neumann, J. von, & Morgenstern, O. (1947). *Theory of games and economic behavior*.

Rachlin, H. (2004). *The science of self-control*. Harvard University Press.

Rachlin, H., & Green, L. (1972). Commitment, choice and self-control 1. *Journal of the experimental analysis of behavior*, *17*(1), 15-22.

Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal of Experimental Psychology: General*, *136*(2), 241-255. https://doi.org/10.1037/0096-3445.136.2.241

Scholz, C., Chan, H.-Y., Poldrack, R. A., de Ridder, D. T. D., Smidts, A., & van der Laan, L. N. (2022). Can we have a second helping? A preregistered direct replication study on the neurobiological mechanisms underlying self-control. *Human Brain Mapping*, *43*(16), 4995-5016. https://doi.org/10.1002/hbm.26065

Seijts, G. H., & Latham, G. P. (2001). The effect of distal learning, outcome, and proximal goals on a moderately complex task. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *22*(3), 291-307.

Shimamura, A. P. (2002). Memory retrieval and executive control. *Principles of frontal lobe function*, *210*.

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science*, *283*(5408), 1657-1661.

Steel, P., & König, C. J. (2006). Integrating theories of motivation. *Academy of management review*, *31*(4), 889-913.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and social psychology review*, *8*(3), Article 3.

Strotz, R. H. (1973). *Myopia and inconsistency in dynamic utility maximization*. Springer.

Sun, R. (2002). *Duality of the Mind: A bottom-up approach toward cognition. Mahwah, NJ: L*.

Sun, R. (2009). Motivational representations within a computational cognitive architecture. *Cognitive Computation*, *1*, 91-103.

Sun, R. (2016). *Anatomy of the mind: Exploring psychological mechanisms and processes with the Clarion cognitive architecture*. Oxford University Press.

Sun, R., Bugrov, S., & Dai, D. (2022). A unified framework for interpreting a range of motivation-performance phenomena. *Cognitive Systems Research*, *71*, 24-40.

Sun, R., & Mathews, R. C. (2012). Implicit cognition, emotion, and meta-cognitive control. *Mind & Society*, *11*(1), Article 1.

Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive science*, *25*(2), Article 2.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological review*, *112*(1), 159.

Sun, R., & Wilson, N. (2014). A model of personality should be a cognitive architecture itself. *Cognitive Systems Research*, *29*, 1-30.

Sun, R., Wilson, N., & Lynch, M. (2016). Emotion: A unified mechanistic interpretation from a cognitive architecture. *Cognitive Computation*, *8*, 1-14.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Wehrt, W., Casper, A., & Sonnentag, S. (2020). Beyond depletion: Daily self-control motivation as an explanation of self-control failure at work. *Journal of Organizational Behavior*, *41*(9), 931-947.

Wilson, N. R., & Sun, R. (2021). A mechanistic account of stress-induced performance degradation. *Cognitive Computation*, *13*, 207-227.