

Running Head: INTENTION, EMOTION, ACTION

Intention, Emotion, and Action: A Neural Theory Based on Semantic Pointers

Tobias Schröder*, Terrence C. Stewart, and Paul Thagard

University of Waterloo, Canada

Schröder, T., Stewart, T. C., & Thagard, P. (forthcoming). Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive Science*.

Word count: 13,087

Acknowledgements:

Order of authorship is alphabetical, as the authors contributed equally. Tobias Schröder was awarded a research fellowship by the Deutsche Forschungsgemeinschaft (SCHR 1282/1-1) to support this work. Paul Thagard's work is supported by the Natural Sciences and Engineering Research Council of Canada. We would like to thank Chris Eliasmith, Zhu Jing, and anonymous reviewers for comments on an earlier version of the manuscript.

***Address correspondence to:**

Dr. Tobias Schröder
University of Waterloo
Department of Philosophy
200 University Avenue West
Waterloo Ontario Canada N2L 3G1
E-Mail: mail@tschroeder.eu

Abstract

We propose a unified theory of intentions as neural processes that integrate representations of states of affairs, actions, and emotional evaluation. We show how this theory provides answers to philosophical questions about the concept of intention, psychological questions about human behavior, computational questions about the relations between belief and action, and neuroscientific questions about how the brain produces actions. Our theory of intention ties together biologically plausible mechanisms for belief, planning, and motor control. The computational feasibility of these mechanisms is shown by a model that simulates psychologically important cases of intention.

Keywords: Intention, Emotion, Action, Implementation Intentions, Automatic, Deliberative, Planning, Neural Engineering Framework, Semantic Pointers

Intention, Emotion, and Action: A Neural Theory Based on Semantic Pointers

1. The Problem of Explaining Intention

The concept of intention is important in many disciplines, including philosophy, psychology, artificial intelligence, cognitive neuroscience, and law. For example, criminal law treats cases where one person intends to kill another very differently from cases where death results unintentionally from negligence. Despite decades of discussions, however, there is no received theory of intention within any of these disciplines, let alone a theory that accounts for all the phenomena identified across all of the disciplines.

We propose a unified theory of intentions as neural processes that integrate representations of states of affairs, actions, and emotional evaluation. We will show how this theory provides answers to philosophical questions about the concept of intention, psychological questions about human behavior, computational questions about the relations between belief and action, and neuroscientific questions about how the brain produces actions. Our theory of intention ties together biologically plausible mechanisms for belief, planning, and motor control. The computational feasibility of these mechanisms is shown by a model that simulates psychologically important cases of intention. These simulations support the plausibility of the claim that human intentions are neurocomputational processes operating in the brains of individuals. Our theory has implications for many vexed issues in the cognitive sciences, such as the nature of the relation between automatic and deliberate processes.

Intention has been an important topic of philosophical discussion since the 1950s (Anscombe, 1957; Bratman 1987; Mele 2009; Setiya; 2010; Ford, Hornsby, and Stoutland,

2011). Debates have concerned questions such as the following. What are intentions? What is the relation between intentions and other mental entities such as beliefs, desires, plans, and commitments? Are intentions causes of actions, or just reasons for actions? What is the relation among intentions about future actions and intentions that are part of actions in progress? What is the difference between intentional and unintentional actions? Why do people sometimes fail to act on their intentions through weakness of will (*akrasia*)? The nature of intention and its relation to action are central to discussions of whether people have free will and whether they should be held responsible for their actions.

Psychologists have been concerned with more practical questions such as how intentions can affect people's behavior in practices such as voting, safe sex, healthy nutrition, and public transport. By far the most influential approach has been the theory of planned behavior of Fishbein and Ajzen (1975, 2010), according to which behaviors result from intentions, which result from a combination of attitudes, subjective norms, and perceived behavioral control, as shown in Fig. 1. This approach, however, is based largely on correlations among empirical measures of beliefs, attitudes, and intentions, and provides no account of the psychological or neural mechanisms by which beliefs and attitudes cause intentions. It also does not specify how intentions cause and sometimes fail to cause behavior. Psychologists use the term "intention-action gaps" for the class of intention failures that philosophers call weakness of will. The psychology of self-control studies the cognitive processes and strategies that help people to reduce intention-action gaps (Baumeister & Tierney, 2011). One such strategy is the use of implementation intentions, i.e. sets of rules that connect anticipated cues in specific situations with previously made commitments to certain behavioral choices (Gollwitzer, 1999).

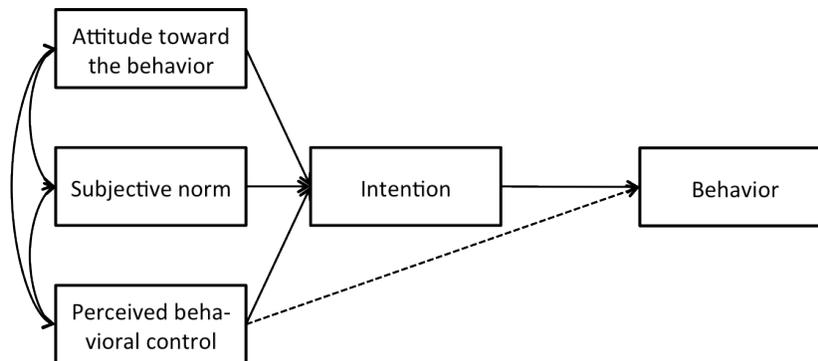


Fig. 1. Theory of planned behavior. Adapted from Fig. 1 in Ajzen (1991, p. 182).

Intention has also become an important topic in cognitive neuroscience, originating with Benjamin Libet’s (1985) controversial claims about the relation of conscious intentions to actions. Subsequent work has concerned use of brain imaging to identify human intentions (e.g., Cunnington, Windischberger, Robinson, & Moser, 2006; Haynes, Sakaai, Rees, Gilbert, Frith, & Passingham, 2007). Work with non-human primates has investigated the relation between intentions in frontal and parietal areas and sensorimotor control (Andersen and Cui, 2009). Understanding intentions is an important part of building neural prosthetics to aid paralyzed patients (Andersen, Hwang, & Mulliken, 2010). However, there has yet to appear a theory of neural processing that can account for the results of neuroscientific experiments concerning intention.

In artificial intelligence, intention has been an important part of attempts to program computers as intelligent agents (e.g. Wooldridge, 2000). Following Bratman (1987), these AI researchers take intentions to be desires to which an agent has become committed as part of a plan. In robotics, investigators have considered how an observer robot infers the intention of a partner to choose a complementary action sequence (Bicho, Louro, and Erlhagen, 2010).

The concept of intention is also central to investigations into legal liability and moral responsibility (Moore, 2009). Actions are considered to be more wrongful if they result from intention rather than negligence or recklessness. Legal scholars are becoming increasingly worried about the challenge posed by neuroscientific findings to the folk understanding of intentions as the result of free decisions. Resolution of this issue requires theoretical understanding of the causes and effects of intentions.

This paper proposes a new neural theory of intention as a brain process that binds together information about situations, emotional evaluations, actions, and sometimes also about the self. We argue that intentions are *semantic pointers*, a powerful kind of neural process proposed by Chris Eliasmith (in press; Eliasmith et al., 2012). The next section outlines the basic claims that we want to make about intentions as semantic pointers, which are subsequently fleshed out using a computational model of how intentions can lead to action. This model is implemented in a computer program that simulates central cases of how intentions sometimes cause actions and sometimes fail to cause actions. Finally, a concluding discussion shows the relevance of this theory and model for issues in psychology and philosophy.

2. Outline of a Neural Theory of Intention

We want to defend the following theoretical claims:

1. Intentions are semantic pointers, which are patterns of activity in populations of spiking neurons that function as compressed representations by binding together other patterns.
2. Specifically, intentions bind representations of situations, emotional evaluations of situations, the doing of actions, and sometimes the self.

3. Intentions can cause actions because of neural processes that connect semantic pointers with motor instructions.
4. Intentions can fail to cause actions because of various kinds of disruptions affecting any of:
 - (a) Evaluation of the situation and doing.
 - (b) Binding of the evaluation, situation, and doing.
 - (c) Processes that connect the intention semantic pointer with motor processes.

Each of these claims requires exposition.

2.1. Semantic Pointers

First we need to say more about the nature of semantic pointers. According to Eliasmith (in press), semantic pointers are patterns of neural firing activity whose structure is a consequence of information compression operations implemented in neural connections. The term “pointer” comes from computer science where it refers to a kind of data structure that gets its value from a machine address to which it points. Semantic pointers thus provide representations of other representations, but those representations are compressed, analogous to JPEG picture files or iTunes audio files, which do not encode the full available information. Neural compression operations bind semantic pointers into complex symbol-like structures. Semantic pointers can be decomposed into the underlying representational structures, thereby enabling the cognitive system to control flows of information across different modalities. For understanding how intentions cause actions, the decompression operation is crucial, since it specifies how high-level symbolic representations set off the low-level motor representations that ultimately govern physical actions (see also Schröder & Thagard, 2013). In Eliasmith’s (in press) terms, semantic pointers connect *shallow semantics* with *deep semantics*. Shallow semantics are given through

symbol-like relations to the world and other representations, while *deep semantics* are constituted by relations to perceptual, motor, and emotional information.

The semantic pointer idea can be understood as a computational specification of various well-known theories that have posited symbolic/sensory connections in human cognitive systems. For example, Barsalou (1999) claims that symbols are higher-level representations of combined perceptual components extracted from lower-level sensorimotor experience. Similarly, the mental models of Johnson-Laird (1983) can be understood as multimodal data structures ultimately grounded in semantic primitives like emotional and kinesthetic representations. Lakoff and Johnson (1980) view cognitive processes as driven by complex conceptual metaphors composed of basic metaphors like affection= warmth that are rooted in ubiquitous sensorimotor experience and thus shared among humans across cultures. Osgood and colleagues have shown that the metaphorical structure of concepts can be described with three universal dimensions representing the basic sensory and emotional experiences of approach vs. avoidance, power/control, and activity/arousal (e.g., Heise, 2010; Osgood, May, & Miron, 1975).

We accordingly conjecture that intentions are high-level cognitive phenomena that model configurations of lower-level representations in multiple modalities. When bound together, they can cause action through routing semantic information to the motor system. Fig. 2 elucidates how we think this works: Intentions are semantic pointers, i.e. patterns of spiking activity which bind together neural representations of situations and their evaluation along with actions and sometimes the self. All of these components are semantic pointers, i.e. patterns of spiking activity on their own. The binding operation relies on neural pattern transitions embedded in the connection weights between the respective populations of neurons. Bindings of semantic pointers are recursive. Therefore, the semantic pointer idea provides a way of reconciling connectionist

accounts of distributed representations with more hierarchical and rule-based perspectives on the control of action (cf. Botvinick & Plaut, 2006; Cooper & Shallice, 2006). Our theory of intentions as semantic pointers thus applies to cases where there are behavioral plans that can be decomposed into smaller component actions (Miller, Galanter, & Pribram, 1960).

Other kinds of mental representations can also be understood as semantic pointers that bind together different sorts of information: intentions are semantic pointers but not all semantic pointers are intentions. Concepts bind together information about examples, prototypical features, and explanatory rules (Blouw, Solodkin, Eliasmith, and Thagard, forthcoming). Emotions bind together cognitive appraisals and physiological perceptions (Thagard and Schröder, forthcoming; Thagard and Stewart, 2011). The priming of behavior requires binding cued concepts with information concerning situations, the self, other persons, and emotions (Schröder and Thagard, 2013).

We thus propose that intentions are a special instance of a general cognitive process whereby a representation emerges from binding other representations. The subsequent section elaborates on the elements we consider crucial for the operation of intentions.

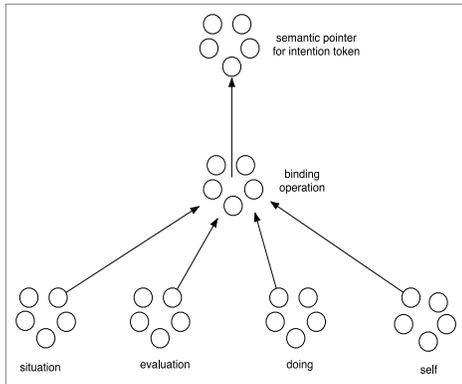


Fig. 2. How intentions are formed by binding representations of a situation, evaluation, doing, and self. The sets of circles indicate neural populations. The arrows indicate flow of information performed by neural firing.

2.2. Components of Intention

Representations of situations include the physical features of the current environment, processed primarily through visual areas of the brain, but sometimes also by olfactory, auditory, and tactile areas. These basic representations are constraints on the formation of intentions. For example, one may want to help a child trapped in a house on fire, but hold back because entrances are inaccessible. There are also important symbolic aspects about how we represent situations, as the choice of behaviors in situations is equally strongly constrained by culturally shared knowledge about identities and social institutions (MacKinnon & Heise, 2010). For example, one would easily recognize the presence of firefighters by visual cues (uniforms, fire trucks, equipment) and immediately know that it is their responsibility, not one's own, to rescue the child in danger. Representations of situations are thus complex compounds of physical as well as symbolic features of the environment (i.e., deep and shallow semantics). The semantic pointer

architecture provides a set of mathematical principles stating how the integration of such different representations can be achieved in populations of spiking neurons (Eliasmith, in press).

Humans constantly evaluate situations with the emotion system of the brain, and we believe these evaluations to be an important building block of intentions. Brain areas with a prominent role in processing emotional evaluation include (but are not limited to) the amygdala, insula, ventromedial prefrontal cortex, and the nucleus accumbens (for reviews, see Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Thagard & Aubie, 2008). The emotion system mirrors the hierarchical nature of cognition, with more basic and ubiquitous emotions like anger and fear more tied to immediate sensorimotor experience, and more complex and culturally shaped emotions like guilt and shame of a more symbolic nature. Extending this analogy, we have applied the semantic pointer idea to emotion elsewhere (Thagard & Schröder, forthcoming).

Emotional evaluations of situations vary along a continuum of more automatic/implicit vs. deliberative/explicit appraisal (Cunningham & Zelazo, 2007). Most representations of symbolic concepts elicit spontaneous affective evaluations that reflect common cultural knowledge (Heise, 2010; Osgood et al., 1975). Elsewhere, we have argued that those affective meanings of concepts play a major role in behavioral priming, where subtle cues in the environment cause people to align their behaviors automatically and without conscious awareness (Schröder & Thagard, 2013; cf. Bargh, 2006; Bargh & Chartrand, 1999). However, people might also deliberately choose to ignore automatic emotional associations as a source of information for their judgments, if they conflict with consciously endorsed goals and values (Gawronski & Bodenhausen, 2007). In current psychological theorizing, such dissociations between implicit and explicit evaluations play a major role in explaining intention-action gaps. For example, one might intend to quit smoking or excessive eating, as one actively evaluates

these behaviors as bad for one's health, but nevertheless have implicit positive representations of these behaviors. Especially under limitations of cognitive resources, the implicit positive attitudes defeat the explicit negative ones, causing a failure to implement intentions (Chassin, Presson, Sherman, Seo, & Macy, 2010; Friese, Hofmann, & Wänke, 2008; Hofmann & Friese, 2008; Hofmann, Gschwendner, Friese, Wiers, & Schmitt, 2008; Ward & Mann, 2000). This kind of contest is consistent with the proposal by Norman and Shallice (1986) that actions under conscious control involve a competitive mechanism in addition to those used in automatic actions.

Intentions also require representations of the intended actions themselves. We understand them not just as linguistic descriptions but also as patterns of activation in areas of the brain involved in processing motor instructions. Neuroscientific evidence corroborates the notion of a non-verbal "action vocabulary" in pre-motor cortex, consisting of abstract representations of underlying motor programs in relation to goals (Fogassi, 2011; Gallese, 2009; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). The analogy to semantic pointers as compressed models of deeper sensorimotor representations is straightforward (see DeWolf & Eliasmith, 2011, on motor control within the semantic pointer architecture), and it is just another step up in the hierarchy of the cognitive system to a symbolic representation of actions with language. Indeed, there is abundant empirical evidence for the priming of verbal concepts to facilitate mental simulations of movements (e.g., Springer & Prinz, 2010) as well as action itself (for review, see Bargh, 2006). The semantic pointer idea provides a mechanistic explanation of the neural processes underlying those priming effects (Schröder & Thagard, 2013).

Finally, we believe that intentions sometimes involve representations of the self, on occasions when people explicitly think of themselves as planning to do something. Intentions are

about one's own actions in specific situations. Self-representations are semantic pointers that result from binding together self-related information in various modalities, from abstract verbal characterizations such as *professor* to the associated emotional meanings to kinesthetic representations such as *swinging a golf club*. The resulting dynamic neural process theory of the self reconciles conflicting philosophical views such as Kantian unified consciousness and Humean non-unified bundles of perceptions (Thagard, in press).

Some of the components of self-representations are self-concepts, emotional memories, and the sensorimotor experience of agency. Self-concepts are linguistic labels that people apply to themselves. In so doing, they make use of culturally constructed categories, crystallized in language, to make sense of themselves and their social experiences (MacKinnon & Heise, 2010). Through binding representations of past emotional episodes into the current self-representation, people experience a sense of continuing coherence of their affective states. At the core of self-representations lies a sense of agency, which results from "intentional binding" of afferent motor information with efferent perceptual input (Tsakiris & Haggard, 2004). As a result, individuals experience themselves as causes of changes in their environments. Perceived agency goes along with characteristic shifts in time perception: Subjects who believe that they caused a tone through pressing a button voluntarily judge the time elapsed between action and tone to be shorter than subjects who knew that their pressing the button was caused by transcranial magnetic stimulation (Haggard, Clark, & Kalogeras, 2002). Such effects can be interpreted as experimental evidence for binding of efferent and afferent information to underlie the sense of agency. We conjecture that this process provides the basis for the representation of self. The result of efferent-afferent binding is a semantic pointer that can be stored in memory and later be retrieved and itself recursively bound into a different higher-level semantic pointer. Thus,

previous sensorimotor experiences of agency form the basis for later inclusion of the self in complex intentions.

3. A Neurocomputational Model of Intention

To develop our theory of intention further, we now present a computational model of interacting neurons that yields simulations of important psychological phenomena. We use the Neural Engineering Framework of Eliasmith & Anderson (2003) to simulate flows of current in different, interconnected populations of neurons. All neurons are modeled as standard Leaky Integrate-and-Fire neurons that receive current from input neurons, integrate these inputs with some loss, and produce as outputs firing behaviors that send current to other neurons.

Mathematical details are explained in Eliasmith & Anderson (2003) and in the appendix below.

The model consists of six different groups of interacting neurons, meant to represent six different brain areas: sensory cortex, prefrontal cortex, the basal ganglia, the amygdala, anterior cingulate cortex, and the supplementary motor area. The connections among these areas are shown in Fig. 3, consistent with neural anatomy. The model is loosely based on Tsakiris and Haggard's (2010) review of the neural structures underlying the control of intentional action. It is also compatible with Cunningham and Zelazo's (2007) iterative cycle of evaluative reprocessing, a neuroanatomical model of the interplay of automatic (implicit) vs. deliberate (explicit) evaluation of situations. We will see that the automatic/deliberate distinction is crucial to psychological understanding of why intentions sometimes fail to produce actions. We acknowledge that our model is extremely simplified, and we do not claim to explain the neural data that support the relevance of these structures to intention and action. Our model is consistent

with these data, but the empirical support for the model comes primarily from the simulation of the results of psychological experiments in section 4.

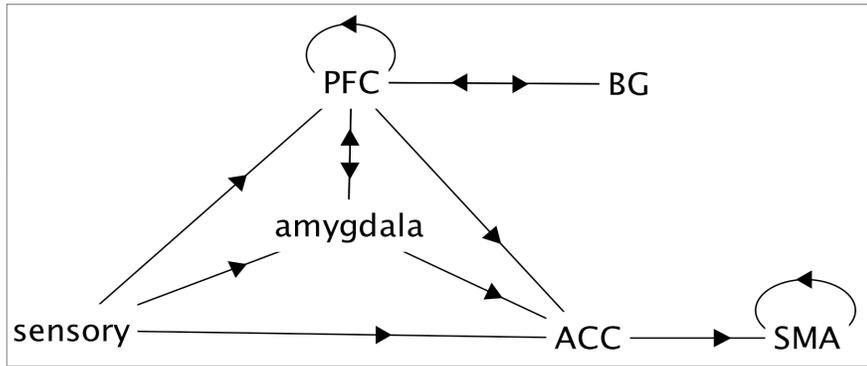


Fig. 3: Functional components of the model of intention, consisting of six groups of neurons and synaptic connections shown by arrows. Abbreviations: PFC - prefrontal cortex; BG - basal ganglia; ACC - anterior cingulate cortex; and SMA - supplementary motor area.

The input to the model is entirely through sensory cortex, where we trigger different patterns of firing for the different stimuli that can be given to the model. Output is from the supplementary motor area (SMA), where different patterns of neural firing represent the different actions the model can take. Taking SMA as the output structure of our model is consistent with research on readiness potentials: Activation over the SMA, measured with EEG, correlates with participants reporting a felt “urge” to start an action. This has been interpreted as the neural process underlying phenomenological intentions (Libet, 1985; see Tsakiris & Haggard, 2010, for review). All of the firing patterns in the components of our model are randomly initiated.

In order to have the model perform complex tasks, we need to manipulate these patterns internally. To do this, we treat each pattern of firing as a different semantic pointer, allowing us to define computations to combine and extract information from these patterns. In particular, our

model relies on neural pattern transitions: creating synaptic connections between two groups of neurons such that if a particular pattern is part of the activity in the first group, then the second group of neurons will be driven to some other pattern of activity. This allows us to transform and manipulate semantic pointers using pattern transitions: for example, we may say that if the pattern for “the letter A” is in the sensory system, then we want the pattern for “press button 1” to appear in the ACC. We may also combine different input patterns (e.g., semantic pointers for sensory input and for emotional evaluation) to produce an output pattern (e.g., semantic pointer for intention). Once we have defined what pattern transitions we want, we use the Neural Engineering Framework to calculate the optimal synaptic connection weights to give us those transitions (Eliasmith & Anderson, 2003). Mathematical details are outlined in the appendix below.

We also define a few fixed sets of connections regardless of the pattern transition rules. For the prefrontal cortex and the supplementary motor area, we include feedback connections that cause these neurons to maintain whatever pattern they are currently producing. This feedback provides a memory (since a pattern can be maintained even if the input is removed), and gives a gradual transition between patterns (i.e. if there is an input, the pattern will slowly change to match that desired pattern). This allows us to store an arbitrary semantic pointer over time.

For the connection between the anterior cingulate cortex (ACC) and the supplementary motor area, we combine the pattern in the ACC with the pattern in the amygdala. The pattern in the amygdala models the *value* of the current action. The stronger this value, the more the SMA will be driven to store whatever pattern is in the SMA. This preference allows the model to

quickly perform actions if they are thought to be very good, and even to decide not to do an action if it realizes it would have a low value.

Finally, the basal ganglia area allows the model to choose one action out of a list of possible actions. The neural connections for this group are more complex, using an existing selection model of the basal ganglia (Stewart, Bekolay, & Eliasmith, 2012). This model follows a similar process of having rules that map one pattern onto another pattern, but forces only one rule to be active at a time. This is responsible for providing a “serial bottleneck” to cognition, and has been used to model complex cognitive tasks such as solving the Tower of Hanoi problem (Stewart & Eliasmith, 2011).

4. Simulations

A neurocomputational model of intention should apply to a wide array of phenomena that have not previously been connected. On the one hand, social psychology has treated intentions as high-level symbolic phenomena involving planning for the future, without caring about the details of implementation in the brain (e.g., Fishbein & Ajzen, 1975, 2010). On the other hand, intention-related work in cognitive neuroscience has predominantly dealt with low-level tasks like moving hands or fingers or adding numbers in present situations (e.g., Cunnington et al., 2006; Haynes et al., 2007; Libet, 1985). We believe that semantic pointers allow us to bridge this gap, resulting in computational models that address both high-level and low-level accounts of intention. To demonstrate this, we now present a series of five simulations, starting with a simple model and adding to it, resulting in a single model that accounts for five different types of intentional activity.

First, we simulate an experiment where the participants were expected to intentionally choose one among six specific finger gestures to produce while their brain activity was recorded with fMRI (Cunnington et al., 2006). The simulation involves causing an action by connecting a representation of the situation with a representation of doing. The second simulation additionally involves emotional evaluation. We model a situation where a person drinks alcoholic beverages at a party after forming the deliberate intention to do so, which results from favorable attitudes and social norms towards drinking (Fishbein & Ajzen, 2010; Glindemann, Geller, and Ludwig, 1996). The third simulation deals with a dissociation between automatic and deliberative emotional evaluation (Cunningham & Zelazo, 2007; Deutsch & Strack, 2006). We model how a person initially feels inclined to smoke a cigarette but then refrains from it because of the deliberate intention to quit smoking due to negative health effects. Fourth, we simulate how intentions can fail when cognitive load prevents the deliberative pathway from interrupting an initial affective action tendency (e.g., Friese et al., 2008; Hofmann & Friese, 2008; Ward & Mann, 2000). Finally, we show how neural representations can be combined, stored in a semantic pointer, and replayed later to produce actions. This simulation models implementation intentions, a special case of planning and future intentions that have been effective as a strategy in psychotherapy to overcome intention-action gaps (Gollwitzer, 1999).

To create this model, we use a software package called Nengo that generates neural networks in accord with the Neural Engineering Framework (<http://www.nengo.ca>). These simulations are very different from conventional connectionist models using hand-coded localist representations or distributed representations produced by training. Instead, networks are produced analytically by specifying neural populations and the mathematical functions that they are required to compute. Details as to how to represent patterns using spiking neurons and how

to compute the connection weights required to connect these neurons so as to perform the functions described below are provided in the appendix. This results in a model using 11,648 spiking neurons in total. Since these neurons are organized to represent and transform semantic pointers in general (rather than particular patterns of activity), the model can respond appropriately to a widely varying range of stimuli, rather than being restricted to those representations that it was trained on.

Using semantic pointers within the Neural Engineering Framework provides an approach to understanding the relation between representation and behavior that is intermediate between explicit goal and schema representations (Cooper and Shallice, 2006) and distributed representations in recurrent networks (Botvinick and Plaut, 2006). Semantic pointers are fully distributed across a neural population, but the following simulations show how distributed representations can function much like symbols. To demonstrate the behavior of the models over time, we show the spiking output of different groups of neurons, along with an indication of the semantic pointer that mostly closely matches the current firing pattern of those neurons. For example, in Fig. 4 we show just the sensory system of our model as we change the input to be the randomly chosen semantic pointers for “A”, “B”, and then “A” again. The pattern of firing activity for each semantic pointer is different, but interestingly the overall average firing rate across the population is similar for each one. Every semantic pointer will have its own unique firing pattern.

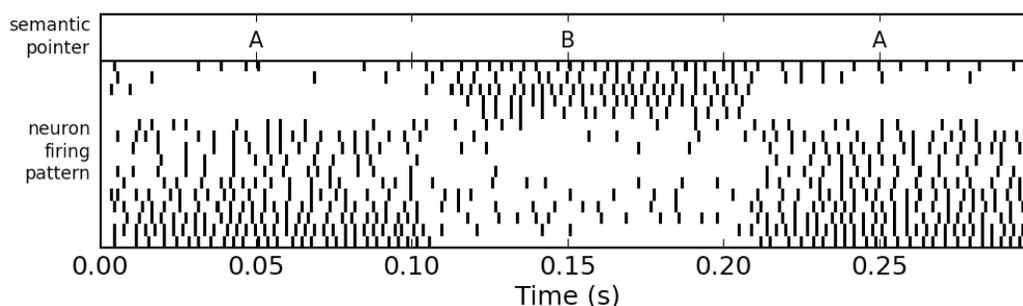


Fig. 4: Neural response of 16 sensory neurons (see Fig. 3) representing the randomly generated semantic pointers “A” and “B”. The box for neuron firing pattern has 16 rows, one for each neuron. A mark in a row indicates that the neuron is firing at a particular time. The neurons have some random variability, but distinct overall patterns correspond to distinct semantic pointers.

A crucial feature of these semantic pointer models is that we can build models that are generic across semantic pointers. That is, we can create a neural model that will, for example, pass a semantic pointer from one population to another, and this will work *even for semantic pointers that it has never seen before*. That is, the model is not limited to a particular small set of patterns of activity that it is “trained” on. Rather, we use the Neural Engineering Framework to find a set of connection weights that will reliably transfer information for any possible semantic pointer. This feature is vital to the following simulations, since at each stage we add new semantic pointers for new conditions.

4.1. Simulation 1: Motor Intentions

Our first simulation is based on the free choice task from Cunnington et al. (2006). In this task, certain stimuli are paired with certain actions (in the original study, hand gestures from American Sign Language). For example, if the subjects see , they must respond in kind by

making the same gesture . Similarly, if they see , they must respond with . However, when shown a special stimulus (in our simulations, a question mark [?]), the subject must *choose* to respond with either gesture. This is meant to show the neural difference between a free choice and a forced response: more neural activity is seen in the pre-frontal cortex (PFC) and basal ganglia (BG) when making a free choice than in the forced condition (Cunnington et al., 1996, p. 1297).

We implement this task in our model by defining semantic pointers for each stimulus (, , and ?) and each response ( and ). These can be arbitrarily complex combined representations of the visual stimulus and the motor commands needed to create these gestures. Since a full model of this process would require a complete model of the human visual and motor systems (and thus be well outside the scope of this paper), we select an arbitrary firing pattern for each stimulus (shown in the top row of Fig. 5) and each motor action (shown in the bottom row of Fig. 5). It should be noted that, as expected, the firing pattern for the visual stimulus  is quite dissimilar from the motor command needed to generate the same gesture (Fig. 5, left-most column, top and bottom row).

Once these semantic pointers are defined, we need to construct the neural connections that will cause the model to perform as desired. For the forced actions, this is done by forming connections between the sensory area and the ACC that implement the desired pattern transitions. In particular, we add the transition rules “visual()→motor()” and “visual()→motor()”. That is, we use the Neural Engineering Framework (Eliasmith & Anderson, 2003) to create neural connections between the sensory and ACC areas such that if the semantic pointer in the sensory system contains the visual representation of , the neurons for

the corresponding pattern in ACC will be stimulated (and the same for ). For mathematical details, see the appendix.

To implement the choice behavior, we add further neural connections. First, between sensory and pre-frontal cortex (PFC) we add “? \rightarrow ?”, so that the fact that we have to make a choice is transferred to PFC. Then in the basal ganglia (BG) we add the two neural transition rules “? \rightarrow ” and “? \rightarrow ”. Thus, if the “?” is shown to the sensory system, a corresponding semantic pointer will be transferred to PFC. In turn, this will stimulate the BG neurons to drive the PFC to initiate either  or  (randomly chosen based on noise in the neural representation). Finally, we add transition rules between PFC and ACC that simply transfer the patterns: “ \rightarrow ” and “ \rightarrow ”. This scenario does not use the amygdala, since none of these patterns have an associated emotional value representation.

The resulting behavior is shown in Fig. 5, displaying the firing activity for 128 neurons in each of the three brain areas relevant to this task (sensory, PFC, and ACC). The different patterns of activity represent different stimuli (sensory) and actions (PFC and ACC). For each brain area and time interval, the degree of firing of each of the neurons is shown by dark shading. For example, the row for sensory neurons shows how they each fire (or fail to fire) in response to different sensor stimuli. Activity in the other areas is entirely driven by synaptic connections as discussed. Notice that when the model sees a  or a  (top row), it accurately produces the appropriate output pattern (bottom row). Furthermore, when shown a ?, it will produce one of the two possible patterns. We note that the PFC is only strongly active when it is making a free choice. This behavior of the model is compatible with the fMRI data from Cunnington et al. (2006).

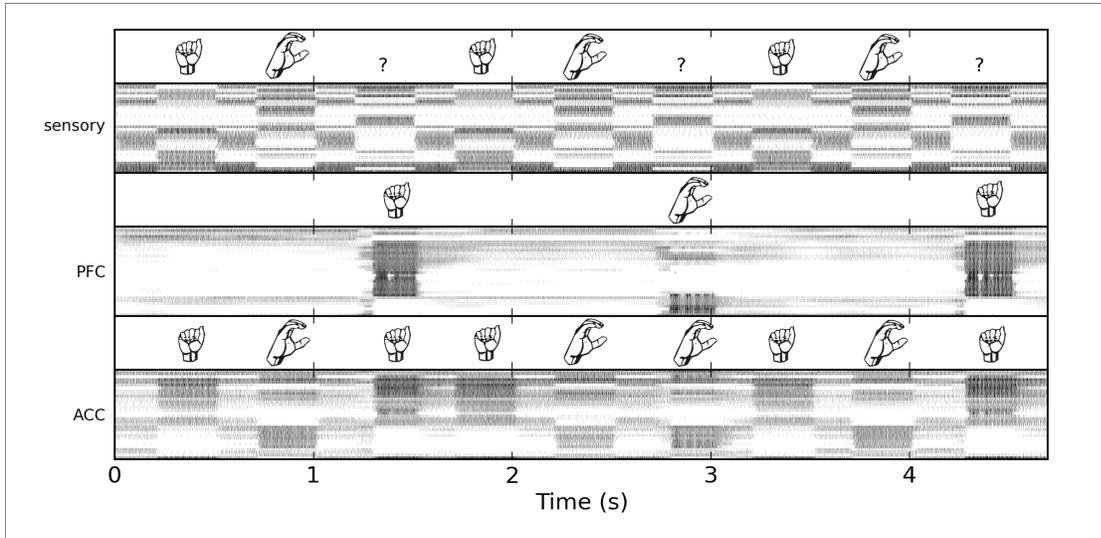


Fig. 5: Behavior of the model when performing the free choice task. When shown a , the model responds with the motor pattern for . When shown a , the model responds with the motor pattern for . When shown a , the model chooses either  or  (via the PFC), and then performs that action.

4.2. Simulation 2: Intentions involving Emotional Evaluation

For our second example, we examine a social situation that includes emotional evaluation. For this task, we assume that the action produced by the automatic direct behavior pathway (the connection between sensory cortex and ACC) is in accord with the deliberative pathway (the connection via PFC). To match the situation from a study by Glindemann et al. (1996), we consider a situation where the subject is offered a drink and acts autonomously.

To control this behavior, we add pattern transition rules to the model. These are new transformations in addition to those rules considered in the previous simulation. Since these are implemented as semantic pointers in the NEF, we can use the NEF to adjust the existing synaptic connections to implement these new rules as well, rather than creating entirely new connections

for each rule. From sensory to PFC we add a rule DRINK→DRINK, which simply passes the pattern for the DRINK semantic pointer into working memory. We also add a rule OFFER→TAKE between sensory and ACC, representing a standard default action of taking something if it is offered. This corresponds to a social norm (Fishbein & Ajzen, 2010). Importantly, since semantic pointers can be combined, we can now provide a single sensory input of “OFFER+DRINK” and this combined pattern of neural activity will correctly trigger the two separate rules DRINK→DRINK and OFFER→TAKE.

For this simulation, we must also consider the behavior of the amygdala and SMA. Connections from the sensory cortex and PFC are configured so that both follow the transition rule “DRINK→GOOD”. Fig. 6 illustrates the simulation. The patterns for OFFER and DRINK are both presented at $t=0.2s$. This presentation results in the PFC getting the pattern for DRINK, which is evaluated in the amygdala as GOOD. This can be seen in the chart by the change in neural activity in the amygdala around 0.25s. This evaluation allows the automatically chosen action TAKE to be quickly passed to the SMA (by $t\approx 0.3s$), which would then trigger the appropriate response.

The overall idea, then, is that when offered something (represented by presenting the sum of the patterns for OFFER and DRINK to the sensory area), the default action is to take it. This does not require cognitive effort (i.e. it does not require the deliberative activity of the PFC). However, in this case the PFC is in agreement with the automatic pathway and increases the strength of the pattern being sent to SMA, resulting in a fast decision to take the drink.

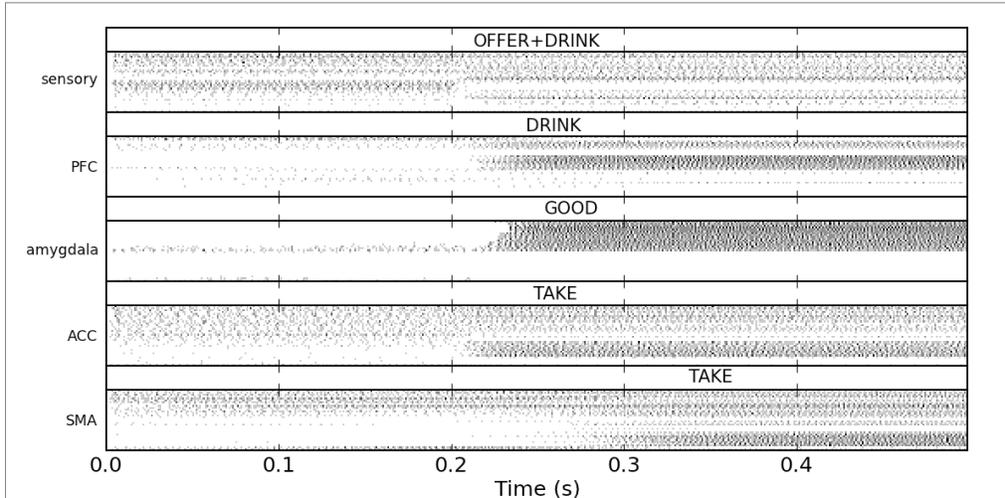


Fig. 6: Behavior of the model when the automatic and deliberative pathways for emotional evaluation are in accord. For each brain area such as PFC, the chart shows spiking of each of 128 neurons: darker means more spiking.

4.3. Simulation 3: Intentions Override Affective Action Tendencies

The third simulation (Fig. 7) considers a situation where the deliberative pathway overrides the automatic pathway. In this example, the subject is offered a cigarette. We model this by presenting both the patterns for OFFER and SMOKE to sensory at $t=0.2s$. As before, the automatic pathway will perform its default action to TAKE the cigarette. The nature of semantic pointers is such that the combined semantic pointer OFFER+SMOKE will trigger exactly the same activity in ACC as was seen in the previous simulation, even though the spiking activity OFFER+DRINK is different from the spiking activity for OFFER+SMOKE. In this case, however, at the same time the pattern for SMOKE will be passed to working memory (PFC), rather than the pattern for DRINK as in the previous case. The basal ganglia have a transition rule for SMOKE→UNHEALTHY (representing explicit knowledge), and there is a transition rule between PFC and the amygdala for UNHEALTHY→BAD, overriding the initial evaluation

of SMOKE as GOOD (at $t=0.25s$ in the amygdala). The presence of this negative evaluation stops the TAKE action from being passed from the ACC to the SMA, thus preventing the action from occurring. This prevention is an instance of successful self-control (cf. Baumeister & Tierney, 2011; Vohs & Baumeister, 2010).

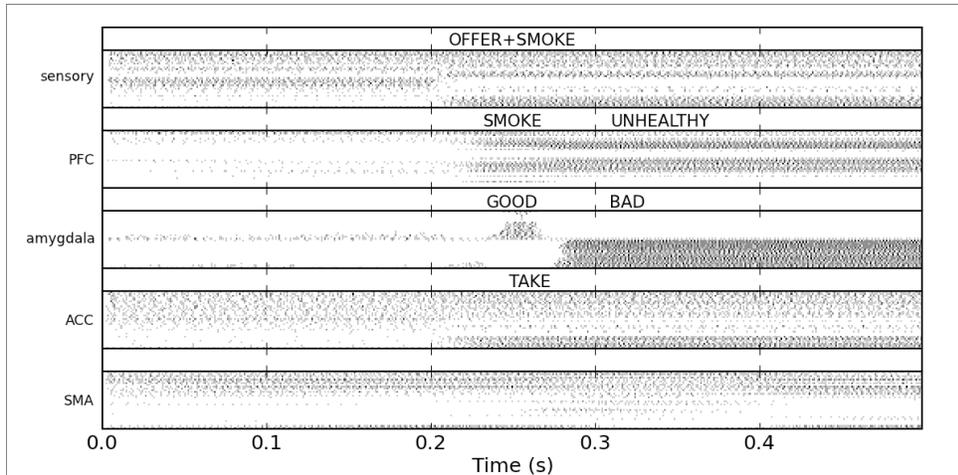


Fig. 7: Behavior of the model when the automatic and deliberative pathways are not in accord.

4.4. Simulation 4: When Intentions Fail

We next consider the case where there is a heavy cognitive load that stops the deliberative pathway from overriding the automatic pathway (e.g., Frieese et al., 2008). Here, we add a transition rule for the PFC back to itself (via the basal ganglia) that says $WORK \rightarrow WORK$. Once the PFC contains the pattern for WORK, it will continue thinking about work. We now continue with exactly the same stimulus as in simulation 3. In this case, however, when OFFER+SMOKE is presented to the sensory cortex, the pattern for SMOKE will not be successfully transferred to PFC (or at least it will be much weaker than the pattern for WORK). This, in turn, will mean that the deliberative pathway will not pass its evaluation on to the amygdala and ACC, and so the

automatic TAKE action will occur. Hence a subject who is distracted by thinking about other things will not follow through on the intention to avoid smoking. This result is shown in Fig. 8.

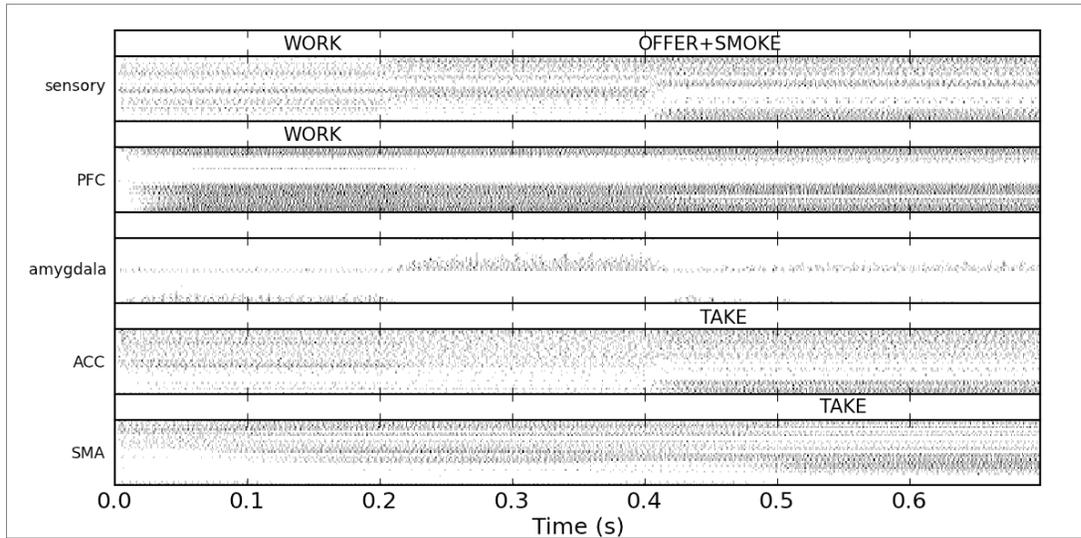


Fig. 8: Behavior of the model when the automatic and deliberative pathways are not in accord, but the deliberative pathway is busy. For the first 0.2s, the pattern for WORK is presented. This locks the PFC into the pattern for WORK. We now present (at $t=0.4s$) the pattern for OFFER+SMOKE. Since the PFC is busy, it is unable to interrupt the automatic pathway as it could in Fig. 7. As a result, the TAKE action is selected.

4.5 Simulation 5: Implementation Intentions

Finally, we turn to a case where neural representations must be combined, stored, and replayed when appropriate. As discussed in section 2.1, it is possible to combine the representations in different parts of the brain into a single semantic pointer. Furthermore, this compressed representation can also be split back apart, re-stimulating an approximation of the original neural state. Simulation 5 shows how to model future intentions, and makes explicit the role that

semantic pointers can play in producing actions. In particular, the semantic pointer binds together many different representations in different parts of the brain, producing a new pattern: a single compact representation. This new pattern can be stored and recalled efficiently, allowing the brain to recreate an approximation of a previous mental state.

We use this capability to model implementation intentions, which are cognitive rules that take an environmental cue and turn it into a commitment to a particular course of action (Gollwitzer, 1999). Consider someone who wants to form an intention to not smoke when offered a cigarette. Importantly, since an implementation intention is based on sensory input (the environmental cue), then this should succeed even if the individual is currently distracted thinking about other things, as in Simulation 4. Instead of relying on PFC to follow the reasoning SMOKE→UNHEALTHY, here the model relies on a stored semantic pointer that can be triggered to recreate the original intention to not smoke. The new “memory” component for storing and replaying this compressed representation is shown in Fig. 9.

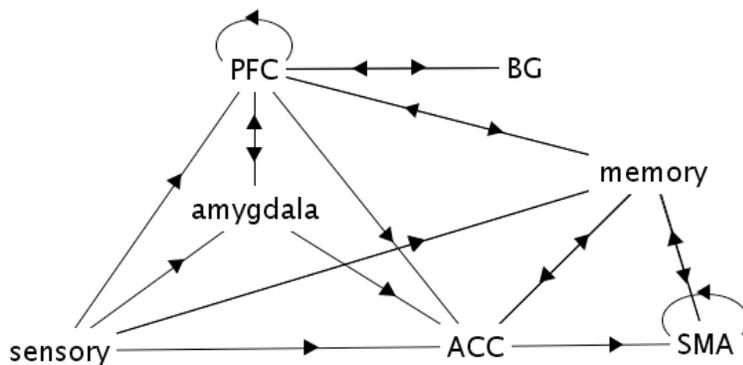


Fig. 9: The model extended for implementation intentions. The memory system combines representations from different cortical areas (as per Fig. 2), and reconstructs the original pattern when triggered by a sensory cue of OFFER+SMOKE.

It is important to remember that we can create this memory component to work for *any* semantic pointer. That is, we can use the NEF to find connection weights to and from the memory that work without explicit training on the data to be compressed and decompressed. This is a key advantage of semantic pointers: since they are built up via a compression and decompression process, we can build neural systems that correctly function for any input values, allowing the intention system to create new implementation intentions and apply them without retraining or adjusting the connection weights in the rest of the model.

As with Simulation 4, we test the model by first presenting it with the sensory stimulus for WORK. This is passed to the PFC and simulates the heavy cognitive load that caused the intention in Simulation 4 to fail. In this extended simulation, however, we have added to the memory a semantic pointer representation of the global pattern of neural activity from Simulation 3 (in which the intention was successful). Now, when the OFFER+SMOKE stimulus occurs, that memory is decompressed, pushing the spiking patterns of the PFC, ACC, and SMA back to successful patterns from Fig. 7. Our model thus explains why implementation intentions can be an effective strategy to reduce intention-action gaps: Semantic pointers allow brains to divert the cognitively demanding intentional decision-making process to a point in time prior to the critical situation.

5. Discussion

Our model is compatible with current theorizing in psychology of the relationship between intention and action. We propose that it is a computational specification of contemporary views of action control as resulting from interactive competition between at least two different ways of

processing information: deliberative (reflective, explicit, controlled, system 2) vs. automatic (impulsive, implicit, unconscious, system 1) (e.g., Cunningham & Zelazo, 2007; Deutsch & Strack, 2006; Fazio & Towles-Schwenn, 1999; Kahneman, 2011; Lieberman, 2003; Norman & Shallice, 1986; Smith & DeCoster, 2000; Strack & Deutsch, 2004). The theory of planned behavior is the most influential psychological account of deliberative intentional action (see Fig. 1; Ajzen, 1991; Fishbein & Ajzen, 1975; 2010). It is largely compatible with philosophically influential views of the function of intentions for planning and coordination (Bratman, 1987). We first discuss the relations of our model to this perspective, before we turn to the contrasting vision of action as controlled by automatic, implicit processes. As demonstrated in our simulation 4, dissociations between the two systems of action control can explain instances of intention-action gaps, called weakness of will or *akrasia* in philosophy.

The theory of planned behavior (TPB) has been applied widely, mostly in contexts where psychology is used to change people's behaviors in ways deemed desirable by governments, action groups, marketers, doctors, or other stakeholders (for review, see Fishbein & Ajzen, 2010). The theory is conceptually similar to the belief-desire-intention model of action control, influential in philosophy and artificial intelligence (e.g., Bratman, 1987; Woolridge, 2000). Fishbein and Ajzen posit that actions follow from behavioral intentions. In turn, attitudes toward a behavior, resulting from beliefs about its expected outcome combined with the value (\approx desire) of that outcome, predict intentions. However, as in Bratman's (1987) model, beliefs and desires (i.e., attitudes) are not sufficient to form a commitment to an action (see Fig. 1). Perceived social norms, reflecting the anticipated reaction of significant others, and perceived behavioral control, reflecting a subjective assessment of whether one is able to carry out the action, are the two additional components.

Despite its influence, the TPB has important conceptual limitations, as it leaves open what intentions actually are. Fishbein and Ajzen construe intentions as “the subjective probability of performing a behavior” (Fishbein & Ajzen, 2010, p. 40). They call for the empirical operationalization of an intention to be as close as possible to the behavior itself in order to enable predictive success (they call this matching “levels of generality”; Fishbein & Ajzen, 2010, p. 30). The problem with this definition is that intentions are not conceptually different from the corresponding actions, and therefore it is hard to argue that intentions cause actions (Greve, 2001). In contrast, we argue that intentions are semantic pointers, i.e. neural processes emerging from binding different representations, and we showed in simulations how intentions as semantic pointers can cause actions by routing information to the motor areas of the brain. We also showed how the semantic pointer hypothesis of cognition enables us to relate the conceptual components of high-level theories like the TPB and the similar belief-desire-intention model to neural processes. For example, in our simulation 2, we implemented Fishbein and Ajzen’s concept of social norms as transition patterns between neural populations. A pattern of neural activity representing someone offering a drink at a party caused the emergence of another firing pattern representing the action of taking the drink. Hence, we showed in principle how social norms can be embedded in the connection weights between neural populations. Similarly, the neural representations of situations and emotional evaluations are required for the beliefs and desires, respectively, in philosophical theorizing about intentions. Intentions can contribute to planning, as argued by Bratman (1987), because the semantic pointers that we take to constitute intentions are fully capable of participating in the partial, hierarchical, and conduct-controlling mental states that Bratman describes. Intentions include a kind of commitment not found in either beliefs or desires because they require binding together the representations of situations in

sensory and prefrontal cortices and emotional evaluation in the amygdala with links to action shown by the involvement of the supplemental motor area.

The TPB also has empirical limitations. Meta-analytic reviews of empirical studies under the TPB paradigm revealed that behavioral intentions roughly account for between a fourth and a third of the variance in actual behaviors – the predictive success is higher when self-reports of behaviors are used as criterion variables, and lower for objective measures (Armitage & Conner, 2001; Shepperd, Hartwick, & Warshaw, 1988). For the standards of social science, predictive accuracy of that size is certainly notable and makes the theory a suitable framework in many applied contexts. However, it is also apparent that the TPB is far from providing a complete picture of the intention-action relationship since two thirds or more of behavioral variance remain open to further inquiry.

These limitations are unsurprising in light of abundant empirical studies that have demonstrated behavior to be controlled by automatic, unconscious processes rather than deliberative decision-making. For example, studies under the influential behavioral priming paradigm have demonstrated how people's actions are often biased by the mere cognitive activation of concepts through cues in the environment (for reviews, see Bargh, 2006; Bargh & Chartrand, 1999). At first sight, this perspective on behavioral control differs sharply from any approach that emphasizes the role of deliberative intentions, but the neural mechanisms underlying both forms of action generation appear to be surprisingly similar. Elsewhere, we have proposed a neurocomputational model of automatic social behavior, which is also based on semantic pointers and whose architecture overlaps with the present model of intention (Schröder & Thagard, 2013). Based on the theory that all concepts are grounded in culturally shared affective meanings (Heise, 2010; Osgood et al., 1975), we have argued that behavioral priming

effects occur because primed concepts automatically elicit specific evaluations in the affective networks of the brain, which, in turn, activate representations of emotionally congruent actions. This process was modeled in the same way as the automatic pathway in the present model of intention, with primed concepts and related behaviors implemented as semantic pointers in the sensory and supplemental motor area networks, respectively; the amygdala and anterior cingulate cortex provided the connections (Schröder & Thagard, 2013).

The essential difference between the two models is that the intention model has an additional deliberative pathway, consisting of the prefrontal cortex and basal ganglia. In our simulation 3, we showed how intentions operate as semantic pointers in prefrontal cortex, binding underlying representations in ways that interrupt and change impulsive action tendencies by overriding the initial emotional evaluation of the action. This cortico-limbic feedback loop is compatible with Cunningham and Zelazo's (2007) iterative reprocessing model of evaluation, based on a review of the neural structures that may underlie the fundamental dichotomy between impulsive and intentional control of action. The dynamic competition of automatic and deliberative action control in the brain is currently the most widely believed psychological explanation for the frequent failure of intentions to produce actions. In our simulation 4, we showed accordingly how affect-driven action tendencies win over intentional choices when working memory capacity is limited, in line with evidence from psychological studies on health-related behaviors (Chassin et al., 2010; Friese et al., 2008; Hofmann & Friese, 2008; Hofmann et al., 2008; Ward & Mann, 2000).

Similarly, our model can readily explain procrastination, an important psychological phenomenon where people delay working on their tasks despite their deliberate commitment to get those tasks accomplished (for review, see Steel, 2007). It was shown that procrastination is

caused by the aversiveness of the task in question itself along with a cognitive inability to override the resulting negative affect with more positive evaluations that stem from the goals associated with finishing the task (e.g., Ferrari, 2001; Onwuegbuzie & Collins, 2001; Steel, 2007). This behavior is exactly the reverse of what happens in our simulations 3 and 4, where the immediate, impulsive emotional evaluation of the action (smoking a cigarette) was positive and needed to be replaced by more negative appraisals of the long-term consequences of the action. In the case of procrastination, the initial negative affect associated with the task needs to be replaced with more positive appraisals of the long-term consequences of tackling the task, and this requires cognitive effort and capacity.

To summarize, our model presents a detailed hypothesis about the neural mechanisms that may underlie the control of action according to recent social psychological theories, contributing to the new field of social neuroscience (Todorov, Fiske, & Prentice, 2011). We think that Eliasmith's (in press) semantic pointer hypothesis and the computational tools that implement it provide a framework for going beyond purely data-driven research in social neuroscience. Rather than merely correlating brain areas to psychological functions, we described neurocomputational mechanisms that plausibly cause psychological phenomena. Moreover, we have shown how automatic and deliberate processes can interact. It is important to note that our model does not assume qualitatively distinct mechanisms for these processes, but rather, the competition between implicit and explicit aspects of action control emerges from the dynamical binding and feedback mechanisms of semantic pointers within the same information-processing system. Hence our approach is compatible with the view that the automatic-deliberative dichotomy is more phenomenological than based on two clearly distinguishable

systems in the brain (cf. Cunningham & Zelazo, 2007; Kruglanski & Thompson, 1999; Newell & Shanks, in press).

Our methodological approach to the nature of intentions contrasts with the usual philosophical one of analyzing the everyday concept of intention by attention to how people talk about their intentions and other mental states. Instead, we look at robust phenomena about intention revealed by controlled experiments in psychology and neuroscience, and seek to explain these phenomena by describing neural mechanisms that can produce these phenomena. The connection between the postulated mechanisms and the phenomena to be explained is shown by the development of a computational model that employs the proposed mechanisms to simulate the phenomena of interest (Thagard, 2012b, ch. 1).

Our hypothesis that intentions are semantic pointers may seem rather audacious given the currently limited extent of knowledge about how brains carry out complex mental tasks. The procedure we have employed is increasingly fruitful in cognitive science and operates as follows. First, identify an important mental phenomenon such as the ways in which intentions can lead and fail to lead to behavior. Second, use what is known about brain operations to form conjectures about the kinds of representations and processes that might produce the phenomena, for example semantic pointers and their associated neural operations. Third, spell out these conjectures with sufficient rigor that they can be implemented in computer simulations, as we have done using the Nengo simulation software. Fourth, determine whether the computer simulations match the behavior of people in psychological experiments, as we have done in 5 cases. Fifth, argue that the mechanisms specified provide the best available explanation of the mental phenomena, which justifies the tentative identification of a familiar mental process (intention) with a novel neural process (semantic pointers). Of course, like all theoretical claims

in science, the proposed identification is fallible and may be found wanting either because there are important phenomena for which it cannot account or because better theories come along. The procedure for identifying mental processes with neural processes is no different from the many cases in the history of science where everyday notions become understood scientifically through their identification with newly proposed mechanisms; for example, fire is rapid oxidation and electricity is the flow of electrons (Thagard, forthcoming). Philosophical arguments that mental states cannot be identified with neural processes are dealt with in Thagard (2010).

We have argued that intentions are patterns of activity in populations of spiking neurons that function as compressed representations by binding together representations of situations, emotional evaluations of situations, the doing of actions, and the self. This account provides an answer to the central puzzle addressed by Anscombe (1957) of how the same concept of intention can apply to different forms such as intentions for the future and current intentional actions. On our view, what such cases have in common is the same underlying neural mechanisms involving representations of situations, evaluations, doings, and the self. Due to the recursive nature of semantic pointers, current intentions can be combined with anticipated cognitive cues of future situations, stored in memory and later retrieved as in simulation 5, where we modeled Gollwitzer's (1999) implementation intentions.

There has been much debate about the nature of shared intentions (e.g. Alonso, 2009; Tomasello, 2008). From a neurocomputational perspective, the question of whether two people can have the same intention is no different from whether they have in common other mental states such as beliefs, desires, and sensory experiences. In all these cases, sameness cannot mean having identical patterns of neural activity, because no two people have

exactly the same neural connections or sensory inputs. Nevertheless, most people's brains have much commonality in structure and process, and people in similar circumstances can have functionally similar semantic pointers that bind together neural representations of situations, evaluations, and actions that have much in common across different people. In such cases, it makes sense to talk loosely and metaphorically of shared intentions.

By far the most contentious philosophical issue connected with the nature of intention concerns the existence of free will, a topic important for ethics because of the common view that moral and legal responsibility require free action. Some neuroscientists and psychologists have argued that empirical findings make it implausible that free will exists (e.g., Harris, 2012; Libet, 1985, 2004; Wegner, 2003). Dualist philosophers reject these claims out of hand, but even some non-dualists such as Dennett (2003) and Mele (2009) argue for conceptions of free will that they think are compatible with increased neuropsychological understanding of mental causation. All of these debates have taken place without any specification of the neural mechanisms that plausibly link intention and action. Our model of intention has strong implications for questions about free will and responsibility, but these will receive extended discussion elsewhere.

6. Conclusion

This paper has developed the first detailed neurocomputational account of how intentions and emotional evaluations can lead to action. We have proposed that actions result from neural processing in brain areas that include the basal ganglia, prefrontal cortex, anterior cingulate cortex, and supplementary motor area. Undoubtedly there are interactions with other brain areas, for example the mid-brain dopamine system that is also important for emotional evaluations (Litt, Eliasmith, and Thagard, 2008; see also Lindquist et al., 2012). Nevertheless, we have

shown by simulations that a simple model can account for intention-action effects ranging from gesturing to failing to act to anticipating future situations. The new model illuminates psychological issues about the relations between automatic and deliberative control of action, and helps to answer philosophical questions about the nature of intention. The result, we hope, is support for our theory that intentions are semantic pointers that bind together representations of situations, emotional evaluations of situations, the doing of actions, and the self. This account serves to unify philosophical, psychological, neuroscientific, and computational concerns about intentions.

We have made extensive use of Eliasmith's new idea of semantic pointers, which we think is useful for general issues about cognitive architecture and more specific issues about intention and action, as well as for computational modeling. For several decades, there has been ongoing debate between advocates of symbolic, rule-based cognitive architectures and advocates of neural network architectures (for a survey, see Thagard 2012a). Eliasmith's Semantic Pointer Architecture provides a new synthesis that shows how sufficiently complex neural networks can process symbols while retaining embodied information concerning sensory and motor processes, with applications that range from image recognition to reasoning. This synthesis is very helpful for understanding how intention-action couplings can operate with both verbal representations and sensory-motor ones. Our computer simulations, especially the fifth one concerning implementation intentions, show how neural representations can be combined, stored, and replayed. The theory of semantic pointers shows how intentions can bind together representations of situations, emotions, actions, and the self in ways that explain how intentions can both lead and fail to lead to behavior.

Of course, much remains to be done. There are numerous psychological and neural experiments about intention that we have not yet attempted to simulate, and undoubtedly a richer neurological account would introduce more brain areas and connections. We have only scratched the surface in discussing the philosophical ramifications of neural accounts of intention and action, and completely neglected the potential implications for robotics. Nevertheless, we hope that a specific proposal for empirically plausible brain mechanisms that link intention, emotional evaluation, and action will contribute to theoretical progress.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179-211.
- Alonso, F. M. (2009). Shared intention, reliance, and interpersonal obligations. *Ethics*, 119, 444-475.
- Andersen, R. A., & Cui, H. (2009). Intention, action planning, and decision making in parietal-frontal circuits. *Neuron*, 63, 568-583.
- Andersen, R. A., Hwang, E. J., & Mulliken, G. H. (2010). Cognitive neural prosthetics. *Annual Review of Psychology*, 61, 169-190.
- Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.
- Armitage, C. J. & Conner, M. (2001). Efficacy of the theory of planned behavior: A meta-analytic review. *British Journal of Social Psychology*, 40, 471-499.
- Blouw, P., Solodkin, E., Eliasmith, C., & Thagard, P. (forthcoming). Concepts as semantic pointers: A theory and computational model. *Unpublished manuscript, University of Waterloo*.
- Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, 36, 147-168.
- Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54, 462-479.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.
- Baumeister, R., & Tierney, J. (2011). *Willpower: rediscovering the greatest human strength*. New York: Penguin Press.
- Bicho, E., Louro, L., & Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction. *Frontiers in Neurorobotics*, 4.
- Botvinick, M. M., & Plaut, D. C. (2006). Such stuff as habits are made on: A reply to Cooper and Shallice (2006). *Psychological Review*, 113, 917-928.
- Bratman, M. E. (1987). *Intention, plans, and practical action*. Cambridge, MA: Harvard University Press.
- Chassin, L., Presson, C. C., Sherman, S. J., Seo, D.-C., & Macy, J. T. (2010). Implicit and explicit attitudes predict smoking cessation: Moderating effects of experienced failure to control smoking and plans to quit. *Psychology of Addictive Behaviors*, 24, 670-679.
- Cooper, R. P., & Shallice, T. (2006). Hierarchical goals and schemas in the control of sequential behavior. *Psychological Review*, 113, 887-916.
- Cunningham, W. A. & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, 11, 97-104.
- Cunnington, R., Windischberger, C., Robinson, S., & Moser, E. (2006). The selection of intended actions and the observation of others' actions: A time-resolved fMRI study. *NeuroImage*, 29, 1294-1302.
- Dennett, D. (2003). *Freedom evolves*. New York: Penguin.
- DeWolf, T., & Eliasmith, C. (2011). The neural optimal control hierarchy for motor control. *The Journal of Neural Engineering*, 8, 21.

- Deutsch, R. & Strack, F. (2006). Duality models in social psychology: From dual processes to interacting systems. *Psychological Inquiry*, 17, 166-172.
- Eliasmith, C. (in press). *How to build a brain: A neural architecture for biological cognition*. New York: Oxford University Press.
- Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338, 1202-1205.
- Eliasmith, C. & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge: MIT Press.
- Fazio, R. H. & Towles-Schwenn, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual Process Theories in Social Psychology* (pp. 97-116). New York: Guilford.
- Ferrari, J. R. (2001). Procrastination as self-regulation failure of performance: Effects of cognitive load, self-awareness, and time limits on “working best under pressure.” *European Journal of Personality*, 15, 391-406.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading: Addison-Wesley.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York: Psychology Press (Taylor & Francis).
- Fogassi, L. (2011). The mirror neuron system: How cognitive functions emerge from motor organization. *Journal of Economic Behavior & Organization*, 77, 66-75.
- Ford, A., Hornsby, J., & Stoutland, F. (Eds.). (2011). *Essays on Anscombe's Intention*. Cambridge, MA: Harvard University Press.
- 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 behavior. *British Journal of Social Psychology*, 47, 397-419.
- Gallese, V. (2009). Motor abstraction: A neuroscientific account of how action goals and intentions are mapped and understood. *Psychological Research*, 73, 486-498.
- Gawronski, B. & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition*, 25, 687-717.
- Georgopolous, A.P., Schwartz, A., & Kettner, R.E. (1986). Neuronal population coding of movement direction. *Science*, 233, 1416-1419.
- Glindemann, K. E., Geller, E. S., & Ludwig, T. D. (1996). Behavioral intentions and blood alcohol concentration: A relationship for prevention intervention. *Journal of Alcohol and Drug Education*, 41, 120-134.
- Gollwitzer, P. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, 54, 493-503.
- Greve, W. (2001). Traps and gaps in action explanation: Theoretical problems of a psychology of human action. *Psychological Review*, 108, 435-451.
- Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5, 382-385.
- Harris, S. (2012). *Free will*. New York: Free Press.
- Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the brain. *Current Biology*, 17, 323-328.

- Heise, D. R. (2010). *Surveying cultures: Discovering Shared Conceptions and Sentiments*. Hoboken: Wiley.
- Hofmann, W., Gschwendner, T., Friese, M., Wiers, R. W., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: Toward an individual difference perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology*, *95*, 962-977.
- Hofmann, W. & Friese, M. (2008). Impulses got the better of me: Alcohol moderates the impact of implicit attitudes toward food cues on eating behavior. *Journal of Abnormal Psychology*, *117*, 420-427.
- Isokawa, M. (1997). Membrane time constant as a tool to assess cell degeneration. *Brain Research Protocols*, *1*(2), 114-116. doi:10.1016/S1385-299X(96)00016-5
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard Press.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.
- Kruglanski, A. W., & Thompson, E. P. (1999). Persuasion by a single route: A view from the Unimodel. *Psychological Inquiry*, *10*, 83-109.
- Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. New York, NY: Oxford University Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*, 529-566.
- Libet, B. (2004). *Mind time*. Cambridge, MA: Harvard University Press.
- Lieberman, M. D. (2003). Reflexive and reflective judgment processes: A social cognitive neuroscience approach. In J. P. Forgas, K. D. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44-67). Cambridge, England: Cambridge University Press.
- Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, *35*, 121-143.
- Litt, A., Eliasmith, C., & Thagard, P. (2008). Neural affective decision theory: Choices, brains, and emotions. *Cognitive Systems Research*, *9*, 252-273.
- MacKinnon, N. J. & Heise, D. R. (2010). *Self, identity, and social institutions*. New York: Palgrave Macmillan.
- Mele, A. R. (2009). *Effective intentions*. Oxford: Oxford University Press.
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Henry Holt & Co.
- Moore, M. S. (2009). *Causation and responsibility*. Oxford: Oxford University Press.
- Newell, B. R., & Shanks, D. R. (in press). Unconscious influences on decision-making: A critical review. *Behavioral and Brain Sciences*.
- Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1-18). New York: Plenum Press.
- Onwuegbuzie, A. J., & Collins, K. M. (2001). Writing apprehension and academic procrastination among graduate students. *Perceptual and Motor Skills*, *92*, 560-562.
- Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3, 131-141.
- Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review*, 120, 255-280.
- Setiya, K. (2010). Intention. *Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/entries/intention/>
- Shepperd, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, 15, 325-342.
- Smith, E. R. & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, 4, 108-131.
- Springer, A. & Prinz, W. (2010). Action semantics modulate action prediction. *The Quarterly Journal of Experimental Psychology*, 63, 2141 – 2158.
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133, 65-94.
- Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Frontiers in Decision Neuroscience*, 6.
- Stewart, T. C., & Eliasmith, C. (2011). Neural cognitive modeling: A biologically constrained spiking neuron model of the Tower of Hanoi task. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *33rd Annual Conference of the Cognitive Science Society*.
- Strack, F. & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220-247.
- Thagard, P. (2010). *The brain and the meaning of life*. Princeton: Princeton University Press.
- Thagard, P. (2012a). Cognitive architectures. In K. Frankish & W. Ramsay (Eds.), *The Cambridge handbook of cognitive science* (pp. 50-70). Cambridge: Cambridge University Press.
- Thagard, P. (2012b). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.
- Thagard, P. (forthcoming). Explanatory identities and conceptual change. *Unpublished manuscript, University of Waterloo*.
- Thagard, P. (in press). The self as a system of multilevel interacting mechanisms. *Philosophical Psychology*.
- Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition*, 17, 811-834.
- Thagard, P., & Schröder, T. (forthcoming). Emotions as semantic pointers: Constructive neural mechanisms. In L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotions*. New York: Guilford.
- Thagard, P., & Stewart, T. C. (2011). The AHA! experience: Creativity through emergent binding in neural networks. *Cognitive Science*, 35, 1-33.
- Todorov, A. B., Fiske, S. T., & Prentice, D. A. (2011). *Social neuroscience: Toward understanding the underpinnings of the social mind*. New York: Oxford University Press.
- Tomasello, M. (2008). *Origins of human communication*. Cambridge: MIT Press.

- Tsakiris, M. & Haggard, P. (2010). Neural, functional, and phenomenological signatures of intentional actions. In F. Grammont, D. Legrand, & P. Livet (Eds.), *Naturalizing intention in action* (pp. 39-64). Cambridge: MIT Press.
- Vohs, K. D., & Baumeister, R. F. (Eds.) (2010). *Handbook of self-regulation, 2nd edition: Research, theory, and applications*. New York: Guilford.
- Ward, A. & Mann, T. (2000). Disinhibited eating under cognitive load. *Journal of Personality and Social Psychology*, 78, 753-763.
- Wegner, D. M. (2003). *The illusion of conscious will*. Cambridge, MA: MIT Press.
- Wooldridge, M. (2000). *Reasoning about intelligent agents*. Cambridge, MA: MIT Press.

Appendix: Neural Modelling

To construct the computational models shown in this paper, we make use of the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003). In this approach, we specify a type of distributed representation for each group of neurons, and we analytically solve for the connection weights between neurons that will produce the desired computations between groups of neurons. While this approach does encompass neural learning techniques (e.g. Stewart, Bekolay, & Eliasmith, 2012), we do not use any learning in the models presented here.

More formally, the “patterns” for the various different stimuli (e.g. , OFFER, SMOKE), motor actions (e.g. , TAKE), and internal concepts (e.g. WORK, GOOD) are all defined as randomly chosen 64-dimensional unit vectors. This gives a unique randomly-generated vector for each concept. To use these patterns in a neural model, we must define how a group of neurons can store a vector using spiking activity, and how this spiking activity can be decoded back into a vector.

To define this neural encoding, the NEF generalizes standard results from sensory and motor cortices (e.g. Georgopoulos, Schwartz, and Kettner, 1986) that in order to represent a vector, each neuron in a population has a random “preferred direction vector” – a particular vector for which that neuron fires most strongly. The more different the current vector is from that preferred vector, the less quickly the neuron will fire. In particular, Eq. 1 gives the amount of current J that should enter a neuron, given a represented vector \mathbf{x} , a preferred direction vector \mathbf{e} , a neuron gain α , and a background current b . The parameters α and b are randomly chosen, and adjusting their statistical distribution produces neurons that give realistic background firing rates and maximum firing rates (Eliasmith & Anderson, 2003; Figure 4.3). These parameters also impact the model itself; for example, having an overall lower average firing rate means that the model will require more neurons to produce the same level of accuracy.

$$J = \alpha \mathbf{e} \cdot \mathbf{x} + b \quad (\text{Eq. 1})$$

This current can then be provided as input to any existing model of an individual neuron, to determine the exact spike pattern for a particular input vector \mathbf{x} . For this paper, we used the standard Leaky Integrate-and-Fire neuron model, which is a simple model that captures the behaviour of a wide variety of observed neurons (Koch, 1999, chp. 14). Input current causes the membrane voltage V to increase as per Eq. 2, with neuron membrane resistance R and time constant τ_{RC} . For the models presented here, τ_{RC} was fixed at 20 ms (Isokawa, 1997). When the voltage reaches a certain threshold, the neuron fires (emits a spike), and then resets its membrane voltage for a fixed refractory period. For simplicity, we normalize the voltage range such that the reset voltage to 0, the firing threshold is 1, and R is also 1.

$$\frac{dV}{dt} = \frac{JR - V}{\tau_{rc}} \quad (\text{Eq. 2})$$

Given Eqs. 1 and 2, we can covert any vector \mathbf{x} into a spiking pattern across a group of realistically heterogenous neurons. Furthermore, we can use Eqs. 3 and 4 to convert that spiking pattern back into an estimate of the original \mathbf{x} value. This lets us determine how accurately the neurons are representing given values. More neurons leads to higher accuracy. The idea behind Eq. 3 is that we can take the average activity a of each neuron i , and estimate \mathbf{x} by finding a fixed weighting factor \mathbf{d} for each neuron. Eq. 4 shows how to solve for the optimal \mathbf{d} as a least-squared error minimization problem, where the sum is over a random sampling of the possible \mathbf{x} values.

$$\hat{\mathbf{x}} = \sum a_i \mathbf{d}_i \quad (\text{Eq. 3})$$

$$\mathbf{d} = \Gamma^{-1} \Upsilon \quad \Gamma_{ij} = \sum_x a_i a_j \quad \Upsilon_j = \sum_x a_j \mathbf{x} \quad (\text{Eq. 4})$$

These two equations allow us to interpret the spiking data coming from our models. In Figs. 4 through 8, we take the spike pattern, decode it to an estimate of \mathbf{x} , and compare that to the ideal vectors for the various concepts in the model. If these vectors are close, then we add the text labels (e.g. WORK, OFFER, TAKE) to the graphs, indicating that the pattern is very similar to the expected pattern for those terms.

It should be noted that this produces a generic method for extracting \mathbf{x} from a spiking pattern without requiring a specific set of \mathbf{x} values to optimize over. That is, we can accurately use \mathbf{d} to determine if a particular pattern of activity means WORK even though we don't use the WORK vector to compute \mathbf{d} . The sums used to compute \mathbf{d} in Eq. 4 are over a random sampling of \mathbf{x} . Since \mathbf{x} covers a 64-dimensional vector space and since we use only 5000 samples in that space (increasing this number does not affect performance), it is highly unlikely that the sampling includes exactly the vector for WORK (or any other semantic pointer), but as shown in the Figs. 4 through 8, we can still use \mathbf{d} to identify the presence of those semantic pointers (or any others).

Importantly, we also use Eq. 4 to compute the connection weights between groups of neurons. In contrast to other neural modelling methods which rely on learning, the NEF optionally allows us to directly compute connection weights that will cause neural models to behave in certain ways. For example, given two groups of neurons, we can form connections between them that will pass whatever vector is represented by one group to the next group by using the connection weights given in Eq. 5 (see Eliasmith & Anderson, 2003 for the detailed proof).

$$\omega_{ij} = \alpha_j \mathbf{e}_j \cdot \mathbf{d}_i \quad (\text{Eq. 5})$$

However, simply passing information from one group to another is insufficient to implement the transition rules needed for our simulations. Fortunately, the NEF shows that you can find alternate \mathbf{d} values to estimate complex nonlinear functions. That is, instead of simple passing a value from one group to another, we can define an arbitrary function $\mathbf{f}(\mathbf{x})$ and compute \mathbf{d}^f as per Eq. 6. Now, if synaptic connections are formed via Eq. 5, if the first neural population fires with the pattern for \mathbf{x} , then the connections will cause the second population to fire with a pattern representing the result of $\mathbf{f}(\mathbf{x})$.

$$\mathbf{d}^f = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \sum_x a_i a_j \quad \Upsilon_j = \sum_x a_j \mathbf{f}(\mathbf{x}) \quad (\text{Eq. 6})$$

This approach allows us to define the various transition rules given in the paper, and the compression/decompression operation (Fig. 2). The transition rules are converted into a function that maps the particular input vectors to particular output vectors. This function is used to compute \mathbf{d}^f (Eq. 6), which is then used to compute the synaptic connection weights (Eq. 5). The model is then run. To provide input to the model, we generate input current into the sensory neurons for the particular sensory stimuli (Eq. 1). To analyze and interpret the spiking patterns, we convert the spikes back into a vector (Eq. 3) and compare it to the ideal vectors for each concept.

The compression function used here is *circular convolution*. This takes two vectors (\mathbf{x} and \mathbf{y}) and produces a third vector \mathbf{z} as per Eq. 7. This vector \mathbf{z} can be thought of as a compressed representation of \mathbf{x} and \mathbf{y} , forming the basis of our semantic pointers. Importantly, given \mathbf{z} and \mathbf{y} (or \mathbf{x}) we can recover an approximation of \mathbf{x} (or \mathbf{y}) by computing the circular correlation (Eq. 8). This is how semantic pointers can be decompressed into their constituents.

$$\mathbf{z}_i = \sum_j \mathbf{x}_j \mathbf{y}_{i-j} \quad (\text{Eq. 7})$$

$$\hat{\mathbf{x}}_i = \sum_j \mathbf{z}_j \mathbf{y}_{i+j} \quad (\text{Eq. 8})$$

In general, it is possible to use the Neural Engineering Framework to build a network where there are two input populations (one for \mathbf{x} and one for \mathbf{y}) and one output population (\mathbf{z}) such that you can input any two arbitrary vectors and get out their convolution. Importantly, this will work for any input vectors, not just the randomly chosen ones used in the optimization (Eq. 6).

However, for the simulations described here, we use a simpler method where a particular neural connection always convolves its input vector \mathbf{x} with a fixed vector. For example, the connection from the sensory area to the memory area in Fig. 9 computes the function $\mathbf{f}(\mathbf{x})=\mathbf{x}*\text{SENSORY}$ where $*$ is the circular convolution and SENSORY is a randomly chosen semantic pointer vector. The synaptic connection weights computed using this function and Eqs. 5 and 6 result in a spiking neural network that accurately combines information into a single memory semantic pointer regardless of what particular vector \mathbf{x} is provided to the sensory system. A similar function is defined for the other connections into the memory system, resulting in a final semantic pointer of $\mathbf{x}*\text{SENSORY}+\mathbf{y}*\text{ACC}+\mathbf{z}*\text{SMA}+\mathbf{w}*\text{PFC}$. To decompress this semantic pointer, we use a circular correlation instead (Eq. 8).