Integrating social cognition into domain-general control

Ward, Rob; Ramsey, Richard

Cognitive Science

DOI:
10.1111/cogs.13415

Published: 01/02/2024

Publisher's PDF, also known as Version of record

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13415

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13415

Integrating Social Cognition Into Domain-General Control: Interactive Activation and Competition for the Control of Action (ICON)

Robert Ward,[a] Richard Ramsey[b]

[a]*Cognitive Neuroscience Institute, Department of Psychology, Bangor University*
[b]*Department of Health Sciences and Technology and Department of Humanities, Social and Political Sciences, ETH Zürich*

Received 1 November 2023; received in revised form 2 February 2024; accepted 6 February 2024

Abstract

Social cognition differs from general cognition in its focus on understanding, perceiving, and interpreting social information. However, we argue that the significance of domain-general processes for controlling cognition has been historically undervalued in social cognition and social neuroscience research. We suggest much of social cognition can be characterized as specialized feature representations supported by domain-general cognitive control systems. To test this proposal, we develop a comprehensive working model, based on an interactive activation and competition architecture and applied to the control of action. As such, we label the model "ICON" (interactive activation and competition model for the control of action). We used the ICON model to simulate human performance across various laboratory tasks. Our simulations emphasize that many laboratory-based social tasks do not require socially specific control systems, such as those that are argued to rely on neural networks associated with theory-of-mind. Moreover, our model clarifies that perceived disruptions in social cognition, even in what appears to be disruption to the control of social cognition, can stem from deficits in social representation instead. We advocate for a "default stance" in social cognition, where control is usually general, but representation is

Correspondence should be sent to Robert Ward, Cognitive Neuroscience Institute, Department of Psychology, Bangor University, Bangor, Gwynedd, Wales, LL57 2AS, UK. E-mail: r.ward@bangor.ac.uk; Richard Ramsey, Department of Health Sciences and Technology and Department of Humanities, Social and Political Sciences, ETH Zürich, Gloriastrasse 37/39, Zürich, 8092, Switzerland. E-mail: richard.ramsey@hest.ethz.ch

specific. This study underscores the importance of integrating social cognition within the broader realm of domain-general control processing, offering a unified perspective on task processing.

Keywords: Social cognition; Control of action; Stimulus–response compatibility; Computational modeling; Autism spectrum; Imitation; Approach–avoid

1. Introduction

What is "social cognition," as distinguished from cognition more generally? The theoretical position we develop here is that many aspects of social cognition can be accounted for by a combination of specialized feature representations interfacing with domain-general cognitive control (Ramsey & Ward, 2020). We develop and instantiate this proposal as a working model, capable of simulating behaviors from a range of laboratory tasks. We believe the importance of domain-general control has been to a degree neglected in social neuroscience and social cognition research and seek to redress this balance. We begin by describing key aspects of our proposal.

1.1. Domain-specific representation and domain-general control

Research in cognitive science distinguishes between domain-specific and domain-general information processing. Domain-specific systems are specialized for particular stimulus categories or task features, while domain-general ones operate across stimulus or task features (Barrett, 2012).

For our focus on action selection to visual objects, domain-specific visual representations are especially relevant. The human brain responds to visual stimuli with the activation of domain-general and domain-specific regions. For example, the primary visual cortex is arguably domain-general: Cells in the primary visual cortex respond to the orientations of contours, on all sorts of objects (Hubel & Wisesel, 1977). However, as information flows forward from the primary visual cortex, specialized brain regions are recruited for processing specific object categories, like faces, bodies, places (Kanwisher, 2010), and visual word form (Dehaene & Cohen, 2011). The response of these specialized areas is sharply tuned for specific categories, forming what we call domain-specific representations.

Representations of different kinds of objects and events are likely to be domain-specific in important ways because useful representations must capture the unique elements of their subjects. For example, consider two very different classes of objects, human faces and English words. Useful representations of human faces need to allow inferences about things like identity and mood. Useful representations of words need to allow access to semantics on the basis of arbitrary letter combinations. These are very different, domain-specific, properties of the representations, which largely rely on distinct patches of neural tissue in the ventral visual stream (Dahaene & Cohen, 2011; Kanwisher, McDermott, & Chun, 1997). We expect representations of object categories to frequently have domain-specific elements, even while retaining domain-general aspects, such as spatial location.

However, representations are only half the story, as it were. Representations can be distinguished from the computational processes that manipulate them. The general intuition is that a representation of an object "stands in" for that object and can be subject to all kinds of cognitive manipulations (Haugeland, 1992): For example, the representation of an apple might be manipulated to create new semantic associations to an apple; to direct attention toward the location of an apple in the real world; to focus on an attribute of the apple, such as its color; to select an apple as a target for reaching from a crowd of fruit; and so on. The distinction between representations and computational processes can be highly formalized, as in the physical symbol system hypothesis of Newell (1980), and it can be more nuanced, for example, in the context of dynamical systems (van Gelder, 1998; Ward & Ward, 2009).

When we discuss the distinction between representation and processes here, we will be referring specifically to *control* processes: the range of processes regulating the operation of the cognitive system, to allow accurate and flexible responses to task demands (Botvinick et al., 2015; Braver, 2012). This means not only ecologically relevant needs and stimulus–response associations (e.g., "if you are hungry, find something nutritious to eat") but also entirely arbitrary tasks (e.g., "press the green button if you see a letter A"). Attentional selection, response inhibition, executive control of task organization, and task switching would all be examples of domain-general control processes, applicable to many kinds of tasks. As such, domain-generality is the default assumption for research in attention, executive control, and the organization of behavior. Research in these areas tends to use a variety of stimulus categories (pictures, text, letters, auditory stimuli, etc.). For example, investigation about the "early" or "late" nature of selective attention began with auditory stimuli (Cherry, 1953), then proceeded smoothly to the use of tachistoscopic displays and visual stimuli of all sorts (Treisman, 1980), and to neuroimaging studies of executive function in which much of frontoparietal cortex responds to an arbitrary target stimulus (Duncan, 2010).

Sometimes the evidence shows that attentional systems are *not* domain-general. For example, attentional selection in one modality (e.g., selecting a visual target from visual distractors) produces a relatively long-lasting timecourse of interference on subsequent selection in that modality (the "attentional blink"). However, this interference appears to have a modality-specific component. For example, visual selection produces long-lasting interference on subsequent visual targets but has much less effect on auditory ones (e.g., Duncan, Martens, & Ward, 1997), suggesting domain-specific attentional capacity. Domain-specificity when found is subject to detailed investigation (e.g., Arnell & Jenkins, 2004; Wang, Qian, Zhao, Tang, & Zhang, 2022); but domain-specific control processing would be the exception rather than the presumption.

1.2. Domain-specific thinking

Our view is that social cognition, as a field, has over-emphasized domain-specific thinking at the neglect of domain-general thinking. To an understandable extent, much of this emphasis is due to excitement around domain-specific representations of social stimuli, such as brain

regions specialized for bodies and faces (Kanwisher, 2010). We have no concerns about these kinds of socially specific representations and share the excitement around them.

But again, to emphasize, we are distinguishing representations from processes. Unless the field operates carefully and precisely, excitement and discussion about domain-specific representations can incorrectly, unintentionally, and unknowingly drift into presumptions about socially specific control processes. This kind of enthusiasm takes two forms, which vary in what we might think of as researcher awareness. First, to a limited but significant degree, there are *explicit* proposals for domain-specific control processes, which are tied to social processes and do not generalize to other contexts. For example, a dominant view is that the control of conflict in automatic imitation is mediated by a specialized theory-of-mind network (Brass, Zysset, & von Cramon, 2001, 2009; Sowden & Shah, 2014; Spengler, von Cramon, & Brass, 2009; Wang & Hamilton, 2012). Second, and more common, are what might be called *implicit* suggestions for domain-specific social control processes. For example, research into the perception of eye-gaze very often refers to "social" attention (Klein, Shepherd, & Platt, 2009; Nummenmaa & Calder, 2009). This raises the question of whether "social" attention is something distinct from "regular" attention, and if so, in what ways (Braithwaite, Gui, & Jones, 2020; Elsabbagh & Johnson, 2016; Emery, 2000; Frischen, Bayliss, & Tipper, 2007; Langton, Watt, & Bruce, 2000; Mundy & Bullen, 2022; Mundy & Newell, 2007; Nummenmaa & Calder, 2009).

At the very least, we might expect claims about social control processes to begin by examining how they are different from domain-general control processes, but this is rarely the case. Our mission here is to demonstrate that for many laboratory tasks, we have no need to create or refer to socially specific control processing. We want to move the burden of proof to those who argue, explicitly or implicitly, that there are social control processes separate from domain-general ones. Our concern is that while a domain-specific emphasis has allowed considerable progress, social cognition risks becoming isolated from developments in more general and well-established forms of cognitive science. As such, for more effective and efficient progress to be made, we suggest that researchers interested in social cognition should collaborate more directly with researchers who study more general mechanisms of attention and vision.

1.3. Our alternative

We advocate an alternative theoretical and computational position. Consistent with other recent proposals (Barrett, 2012; Binney & Ramsey, 2020; Michael & D'Ausilio, 2015; Ramsey, 2018; Spunt & Adolphs, 2017; van Elk, van Schie, & Bekkering, 2014), control in social tasks comes through domain-general control processes, such as prioritization, selection, and memory (Ramsey, 2018; Ramsey & Ward, 2020). To illustrate our ideas and give guidance about how they might work in practice, we have developed a computational network model of attentional selection and control for many kinds of social and non-social tasks. This is part of an effort to respond to calls for more systematic theory construction in psychology (Gray, 2017; Haig, 2014; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; for a special issue on theory, see Proulx & Morey, 2021).

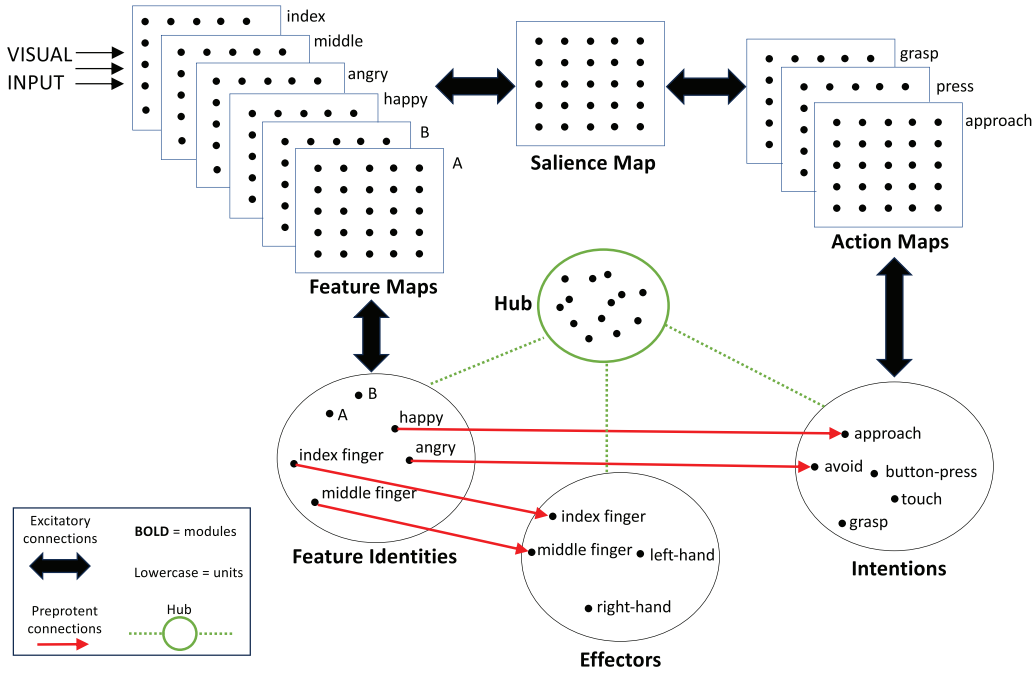

Fig. 1. An overview of the general structure of the ICON model. Modules are labeled in bold font, whereas identification of units and groups are smaller lowercase labels. Within a module, there is inhibition between units. Between modules, there is excitation between corresponding units (bold arrows). For example, all the units within the "A" feature map are reciprocally connected with the "A" identity. Prepotent connections reflect learned associations, and those are indicated by the thin lines. The features were chosen (letters, faces, and fingers) to allow the simulation of a variety of social and non-social laboratory paradigms. Some plausible connections and structures were omitted for simplicity as not required for the simulations: The effector pool might reasonably have direct connections with Intentions and Action maps; and the "button-press" and "avoid" actions are, for the simulation, considered non-spatial and unlike the other intentions do not have a corresponding action map. Not shown are the connections to and from the hub, which in principle could link any unit within any pool.

2. The model

2.1. Overview

As we will show, our interactive activation and competition model for the control of action (ICON) instantiates the framework proposed by Ramsey and Ward (2020) and its three key principles: (a) a variety of domain-specific and domain-general feature representations; (b) localizable domain-general systems for controlled processing; and (c) network-wide biased competition.

The ICON model is a functioning computational system with a broad scope from perception to action. The model reflects an agent designed to live in a crude and desolate world of psychological laboratory experiments (see Fig. 1). The agent's entire environment consists of a 5 × 5 grid of locations. Only a few objects exist: two letters (A and B), and a few body

parts (index and middle fingers; happy and angry faces). The agent is modeled with a "hand" consisting of an index and middle finger effectors and has representations for action plans allowing button pressing with either finger, as well as reaching to grasp objects at any specific location within the 5×5 grid.

Within the bounds of this simple environment, the network is a complete system for perception and action. This simplicity is beneficial in the context of simulations and makes the dynamics of the model and action selection process much easier to understand. The network can simulate a range of traditional social and non-social laboratory tasks, including object-based filtering, location-based filtering, prepotent stimulus–response associations, and goal-directed stimulus–response mappings.

The entire model is built in base R using the freely available iac package (devtools::install_github("rob-ward-psych/iac")), which is designed to "easily" create arbitrary interactive activation and competition (IAC) networks. The R code for running and analyzing all simulations is available from (https://github.com/rich-ramsey/icon-model). The raw data generated by the simulations are reproducible from these files but can also be downloaded directly here from the OSF page (https://osf.io/m2enf/). In the next section, we provide details on the model's architecture.

2.2. The model architecture

2.2.1. A constraint network

ICON is based on the IAC architecture as developed by McClelland and Rumelhart (1981). IAC models are distributed networks of simple units, organized into modular pools. This quote from Rumelhart, Smolensky, McClelland, and Hinton (1986) is slightly extended but provides tremendous clarity about how to best understand the operation of IAC models as a *constraint network*:

> It is often useful to conceptualize a parallel distributed processing network as a constraint network in which each unit represents a hypothesis of some sort (e.g., that a certain semantic feature, visual feature, or acoustic feature is present in the input) and in which each connection represents constraints among the hypotheses. … If the constraints [between hypotheses] are weak, the weights should be small. If the constraints are strong, then the weights should be large. Similarly, the inputs to such a network can also be thought of as constraints. A positive input to a particular unit means that there is evidence from the outside that the relevant feature is present. … The stronger the input, the greater the evidence. If such a network is allowed to run it will eventually settle into a locally optimal state in which as many as possible of the constraints are satisfied, with priority given to the strongest constraints. The process whereby such a system settles into such a state is called relaxation. (pp 8–9)

2.2.2. Bayesian-like

This conception of a constraint network also shows how even the most basic IAC networks embody "Bayesian-like" distinctions between *prior knowledge* about how the world

works, as reflected in the weights between units (i.e., the constraints between hypotheses); *current evidence*, as reflected in the external inputs to the network; and the *plausible interpretation* of evidence, arising as the network relaxes to a stable state of unit activations. Although true Bayesian inference is not needed for our purposes here, McClelland, Mirman, Bolger, and Khaitan (2014) have formally shown how a revised formulation of IAC networks can produce true Bayesian probabilistic inference (see also McClelland, 2013).

2.2.3. Biased competition

An important feature of IAC networks for our purposes is their relationship to biased competition. Models of biased competition have taken a central place in our understanding of controlled cognitive processing. Biased competition assumes that the features of objects are represented across multiple specialized modules, with systematic communication patterns within and between them (Desimone & Duncan, 1995; Duncan et al., 1997; Beck & Kastner, 2009). Within a module, there is competition so that the features of a single object come to dominate the module's activity. Competition within a module may be biased by external inputs, reflecting anything from bottom-up salience to top-down goals. Between modules, there is excitation between corresponding features of the same object or event. The result is that a bias for a winning object in one module propagates to other modules so that the representation of a single object or event comes to dominate across the entire network.

IAC networks are naturally structured to reflect these key principles of biased competition. In fact, their equivalence lets us realize that one of the first computational models of biased competition was the McClelland and Rumelhart (1981) model of word and letter perception. In the McClelland and Rumelhart model, units within the word module competed with one another as did units within the position-specific letter modules. Words had excitatory connections to their corresponding units in the letter modules. This soft architectural constraint meant that a winning feature, arising through competition within one module, activated corresponding features in other models so that the entire network relaxed to a single consistent state.

To help illustrate the process of biased competition, we provide a concrete example using the McClelland and Rumelhart (1981) word and letter perception model. Suppose the visual input corresponded to the letters W*RK, where the star indicates the absence of a letter in the second position. The letters that are present excite corresponding units at the word level, such as WORN to a limited extent and WORK to a greater extent. Activation at the word level flows back to the letters, so that the unit representing the letter O in the second location becomes more active, and inhibits other letter representations in the second position. The network relaxes into a state where the word WORK and its position-specific letter units dominate. This state represents the most plausible interpretation of the evidence provided by external inputs, given what the network knows in the form of its connection weights. As we will see, our model of selection and control of action also embodies biased competition within an IAC network.

2.3. Detailed structure

The ICON network consists of pools of units or *modules*. The operation of the network is defined by the structure of modules and by the connections between units. These properties of the model are fixed, leaving only the external inputs to vary between simulations of different tasks.

2.3.1. Modules

2.3.1.1. Feature maps (What × Where): This module can be thought of as an array of spatial feature maps, activated by external inputs generated by objects in the visual world. The receptive fields of units within the different maps are "tuned" to respond to a specific feature at a specific location: a conjunction of what the feature is and where it occurs. In a more realistic model, one could imagine something closer to stacks of convoluted neural nets, which progressively abstract away spatial information as the complexity of identity information increases. In that case, we assume units would still code a conjunction of what and where information, but as we ascend the stack, units would represent conjunctions of progressively coarser location codes with progressively more elaborate identity codes.

2.3.1.2. Identities (What): The units in this module represent the presence of a specific visual object, somewhere in the environment, but they do not encode where those objects occur. This location-invariant code is useful in tasks such as object recognition. These units are meant to be analogous to specialized visual areas of the ventral visual stream, which respond in a category-specific way to visual stimuli, including social (faces and body parts), and non-social (letter) domains.

2.3.1.3. Salience map (Where): The activity of units in this module indicates the presence of a visual object at a specific location but gives no information about the features of that object. Competition within the map ensures that eventually a single location will tend to dominate, representing the location of most interest in the current context. Activity in this module therefore reflects the salience of objects at specific locations.

2.3.1.4. Intentions (How): This module represents different intentions for action but not a plan for execution. For example, the intention to grasp something is represented by a unit in this module. But the intention to grasp by itself does not indicate what or where to grasp. The full plan for action must be constrained by the identity and location of objects in the environment. Competition within this module tends to keep the network focused on a single intention. Intentions are normally activated through the "hub" module, which we describe below.

2.3.1.5. Action maps and effectors: In general, upper limb responses, like reaching to grasp, involve both shaping distal effectors (e.g., hand and fingers) to the characteristics of the object being manipulated and transporting the effector to a location that allows that interaction. Units in these two modules are meant to provide these different aspects of a motor

plan. Activity within the *action maps* indicates the target location for action in the visual world. The model allows that the target location may depend upon the intended action: For example, approach and avoid intentions have distinct action maps. Activity within the *effector* module specifies which effectors will be used for action, for example, whether the model should respond using its "index finger" or "middle finger."

Simplifying assumptions are made for actions, as they are for other parts of the network. For example, in the current simulations, the model's effectors do not produce visual input, even when moved into the 5 × 5 visual world. There is no forward or backward prediction or online guidance. In effect, the model's mission is complete when it settles to a state in which an action plan is created, specifying the desired transport and effectors.

2.3.1.6. Hub: Influenced by recent neuroscience research (e.g., Cole et al., 2013) as well as computational principles (e.g., Botvinick & Cohen, 2014), the hub provides flexible temporary association between stimuli and actions. The hub is inspired by neurophysiological work examining brain systems for cognitive control (hubs), and the function of the frontoparietal cortex, needed for flexible and organized behaviors (Cole et al., 2013; Duncan, 2010). In the absence of neurophysiological hubs, behavior becomes stereotyped, inflexible, and disorganized. For example, in the syndrome of "utilization behavior," behavior is triggered by associations to the immediate environment rather than current task demands (Riddoch, Edwards, Humphreys, West, & Heafield, 1998).

Units within the hub module encode arbitrary associations between units in any of the other modules. Consider, for example, a task to find the letter A within the visual environment, and then touch it with the index finger. The hub unit representing this task sends excitatory connections to the identity "A," the effector "index," and the intention "touch." This combination of associations encodes a specific task: to touch the letter A. Within the hub are multiple such units, each connected to a particular combination of units in other pools representing the stimulus and response characteristics of a task. These hub weights are bidirectional but asymmetric: the top-down weights (from the hub to other units) are larger than the bottom-up weights (from the units to the hub). The bottom-up weights allow hub units with activated task components to dominate over those without. Although much of the model is based on previous work by Ward (1999), the hub module is a crucial addition, as it gives the model a coherent mechanism for specifying and performing arbitrary tasks.

2.3.2. Connections within and between modules

2.3.2.1. Inhibition and excitation: Units within a module compete with one another. This is instantiated by negative connection weights between all units in the same module. Therefore, within a module, if one unit is activated, the others are suppressed. Between different modules are relatively sparse excitatory connections, for units representing what we refer to as consistent hypotheses. For example, a unit within the visual feature array will carry bidirectional, positively weighted connections with the corresponding identity unit and the corresponding location of the salience map. Take the unit representing the center location of the visual feature array for the letter A: It is connected to both the identity unit for "A" and the central location of the salience map. In terms of a constraint network, the hypothesis that there

is an object at the center location is consistent with the hypothesis that there is a letter "A" at that location. Our network has more restricted connectivity than the general form expressed by Rumelhart et al. (1986), as we do not use inhibitory connections between modules. However, as we will see, this restricted connectivity is sufficient for the network to relax into a state that is fully consistent with task requirements and stimulus evidence.

2.3.2.2. Prepotent associations between stimuli and actions: The model includes some prepotent stimulus–response associations, in the form of positively weighted, unidirectional connections, from specific identity units to action-related units. For example, the identity unit for happy faces activates the intention for approach, whereas identity for angry faces activates the intention for avoid. Other prepotent associations are from the identity of effectors to the effector units themselves (e.g., from an index finger as an identified form to the index finger as an effector), as well as a bias for animate over inanimate objects. These associations are meant to reflect the result of learning about repeated co-occurrences over an extended time. As such, these prepotent associations represent in-built biases within the model.

2.3.2.3. Noise: All connections within the model are subject to some noise. Specifically, the final activation of a unit is a function of the activations of the units it is connected to, the strength of those connections, and random variation. Noise is helpful in these simulations because it creates a range of behaviors for a given set of task conditions. For example, we can vary the overall error rates within a simulation by controlling the level of noise.

2.4. Stable states of the model and what they mean

Having described the structure of the network, we can now address a fundamental question in its operation. As this is a constraint satisfaction network—which will eventually relax into a stable state—what does the stable state represent? We indicated earlier it can be helpful to think about network function in Bayesian-like terms.

2.4.1. Prior knowledge

As mentioned, the structure and connections of this network represent its prior knowledge about how the world works. For the model, this prior knowledge is, roughly speaking, about how to interact with visual objects. The network has the knowledge to interact with every object it knows about using every action it knows about. But it also has some biases. It is biased toward visual activity in the center of its world. This is implemented as a small tonic input to the center of the salience map. It is likewise biased toward animate, or socially relevant, over inanimate ones: as the model only knows about letters, faces, and body parts (fingers), the animate objects also have social significance. The model also carries biases in the form of some prepotent responses to specific visual properties (e.g., approach rather than avoid happy faces). These different biases reflect the model's knowledge that not all things in its world are equally likely nor should they be treated in the same way.

2.4.2. Current evidence

The external inputs to the model specify the visual world, in terms of feature map activity, and task goals, in terms of activation of units in the hub, specifying any current task set. These inputs constitute the *evidence* about what objects are in the world and how the agent should deal with them. In our simulations, we typically provide the model with external inputs representing multiple objects (e.g., an "A" in the center and a flanking face), and hub activation representing multiple tasks (e.g., approach if there's an "A"; avoid if there's a "B").

2.4.3. Best interpretation: What to do

Given the network knows about a range of different actions to different objects, and given the evidence about the objects and tasks relevant to its immediate situation, the final stable state of the network represents a specific action to a specific visual object. The action and object jointly selected in this way represent the network's best conjoint understanding of the current environment and how it should behave. Now, because our particular network does not perform true Bayesian analysis (as in, e.g., McClelland, 2013), this understanding will not always be optimal, but we are not invested at this point in optimality. What we want to highlight is that while the network does have *domain-specific* knowledge, including knowledge about socially relevant stimuli, the process of inference is *domain-general*, arising through network relaxation. Inference about what to do and how to do it is the result of activity circulating throughout the entire network as we detail later.

2.5. Predictions and overview of simulations

The ICON model makes several general predictions, and we present a range of simulations to demonstrate them. The first and most general prediction emerges from the network architecture itself, namely, an "object-based" selection system, in which a single object or episode of attention is selected with its associated features. All properties of the selected object, including its implications for action, become available concurrently, whether the selected object is a social or non-social stimulus. Second, there is no need for a socially specific form of control. Social stimuli—based on domain-specific representations—gain control of behavior through the same general mechanisms as other stimuli. Third, some apparent deficits in social control *processes* can be modeled as deficits in socially specific *representation*.

In addition, we make more detailed predictions tied to the specifics of the different tasks we simulate:

2.5.1. Simulation 1: Automatic imitation

Our first simulations model a social laboratory task argued to index automatic imitation (Brass, Bekkering, Wohlschlager, & Prinz, 2000). We demonstrate how congruency effects in this task can be replicated, contrary to current accounts (Brass et al., 2001, 2009; Sowden & Shah, 2014; Spengler et al., 2009; Wang & Hamilton, 2012), without requiring a socially specific form of control and without processes related to either a self-other distinction or a theory of mind. We also review our criterion-based procedure for generating

reaction-time distributions and condition-dependent accuracy, the object-based nature of selection, and the effect of target and distractor salience on the observed congruency effects.

2.5.2. Simulation 2: Approach–avoidance to social stimuli

Our second set of simulations uses the same network as Simulation 1 to model an entirely different task. In this case, we model the congruency effects resulting when the task requires approaching a pleasant, as compared to approaching an unpleasant, social stimulus. We again demonstrate the object-based nature of selection, and we show multiple ways in which a change in behavior that might appear to be related to socially specific control can arise in the absence of socially specific control processes.

2.5.3. Simulation 3: The autistic spectrum and the "first-year puzzle"

The distinction between socially specific and domain-general cognition has important implications for understanding atypical development relating to autistic spectrum disorders (ASDs; e.g., Mundy & Bullen, 2022). Our third set of simulations considers two important perspectives: "social-specific theories," emphasizing early social deficits leading to disrupted social development, and "domain-general theories," suggesting broader cognitive deficits cause ASD's social differences. The research presents a model addressing both theories and the "first-year puzzle" in ASD. Simulations indicate that early social representation disruptions might become apparent only when coupled with a later domain-general control disturbance.

2.5.4. Simulation 4: Visual search and reach-to-grasp

In our fourth set of simulations, we show some of the versatility of the model, again using the same network, but this time modeling a task in which the participant is meant to reach out and grasp a target letter from a field of non-target letters. This demonstrates how the model's architecture is versatile enough to map between a range of different stimulus sets, task goals, and actions.

3. Simulation results

3.1. Simulation 1: Automatic imitation

3.1.1. The task

We first present simulations of the automatic imitation paradigm as used by Brass and others (Brass et al., 2000). In the simulated task, an imperative target letter is presented centrally, with an adjacent, task-irrelevant, distractor (a middle or index finger stimulus). The task for the model is to press the button "under" its index finger if the target is an "A" or middle finger if the target is a "B."

We simulate Congruent conditions in which an "A" target appears with an index finger distractor and Incongruent with an "A" target and middle finger distractor. Crucially, as described earlier, the network includes prepotent associations from visual stimuli to responses.

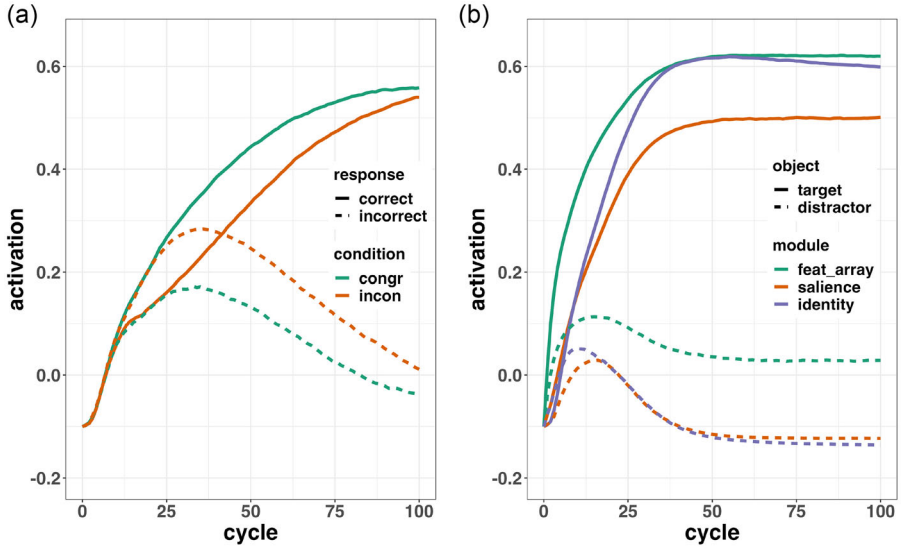

Fig. 2. Timecourse of activity across conditions and parameters. (a) Activation plotted as a function of response and condition (congr = congruent; incon = incongruent). (b) Solid lines show activation corresponding to the target object (the letter target) in the Feature Maps, Feature Indentities, and Salience Map modules. Dashed lines show activation corresponding to the distractor object (finger) in the same modules. The Feature Map activation is not fully suppressed, as it is receiving continuing external input throughout the trial, representing the visual presentation of the finger.

Important in this case are the associations from the finger identities (in the What module) to their corresponding units in the Effectors module. Essentially, the network associates seeing an index finger with movements of the index finger. These associations, in conjunction with the visual characteristics of the display, will drive the congruency effect.

We think of the simulations that follow as analogous to the performance of a single participant. If we were to model multiple participants, we would use the same network architecture but vary the connection strengths and noise to some degree. Although we report tests of significance, we do not place much weight on them. These are interesting in some ways, but trivial in others, as we could change significance and effect size measures by running ever greater numbers of simulated trials and even simulated participants. Instead, we focus on the broad patterns and robust outcomes.

3.1.2. Bias propagation: How the model responds

We first examine the activations for units representing the index (correct) and middle finger (incorrect) response effectors. As a reminder, these units inhibit one another, as they are within the same module (Effectors). Mutual inhibition ensures that as one unit gains an advantage in activation, it will progressively inhibit the other. In Fig. 2a, we see early activation of both effector units in both conditions. With time, activation trajectories diverge, and the correct response is fully activated and incorrect fully inhibited. Interestingly, in the Incongruent

condition, the incorrect response has a small early advantage over the correct one. That is, early in the trial, we see subthreshold activation of the incorrect response associated with the visual distractor. This initial bias toward the incorrect response is corrected on most trials. The trajectories of effector activations shortly cross, and the correct response goes on to fully suppress the incorrect one.

What gives an advantage to the correct response? This is the activity of the hub. In this case, there are two hub units receiving external activation throughout the trial, reflecting the current task set. We will call these press-index-for-A and press-middle-for-B. For example, press-index-for-A has excitatory connections with identity "A," effector "index," and intention "button press." Press-index-for-A and the "A" target present in this trial provide mutual support for one another. In contrast, as the target "B" is not present in the display, press-middle-for-B does not receive the same level of support. The bias in the hub, for one unit over others, propagates through to other modules representing its corresponding attributes: in this case, identity "A," action "button press," and effector "index." By doing so, the hub has allowed the network to map the imperative stimulus to the correct response.

3.1.3. The object-oriented selection of the target and its attributes

We now look more broadly at activations throughout the network. Fig. 2b shows the time-course of attributes relating to the target and distractor stimuli, averaged across all simulated trials. The figure shows the propagation of bias from one module to another. As the competition within one module resolves upon an alternative, the excitatory connections from that module propagate the bias, resulting in a cascade of resolved competitions. The end result, as seen in the figure, is that the attributes of the target come to dominate across the various modules of the network. We can also see the attributes of the distractor are progressively inhibited.

3.1.4. Simulated reaction times

The activations in Fig. 2 represent average activations for 100 trials (50 Congruent; 50 Incongruent). For easier comparison with human performance, these activations can be converted into response latencies. This raises the question, at what point in these graphs should the model make a response, particularly given that there is noise on every trial and that activation levels will vary somewhat unpredictably?

We assumed a process by which a response is selected for execution when the unit representing that response exceeds its alternatives by a criterion difference. Raising the criterion increases the certainty that the correct response is selected but also increases the time to respond. Conversely, lowering the criterion decreases response time but increases the chance of an error. By varying the criterion, we generate a response operating characteristic (ROC) describing how speed and accuracy co-vary in the task.

The resulting ROC curve is shown on the left side of Fig. 3. As we change the criterion, we change both the resulting mean accuracy and mean RT. We choose as our criterion the value producing a required accuracy level of 98%, with the fastest overall RT. The right side of Fig. 3 shows how varying the required accuracy level shifts the congruency effect. These ROC curves bring a first prediction from domain-general thinking to this imitation

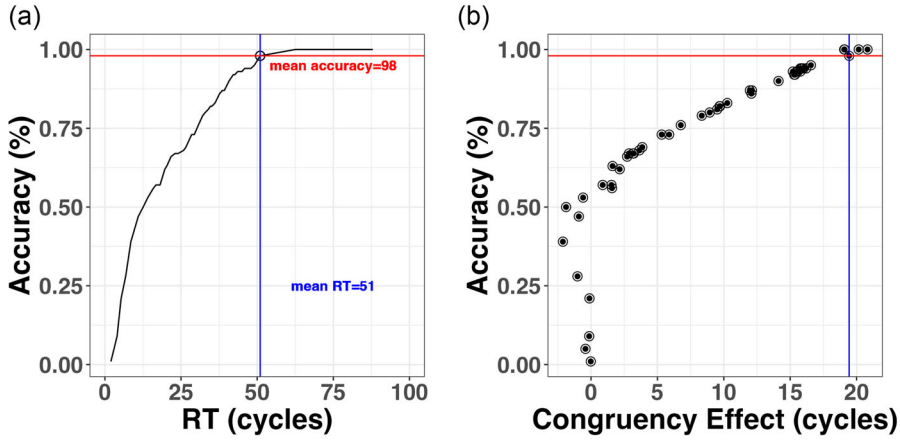

Fig. 3. Response operating characteristic (ROC) curves. (a) Speed versus accuracy and (b) congruency effect versus accuracy.

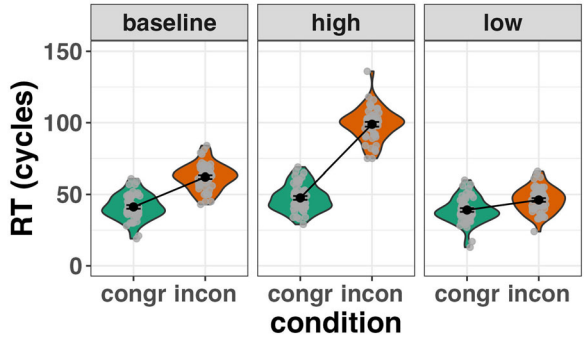

Fig. 4. RT in the imitation task plotted as a function of congruency and stimulus salience. congr = congruent; incon = incongruent; baseline = standard model settings; high = high distractor salience; low = low distractor salience. Error bars = standard error of the mean.

task: We can see how simple response criteria could move participants along a speed-accuracy trade-off, which can then modify the congruency effect they show.

The resulting RTs show a clear advantage for Congruent over Incongruent trials, 41.3 cycles for Congruent (100% accuracy over 50 trials); 60.7 cycles for Incongruent (96% accuracy); $t(92.9) = 9.46, p < .0001, d = 1.6$. These data are illustrated in Fig. 4 in the baseline condition panels.

3.1.5. The effects of relative stimulus salience on the congruency effect

While the inferential statistics based on the congruency effect in the baseline condition (shown in the leftward panel of Fig. 4) would be highly prized by some researchers in a test of human participants, they are less relevant in a simulation, where we have the flexibility to run unlimited numbers of trials and to manipulate internal parameters of the model. Instead,

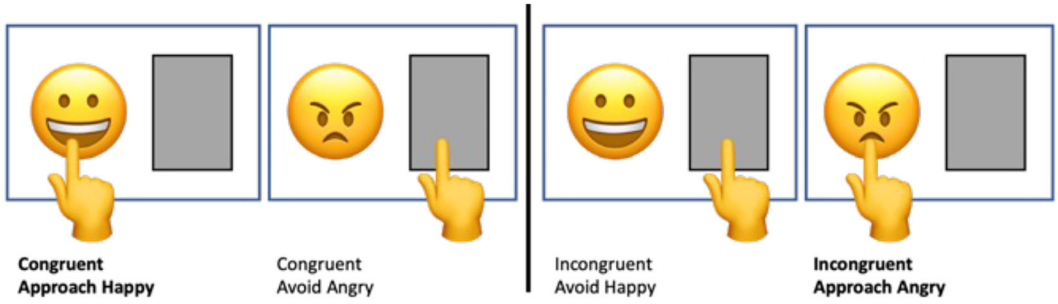

Fig. 5. Sample conditions from Bamford and Ward (2008). Two stimuli were presented on either side of the fixation: a face (angry or happy) and a rectangle. In Congruent blocks, participants approached the happy face (and avoiding angry faces by instead approaching the rectangle); in Incongruent blocks, they approached the angry face. Facial photographs were used in the actual experiment.

we want to emphasize the pattern of results and how these can change based on experimental manipulations.

Here, we show how the size or even existence of congruency effects in our simulations can be manipulated, by increasing the relative salience of the target or the distractor. When distractor salience is increased, so are the effects of prepotent associations; therefore, the congruency effect will increase. When distractor salience is decreased, we increase the suppression on the distractor and therefore reduce its influence. These effects are shown in Fig. 4, where we manipulated distractor salience by increasing or decreasing the bias toward animate stimuli.

These predicted effects of salience in the automatic imitation task will come as no surprise to those who have devoted years or decades of effort to understanding stimulus-response compatibility (SRC) effects. However, this is part of our general point. Domain-general studies have learned a great deal about attention and action. The model informs us about how these concepts from domain-general studies can be brought into investigations of social cognition. The model shows how the imitation experiments simulated here, which are intended to be demonstrations of socially specific control, can be understood as domain-general control processes operating on socially specific representations.

3.2. Simulation 2: Approach–avoidance to social stimuli

Our next simulations are on approach and avoidance to happy and angry faces. Congruency effects in approach–avoid tasks demonstrate prepotent associations between stimuli and responses, that is, spontaneous activation of responses following stimulus detection or identification. Fig. 5 illustrates tasks used by Ward and colleagues (Bamford & Ward, 2008; Bamford, Turnbull, Coetzer, & Ward, 2009; Kramer et al., 2020), similar in principle to other approach–avoidance experiments (e.g., Markman & Brendl, 2005; Solarz, 1960). The paradigm measures a congruency effect reflecting an automatic bias to approach happy or attractive faces and avoid angry, sad, or unattractive faces. In this task, a valenced (positive or negative, e.g., happy or angry face) stimulus appears alongside a neutral stimulus (a black

rectangle). Among several variations, participants in congruent conditions might be asked to approach (i.e., physically reach out and touch) stimuli that were attractive, happy, or otherwise positively valenced; while in incongruent conditions, they would approach negative or angry stimuli.

Our goals in this simulation are: (a) to show the versatility of our model, and how the same network structure and weights used previously can easily accommodate a different laboratory task; (b) to show how damage to socially specific representation might give the appearance of socially specific control processes.

3.2.1. The task

Our simulations address a subset of the manipulations used by Bamford and Ward (2008) and focus on congruent conditions in which participants approach happy faces and incongruent where they approach angry ones.

Recall the model contains domain-specific representations which are relevant to this task, in the form of prepotent connections from visual stimuli to prototypical responses: Specifically, identity units representing happy and angry faces have positive weights to the corresponding approach and avoid intentions. These weights reflect knowledge, presumed to be acquired over much experience, about the usual best courses of action to different aspects of the world. This knowledge means that faces are valenced for the model, and they have positive (approach) and negative (avoid) associations. By contrast, letter stimuli have neutral valence, precisely because they have no direct associations with actions or intentions.

In the absence of any other instruction, the model therefore has a default tendency to approach happy and avoid angry people. However, by activating specific task requirements in the model's hub, we can override this default. On congruent blocks, there is a hub unit activating the "happy" identity and the "approach" intention, reinforcing the prepotent association already present. For incongruent blocks, there is a hub unit activating "angry" and "approach," allowing the model to complete actions contrary to its default.

To respond correctly, it is not enough to simply activate an intention to "approach." An adaptive agent must approach the specific location of a target object. The link between intentions and actions is made possible within the model through "action maps." In this case, the approach intention is associated with a spatial map of all possible locations, which can be approached. We consider the model to have responded successfully if it activates the location within the action map containing the target object. In this way, the stable state arising through relaxation specifies the response location. For example, if the happy face is the target and is presented left of center, then activation of the approach action map in the left-of-center position would mark the correct response.

The model architecture and weights were identical to our first simulations. As in those simulations, we also incorporated domain-specific knowledge in the forms of a bias toward the central location and a bias for animate objects (hands and faces) over inanimate (letters).

3.2.2. Results

Activations for the correct response (approaching the target location) are shown in Fig. 6. We converted activations into response latencies exactly as in our previous simulation (see

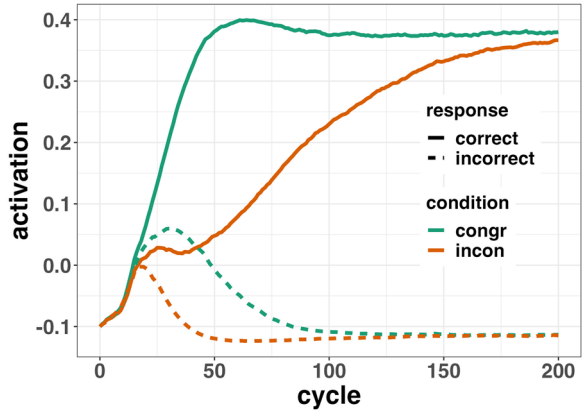

Fig. 6. Mean response activation as a function of task conditions in the approach–avoid task.

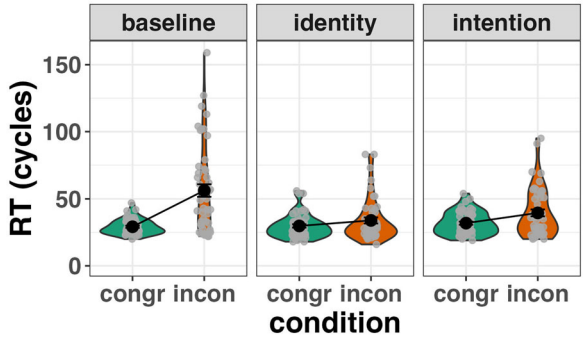

Fig. 7. RT in the approach–avoid task plotted as a function of congruency and with different task settings. congr = congruent; incon = incongruent; baseline = approach–avoid task with standard model settings; identity = approach–avoid task with noise added to identity processing; intention = approach–avoid task with noise added to intention processing. Error bars = standard error of the mean.

Section 3.1.4), to achieve a required overall accuracy of 98%. The resulting RTs and congruency effect are illustrated in Fig. 7 (see baseline panel). Evidently, the model's domain-specific knowledge, in conjunction with domain-general control processes, can produce the typical congruency effect.

3.2.3. Effects of damage and the appearance of domain-specific control

Bamford et al. (2009) suggested that approach and avoidance might be mediated by separate action systems. Their stance was motivated by the work of Davidson and colleagues (e.g., Davidson, 2003), suggesting that approach and avoidance processing is lateralized, the left hemisphere for approach behaviour and the right more involved in avoidance. On the basis of patient data showing reduction of some of the usual congruency effects, Bamford et al. (2009) also argued for lateralized, domain-specific control systems for approach and avoidance.

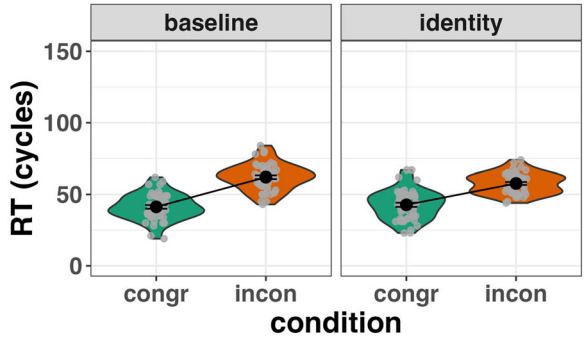

Fig. 8. RT in the imitation task plotted as a function of congruency and task settings. Damage to face stimuli has little effect on baseline performance on the imitation task. congr = congruent; incon = incongruent; baseline = imitation task with standard model settings; identity = imitation task with noise added to face identity units. Error bars = standard error of the mean.

Here, we examine whether different forms of damage to domain-specific representations can give the appearance of impaired domain-specific control. There are many different ways to disrupt the processing of units in an IAC network. We chose simply to increase the noise associated with the activity of units representing either the approach and avoid intentions, or the happy and angry identities: that is, specifically disrupting either the response or stimuli ends associated with approach–avoidance. We will show how, contrary to Bamford et al.'s (2009) reasoning, this kind of result does not mean the two tasks are controlled by different domain-specific processes. A disruption can reduce or eliminate the compatibility effect for the approach task, yet leave unaffected compatibility effects in a different task.

These simulations are summarized in Fig. 7, which shows results when using standard model settings (baseline) and when the effects of disruptive noise "lesions" are applied specifically to identity units (happy, angry) and intention units (approach, avoid). Relative to the baseline, noise disruption virtually eliminated the congruency effect. Perhaps the simplest way to understand this finding is that the signal driving the congruency effect is the prepotent stimulus–response associations between faces and intentions. As the noise within this stimulus-resopnse (SR) chain is increased, the signal-to-noise ratio driving the congruency effect is reduced.

We further confirmed that the effects of the noise "lesion" were specific to the approach task. We compared the response of the lesioned "Identities" network and the normal network on the imitation task used in Simulation 1 (Section 3.1). These results are shown in Fig. 8. As expected, there was no difference in the congruency effect between these two conditions. In the imitation task, unlike the approach task, face identity units remain below the activation threshold in virtually every case, and so additional noise to their activation has very little effect.

Similar to Bamford et al. (2009), we might look at the pattern of results in Figs. 7 and 8 and reason as follows. The "brain lesion" specifically impacted the approach responses to happy and angry faces but did not affect the highly related stimulus–response congruency task of imitation. Therefore, the reasoning might go that the approach task is controlled by a

domain-specific network distinct from the network controlling the imitation task. This kind of reasoning from dissociations is not uncommon, and indeed it is clear that something specific to the approach task has been impaired. But the conclusion is incorrect. Because we are running a simulation, we know the effect of the "lesion" was specific to the *representation* of a stimulus type. The *control* processes—for our model, the process of relaxing to a stable state, and the arbitrary task associations stored in the hub—were not affected. We return to consider the wider implications of this finding in the general discussion.

3.3. Simulation 3: The autistic spectrum and the "first-year puzzle"

Distinctions between domain-specific and general cognition are now important in the study of ASD. ASD is a heterogeneous neurodevelopmental condition, but differences related to control of attention are an important characteristic. These differences relate to social attention, including social motivation (the bias to attend to socially significant stimuli), and joint attention (attending to where others are attending) and also differences in domain-general processes for the control of attention, such as attentional disengagement (e.g., Mundy & Bullen, 2022; Mundy & Newell, 2007).

Because of these social and non-social components, there is ongoing and nuanced debate about the effects of domain-general and domain-specific disruptions in the development of ASD. This debate includes questions about the best way to characterize ASD and related conditions, its early developmental signs, and the ways in which domain-specific and domain-general processes for selection and control of action might interact in typical and atypical function (e.g., Braithwaite et al., 2020).

Debate on the developmental origin of social differences in ASD centers on two perspectives: social-specific theories (e.g., "social first") and domain-general theories (Elsabbagh & Johnson, 2016). Social-specific theories propose that an early socially specific deficit, even in the context of intact domain-general processing, creates knock-on effects disrupting social development (Pelphrey, Shultz, Hudac, & Vander Wyk, 2011). For example, a decrease in social motivation would mean less attention paid to other people, resulting in less social engagement early in life, and ultimately disruption to theory of mind and understanding of others.

The second perspective comes from domain-general theories, which propose that social differences in ASD arise from general cognitive deficits impacting a range of abilities. Domain-general theories recognize that although ASD is at least partly typified by behavioral variation in social contexts, it is not necessary that socially specific cognitive and brain systems are impaired. Social skills might still be differentially affected, compared to other kinds of tasks, as they are generally complex, involve widespread brain systems, and require complex multisensory integration. For example, the nature of joint attention requires general attentional mechanisms of disengagement (e.g., Elsabbagh & Johnson, 2016). This perspective is consistent with the range of domain-general differences seen in ASD (e.g., Robertson & Baron-Cohen, 2017). It also addresses what Elsabbagh and Johnson (2016) refer to as the "first-year puzzle," that is, indicators of ASD in the first year of life can relate more to domain-general differences in attention, perception, and motor skills than to social orienting.

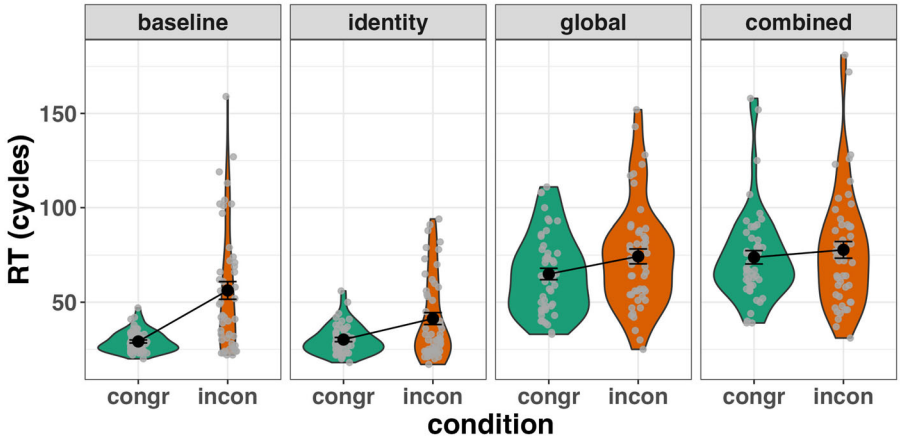

Fig. 9. RT in the approach–avoid task as a function of congruency and with different task settings. congr = congruent; incon = incongruent; baseline: the original intact network; identity: noise disruption to the identity units representing faces (in this case, with less disruption than in Fig. 7); global: increased global noise within the network; combined: the combined effects of the two separate disruptions.

This is an important and complex debate, where we might be able to contribute in two ways. First, our model addresses the important issue raised by Braithwaite et al. (2020): Models of how social-specific and domain-general cognitions might be integrated have been very limited. Our model is meant to address exactly this sort of integration and can help develop alternative ways of thinking about the "first-year puzzle," typical and atypical development, showing combined effects of disruption to control and representation systems.

In Simulation 2, we modeled disruption of approach—avoid effects in a way that would be consistent with a "social first" model. In that simulation, loss of social motivation, as measured by the tendency to approach happy faces, was the result of disrupted social representations. Domain-general control processes for biased competition and the hub were completely intact. We now look at how disruption to domain-general control might exaggerate socially specific deficits.

3.3.1. Effects of domain-general disruption on the use of socially specific representation

Fig. 9 shows the separate and combined effects of noise to specific and general systems. First, let us review the effects of low levels of noise to the identity system. Simulation 2 also applied noise to this system, but here we are using less noise, so we can better observe how this representation-specific disruption will interact with domain-general disruption. Congruency effects for identity disruption (the "identity" panel of Fig. 9), were therefore readily observable, even while reduced, compared to baseline. This pattern looks much like the baseline condition: a clear congruency effect, demonstrating social sensitivity (even if slightly reduced, compared to baseline), and overall latencies very similar to baseline. If we imagine this pattern of results in the context of a behavioral study on young children, showing natural individual variation, the disruption to social representations, which we know to be present here, would be effectively invisible.

We then looked at the effects of increased global noise, on its own (the "global" panel of Fig. 9). By increasing noise throughout the system, we increase the time required to relax into a stable state. Global noise therefore lets us model individual differences in domain-general control: The less noise throughout the system, the more efficient domain-general control processes, such as selection, filtering, and inhibition, will be. In this case, unlike the "identity" panel, the effects of disruption are clearly visible. While there is still demonstrated social sensitivity in the form of a congruency effect (even if reduced somewhat relative to baseline), response latencies are clearly slowed. In the context of tests for early behavioral markers of ASD, a result like this one would correctly flag a deficit in general control, in the relative absence of a social deficit. That is, the congruency effect indicates social sensitivity; but the slowed latencies indicate a general control deficit.

However, when these two forms of disruption are combined (far right panel of Fig. 9), we again see the increased latencies due to disruption to control systems, but now social sensitivity, as reflected in the congruency effect, is almost eliminated.

How does this pattern of results relate to the "first-year puzzle"? Recall that this puzzle is the pattern reported by Elsabbagh and Johnson (2016), in which during the first year, deficits in domain-general control were a better predictor of future ASD than deficits in social-specific processing. A previous proposal is that social tasks are inherently more complex and demanding on domain-general systems than other tasks, and therefore damage or disruption to domain-general systems can produce relatively specific or pronounced social effects (Elsabbagh & Johnson, 2016). One proposal is that social tasks require the simultaneous processing of multiple and sometimes conflicting cues and contexts, straining domain-general systems like attention, memory, and cognitive flexibility. Thus, early deficiencies in domain-general systems make it challenging for children to manage social interactions, making these deficits more noticeable.

Our simulation results suggest an alternative. We do not need to assume that social tasks are inherently more complex than non-social tasks. Instead, we focus on the combined effects of disruption to different cognitive subsystems. That is, domain-general disruption places the system at greater vulnerability to socially specific disruptions. To illustrate, let us assume the minor disruption seen in the "identity" panel of Fig. 9, on its own, would not indicate atypical social behavior. However, if this deficit was concurrent to a domain-general deficit like that in the "global" panel of Fig. 9, it would lead to noticeable deficits (i.e., the "combined" panel of Fig. 9). Based on these results, we suggest that when domain-general deficits arise early in development, the system will be vulnerable to subsequent domain-specific deficits that might otherwise go unnoticed. Or in other words, socially specific deficits require disruption to socially specific systems, but domain-general deficits increase the effect of any specific disruption.

3.4. Simulation 4: Integrating traditional non-social experiments into the model

A strength of our approach is that it integrates the use of socially specific representations into a general-purpose framework for control of behavior. The simulations above were driven by the model's knowledge about the world, encoded in the form of prepotent associations.

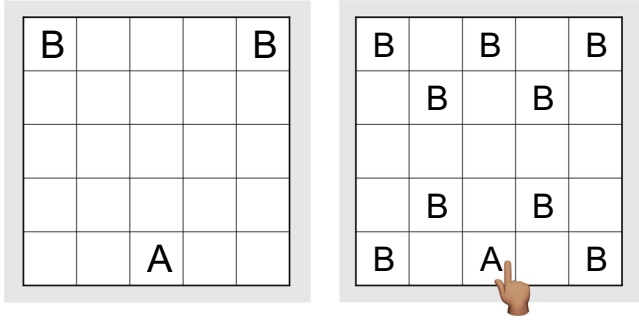

Fig. 10. Visual search task. Two arrays were used, set sizes 3 (left panel) and 10 (right panel). The task was to find the location of target A in an array containing distractor B letters and reach to touch the target's location as indicated in the right panel.

The paradigms being simulated can be placed within the very general category of stimulus–response compatibility or congruency. Here, we present a different task, visual search. Our aim is to illustrate the powerful and general nature of the model's architecture and demonstrate its flexible application to non-social tasks, as well as tasks requiring social representations. We will also show how the relative salience of objects can be changed, and how the trade-off between speed and accuracy is inherent in reaction-time tasks.

Visual search is a classic method for investigating bottlenecks in cognitive performance. A target is to be found among variable numbers of distractors. By increasing the number of distractors, a "search slope" is generated, measuring the increase in response latency with the increase in the total number of items in the array or "set size" (Sternberg, 1966). Search can be highly efficient, with little or no effect of set size, or very costly, in which search slopes approximate the duration of saccades to individual object locations (e.g., Woodman & Luck, 2003). The efficiency of search reflects fundamental constraints and operations of visual processing, including the relative salience of target and distractor (Broadbent, 1982), their featural composition and overlap (Treisman & Gelade, 1980), grouping between and within target and distractor groups (Duncan & Humphreys, 1989), and more (Wolfe, 2020).

3.4.1. The search task

Our task required the model to find a target letter A among distractor Bs, with search arrays of set sizes 3 and 10. A target was present on every trial, and the model was required to "touch" it by activating the location within the touch action map corresponding to the target location as illustrated in Fig. 10. Note that in our simulation, there is no difference between As and Bs other than their labeling. That is, unlike the face identity units (happy associated with approach; angry with avoid), A and B share the same patterns of connections. Responses will be driven to the A by instruction: the active hub unit for this task connects to the identity unit "A," the intention unit "touch," and the effector unit "index finger."

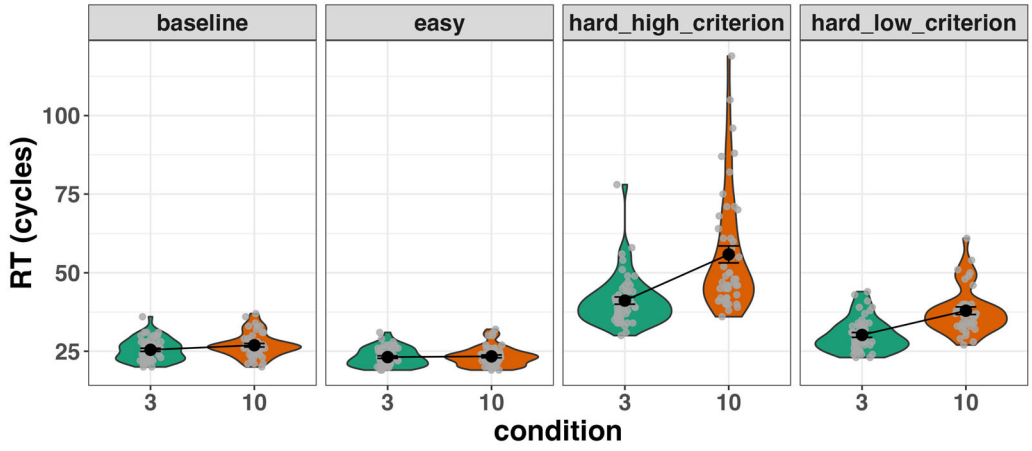

Fig. 11. RT latencies (in simulation cycles; calculated as in Section 3.1.4 to achieve a criterion overall accuracy of 98%). Baseline panels—network parameters were identical to the previous simulations, resulting in a flat search slope. The relative salience of the target was increased (easy panels) or reduced (hard panels). A reduction in saliency is achieved by decreasing competition within the feature maps. This allowed the distractors and target to compete for longer within the salience map, slowing the resolution of competition. The difference between hard_high_criterion and hard_low_criterion is that the criterion for task accuracy was reduced from 98% to 85%. This allows faster responses and a reduced search slope, at a cost to accuracy.

3.4.2. Results: Relative salience and speed-accuracy trade-offs

Our first search simulation was run using the identical network settings and parameters of the previous simulations and produced the flat search slope seen in Fig. 11 (baseline panel). Within the visual search literature, this would indicate a highly efficient search and a highly salient target.

The salience of the target is in part due to the competition in the various spatial maps of the network. Inhibition within the feature maps means that as the number of Bs increases, so too will their mutual inhibition. That is, with set size 3, there are two slightly inhibited Bs, with set size 10, there are nine more thoroughly inhibited objects within the B feature map. Decreased activation within the feature map means decreased activation at the level of the salience map, where representations of all objects in the display will be competing. By reducing the competition within the B feature map, the activity of B objects in the salience map is maintained for longer. This decreases the relative salience of target A, and slows responses, as illustrated in Fig. 11 (hard panels).

A feature of the criterion-based method we use for converting activation patterns into response latencies (Section 3.1.4) is that it is flexible. We can use this criterion in the same way that a person might choose to take extra time so as to ensure accuracy, or sacrifice accuracy to make a faster response. In all simulations so far, we have set the response criterion relatively high, to ensure 98% accuracy for all trials in the task. Functionally, this means that before responding, we require a large difference in the activation between units representing alternative responses. However, we can lower this criterion, allowing faster responses at a cost

to accuracy. This speed-accuracy trade-off is illustrated in Fig. 11. We used the same parameters producing the relatively low salience shown in Fig. 11. But we lowered the response criterion to produce an overall accuracy of 85%. Now although accuracy is lower, latencies and the search slope are clearly reduced.

These are very general patterns of results reflecting fundamental mechanisms in the control of behavior: salience and response criteria. Looking across our simulations, we can see that salience can arise from interactions and representations throughout the model, from the within-module competition seen here to the effects of prior knowledge demonstrated in Simulation 1. Likewise, the speed-accuracy trade-off shown here is inherent within the ROC curves shown in Fig. 3. The ROC shows how, for a given level of observer sensitivity, the observer can change their response criteria to optimize speed, accuracy, or some desired joint function of the two.

4. Discussion

4.1. Control versus representation

The most important conclusion from our work is that simply because a task requires control of social behavior does not mean it requires socially specific control processes. Our simulations show how general control processes operating on social representations replicate important patterns of behavior in social laboratory tasks. Although computational models have been put forward to account for effects within the automatic imitation task (Cooper, Catmur, & Heyes, 2013; Cracco & Cooper, 2019), our approach is novel in addressing the key distinction between representation and control, across a variety of experimental paradigms.

This distinction is also important in assessing deficits and disruptions to social cognition because it is not always obvious how the structure and function of cognitive and brain systems vary across individuals. Our simulations demonstrate that disruption to social representations can give the appearance of a deficit in socially specific control, despite intact control systems. Indeed, we have made a similar erroneous inference ourselves (Bamford et al., 2009).

4.2. The default stance

Our simulations therefore support the claim that the "default stance" in social cognition should be that control is general, but representation is specific (Ramsey & Ward, 2020). As mentioned earlier, this is not to say that domain-specific control systems cannot exist. But in our view, domain-specific control processing would be the exception rather than the presumption, and such extraordinary claims would require extraordinary evidence.

To illustrate, the difference between representations and control is crucial not only for interpreting findings but for deciding research questions. Our interpretation of Brass et al.'s (2000) imitation experiments is essentially that biased competition operates on prepotent SR associations. By this interpretation, the imitation results fit neatly into a much larger literature on SR compatibility. In contrast, the dominant interpretation is that finger responses are being controlled in part by a theory-of-mind network, operating to distinguish the self from other

(Brass et al., 2001, 2009; Sowden & Shah, 2014; Spengler et al., 2009; Wang & Hamilton, 2012).

But when one starts to unpack the idea of a socially specific control system, a range of basic architectural issues arise: How do control processes relating to the theory-of-mind network operate with respect to general control systems? Do they share resources? Are they independent? How are the two systems integrated so that tasks involving social and non-social stimuli can be performed? How do the decades of research describing general principles of attention, executive control, and action relate to the theory-of-mind system?

We have not seen these kinds of questions explicitly addressed in claims around specialized social control. Yet these are the kinds of questions that need answering to overcome what we would argue is appropriate skepticism about a special-purpose control system. As such, much like other tasks that claim to measure processes associated with theory-of-mind, evidence for the specificity or validity of the claims being made is often lacking (Higgins, Ross, Langdon, & Polito, 2022; Higgins, Ross, Polito, & Kaplan, 2023) and can be accounted for by alternative, lower-level or more general processes (Quesque & Rossetti, 2020). Therefore, we support recent suggestions across the board in psychology to take measurement and theory construction far more seriously than is currently practiced (Flake & Fried, 2020; Proulx & Morey, 2021).

4.3. The value of simulating "damage"

Our model was designed as a general framework for perception and action and not as a specific model of atypical function. However, comparing model performance when aspects of the ICON model are "damaged" can be instructive for understanding cognitive performance in typical development and development disorders. For example, in Simulation 2, the representation of faces was disrupted, but control systems were working normally. Nonetheless, based on the resulting reaction time (RT) data in a distractor interference paradigm, one could easily be tempted to conclude that attentional control systems were faulty (Figs. 7 and 8). Indeed, distracter interference is often used to define a loss in attentional function (Moore & Zirnsack, 2017). Thus, a reduction in the strength or quality of visual representation can be sufficient to produce the illusion of an attentional deficit, despite completely intact systems of attention (Squire, Noudoost, Schafer, & Moore, 2013). These effects occur because of the architecture of our model, in which perceptual representations propagate through a connected network via biased competition. Such architectural principles are not unique to the ICON model but are found across many cognitive neuroscience models of attention and vision in both mature systems (Desimone & Duncan, 1995; Moore & Zirnsack, 2017; Reynolds & Heeger, 2009) and developing systems (Amso & Scerif, 2015).

4.4. Integrating social cognition into a wider world of domain-general control processing

Our general approach toward simulating social cognition has been, first and foremost, to develop a general-purpose architecture, capable of simulating a variety of laboratory tasks. Then, within that general architecture, we insert social (and non-social) cognition

experiments. This approach ensures we maintain a unified account of social and non-social task processing.

Our model is therefore a vehicle for exploring how general control systems can impact laboratory tasks with social relevance. We believe this is important because the wealth of research on attention and control of action is a valuable resource, hard-fought and won. Rather than turning first toward novel mechanisms and explanations, as a field, social cognition can turn toward this knowledge base. For example, we demonstrated how the influence of social representation, established in one version of an experiment, can virtually vanish when distractor salience is reduced (as in Fig. 4) or because participants' criteria for accuracy are altered (as in Fig. 3b). Like others recently (Botvinick & Braver, 2015), we see a lot of value to be added by taking core principles and findings from research on general control mechanisms and seeing how they mesh with and inform processes that may mimic real-world behaviors, such as motivation.

4.5. Models and theories

This work is also a direct response to calls for more systematic theory construction in psychology (Proulx & Morey, 2021). One advantage of more formal modeling approaches is that it forces researchers to be explicit about the relationship between parts of the system under investigation, which is valuable when comparing and testing theories (Hintzman, 1991; McClelland & Rumelhart, 1981; Newell, 1992). By doing so we minimize the reliance on narrative accounts of psychological theories, which are easier to misconstrue and therefore harder to evaluate, revise, and update in an efficient manner (Yarkoni, 2022).

Along with other emerging computational social neuroscience approaches (Charpentier & O'Doherty, 2018; Cheong et al., 2017; Hackel & Amodio, 2018; Lockwood & Klein-Flugge, 2021), we see considerable value in providing a platform for others to prove the limitations of our approach. Given that the entire model is open and available to download and run using freely available software also encourages other researchers to build upon this work in the future, which is consistent with the principle that we should try to make science as open and transparent as possible (Munafò et al., 2017).

4.6. Limitations

As argued by Simons, Shoda, and Lindsay (2017), we feel it is valuable to explicitly state the limitations of our approach, and likely constraints on the generality of our findings, to provide a more stable and transparent platform for future research (Ramsey, 2021).

4.6.1. We chose to focus on one type of modeling approach

Our aim was not to optimize performance or to match specific datapoints but to understand the capabilities of a particular cognitive architecture. We therefore focussed on general patterns and general parameters—we have not mucked about with details. For this reason, except when explicitly exploring parameter effects, we use a single set of weights, a single set of network parameters, and a consistent set of external activations, that is, the input to visual objects and hub units was the same across tasks. Furthermore, we have focussed on

cognitive modeling, not statistical learning. This is modeling a specific architecture rather than using statistical learning or other techniques to uncover new solutions (e.g., Ward & Ward, 2008). As experimenters, we supplied the network architecture, and the network itself with knowledge about SR associations. There is no learning taking place in the simulations here, although that may be a valuable extension in future research.

The range of simulations we cover suggests the versatility of the model's architecture. However, even within the realm of a single paradigm like the automatic imitation task (Simulation 1), we have made simplifications. The compatibility effects in our model result from the effects of prepotent associations acting with or in opposition to the transient associations established by the hub. The current model has no prepotent associations between spatial codes and any of the response or effector units. This means we do not simulate the effects of spatial compatibility, which have been identified in the Brass et al. (2001) task (Catmur & Heys, 2011). In contrast, Cooper et al. (2013) model both imitative compatibility (based on whether the distractor finger is index or middle) and spatial compatibility (whether the distractor finger is relatively left or right) in the Brass et al. task and include detailed modeling of the time-course of compatibility effects. Cooper et al. also support the idea of domain-general control processing in this task.

4.6.2. The hub

The hub is central to the operation of the network, allowing arbitrary goals to control behavior. The hub in the model was inspired by recent neuroscientific approaches to flexible hubs in the human brain (Cole et al., 2013), as well as principles from computational neuroscience (Botvinick & Cohen, 2014). The hub as shown here raises two important issues though that extend beyond the scope of the model. First, if the hub is responsible for maintaining the task set, what homunculus is responsible for activating the hub? For example, in the imitation task, we assume the appropriate units in the hub are activated for the task (e.g., the unit representing the combination of "button-press," "index," and "A"). That is, the hub establishes and maintains the task set. But how would the hub know which units to activate for the current set? Some other (domain-general) process would need to decide that button-pressing to A's is in order. We have shown how a hub allowing arbitrary task activations can be integrated into a general action selection system but not how the hub units themselves come to be active. We have simply done that as programmers. Second, the hub is meant to understand all the ways that deliberate action can be made to all kinds of objects. Within the limited world here, that is easy enough. But how would the innumerable combinations be handled in a hub at scale? We do not have answers to these questions.

4.6.3. Specific and general

Why should we call the face identity units "social"? In the brain, socially specific visual areas, such as single cells for faces (Perrett, Rolls, & Caan, 1982), or extrastriate brain regions for bodies (Downing, Jiang, Shuman, & Kanwisher, 2001), are typically reported in terms of a response profile or tuning curve, showing how the intensity of response varies based on stimulus categories and properties. Specificity and intensity are typically graded. For example, a cell in the superior temporal sulcus might respond very strongly to a face and more weakly

to a brush (Desimone, Albright, Gross, & Bruce, 1984). Our identity units (within the "What" module) reflect extreme versions of such tuning profiles: The face and finger units are "social" in that they represent visual categories of the social world and "specific" in that they represent nothing else. The category specificity of the identity units can be contrasted to the generality of the model's spatial salience map, which represents the locations of all categories of objects.

As the model has no dedicated semantic store, the meaning of those objects, beyond what actions can be performed on them, is very limited. The model's social identity units could still be argued to have limited semantics by virtue of their prepotent associations. For example, happy faces have associations with approach actions. In a more detailed model, these associations could be elaborated to include social actions like a smile or an adjustment in posture. However, the focus of our model is on control of action, and rich semantic content is well outside our current scope.

One could reasonably question whether any representation can be entirely domain-general, as long as we are free to vary what we call a "domain." For example, we described early visual regions as domain-general, but this is only within the domain of vision. Similarly, we might "zoom in" to say this region or those units show category-specific tuning, and "zoom out" to say those same processors are part of a larger and more general system. Ultimately, these definitional issues do not change the fact that in the brain, as in our model, there are specialized processors for different categories of objects.

In contrast, it seems clear to us that, within our model, biased competition is a domain-general control process and cannot be reasonably characterized as domain-specific. Biased competition is an emergent property inherent within the very architecture, and it affects every unit and every module. The hub module we also see as domain-general but in a different way. We think of the hub module, in a way that is similar to Cole and colleagues' view of hubs in the human brain (e.g., Cole et al., 2013; Cocuzza, Ito, Schultz, Bassett, & Cole, 2020), as a structure that can effectively reconfigure the connectivity of the system, in a flexible and arbitrary way, in line with current goals.

Taking these considerations into account, we might rephrase our default stance from "domain-general control operating on domain-specific representation" to "domain-general control processes operating on representations specialized for different object categories." That might be technically more accurate, although it might make our general view less clear.

4.6.4. What is a "real" social task?

As we have tried to demonstrate, our model can handle a wide range of laboratory tasks, of the kind that involve a single participant, working with a display and buttonbox, on some well-defined procedure. That is true whether the task involves socially specific stimuli or not. Are such tasks truly "social"? In an important sense, they are very clearly and obviously social tasks, because they investigate and reveal the function and structure of social representations (even if not social control systems). We have emphasized the prepotent associations between social representations and actions, and how these associations need to be controlled. However, in other important ways, they are not particularly social tasks. We agree with the recent framework proposed by Quesque and Rossetti (2020), who argue that many "theory of mind" tests do not require theory of mind at all. They suggest a valid measure of theory of mind would

require the participant to represent the mental state of another person. Tasks like inhibiting a distracting animate object and approaching a positively valenced object do not require a representation of mental state. The true theory of mind tasks would be something that our model could not possibly address, for example, the Movie for the Assessment of Social Cognition (Dziobek et al., 2006), in which participants explain the motivations and feelings of characters in short videos. As such, more social and cognitive neuroscience research that can more effectively span the lab to real life would be welcomed as others have argued previously (Kingstone, Smilek, & Eastwood, 2008).

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/m2enf/ and https://github.com/rich-ramsey/icon-model.

References

Amso, D., & Scerif, G. (2015). The attentive brain: Insights from developmental cognitive neuroscience. *Nature Reviews Neuroscience*, *16*, 606–619. https://doi.org/10.1038/nrn4025

Arnell, K. M., & Jenkins, R. (2004). Revisiting within-modality and cross-modality attentional blinks: Effects of target–distractor similarity. *Perception & Psychophysics*, *66*(7), 1147–1161. https://doi.org/10.3758/BF03196842

Bamford, S., Turnbull, O. H., Coetzer, R., & Ward, R. (2009). To lose the frame of action: A selective deficit in avoiding unpleasant objects following a unilateral temporal lobe lesion. *Neurocase*, *15*(4), 261–270. https://doi.org/10.1080/13554790802680313

Bamford, S., & Ward, R. (2008). Predispositions to approach and avoid are contextually sensitive and goal dependent. *Emotion*, *8*(2), 174–183. https://doi.org/10.1037/1528-3542.8.2.174

Barrett, H. C. (2012). A hierarchical model of the evolution of human brain specializations. *Proceedings of the National Academy of Sciences*, *109*(Supplement1), 10733–10740. https://doi.org/10.1073/pnas.1201898109

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, *49*(10), 1154–1165. https://doi.org/10.1016/j.visres.2008.07.012

Binney, R. J., & Ramsey, R. (2020). Social semantics: The role of conceptual knowledge and cognitive control in a neurobiological model of the social brain. *Neuroscience & Biobehavioral Reviews*, *112*, 28–38. https://doi.org/10.1016/j.neubiorev.2020.01.030

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*(1), 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Braithwaite, E. K., Gui, A., & Jones, E. J. H. (2020). Chapter 13—Social attention: What is it, how can we measure it, and what can it tell us about autism and ADHD? In S. Hunnius & M. Meyer (Eds.), *Progress in brain research* (Vol. 254, pp. 271–303). Amsterdam: Elsevier. https://doi.org/10.1016/bs.pbr.2020.05.007

Brass, M., Bekkering, H., Wohlschlager, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: Comparing symbolic, spatial, and imitative cues. *Brain and Cognition*, *44*(2), 124–143.

Brass, M., Ruby, P., & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2359–2367. https://doi.org/10.1098/rstb.2009.0066

Brass, M., Zysset, S., & von Cramon, D. Y. (2001). The inhibition of imitative response tendencies. *Neuroimage*, *14*(6), 1416–1423. https://doi.org/10.1006/nimg.2001.0944

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Broadbent, D. E. (1982). Task combination and selective intake of information. *Acta Psychologica*, *50*(3), 253–290. https://doi.org/10.1016/0001-6918(82)90043-9

Catmur, C., & Heyes, C. (2011). Time course analyses confirm independence of imitative and spatial compatibility. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 409–421. https://doi.org/10.1037/a0019325

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience*, *13*(6), 637–647. https://doi.org/10.1080/17470919.2018.1518834

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models in social neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior* (pp. 229–244). Hoboken, NJ: John Wiley & Sons.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979. https://doi.org/10.1121/1.1907229

Cocuzza, C. V., Ito, T., Schultz, D., Bassett, D. S., & Cole, M. W. (2020). Flexible coordinator and switcher hubs for adaptive task control. *Journal of Neuroscience*, *40*(36), 6949–6968.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, *16*, 1348–1355. https://doi.org/10.1038/nn.3470

Cooper, R. P., Catmur, C., & Heyes, C. (2013). Are automatic imitation and spatial compatibility mediated by different processes? *Cognitive Science*, *37*(4), 605–630. https://doi.org/10.1111/j.1551-6709.2012.01252.x

Cracco, E., & Cooper, R. P. (2019). Automatic imitation of multiple agents: A computational model. *Cognitive Psychology*, *113*, 101224. https://doi.org/10.1016/j.cogpsych.2019.101224

Davidson, R. J. (2003). Affective neuroscience and psychophysiology: toward a synthesis. *Psychophysiology*, *40*, 655–665.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*, 2051–2062.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458. https://doi.org/10.1037/0033-295X.96.3.433

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current Opinion in Neurobiology*, *7*(2), 255–261. https://doi.org/10.1016/S0959-4388(97)80014-1

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, *387*(6635), 808–810.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A Movie for the Assessment of Social Cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

Elsabbagh, M., & Johnson, M. H. (2016). Autism and the social brain: The first-year puzzle. *Biological Psychiatry*, *80*(2), 94–99. https://doi.org/10.1016/j.biopsych.2016.02.019

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioural Reviews*, *24*(6), 581–604.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*, 456–465. https://doi.org/10.1177/2515245920952393

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science*, *12*(5), 731–741. https://doi.org/10.1177/1745691617691949

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92–97.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, MA: MIT Press.

Haugeland, J. (1992). Representational genera. In L. Burkholder (Ed.), *Philosophy and the computer* (pp 105–134). London: Routledge.

Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2022). The "Reading the Mind in the Eyes" test shows poor psychometric properties in a large, demographically representative U.S. sample. *Assessment*, *30*(6), 10731911221124342. https://doi.org/10.1177/10731911221124342

Higgins, W. C., Ross, R. M., Polito, V., & Kaplan, D. M. (2023). Three threats to the validity of the Reading the Mind in the Eyes Test: A commentary on Pavlova and Sokolov (2022). *Neuroscience & Biobehavioral Reviews*, *147*, 105088. https://doi.org/10.1016/j.neubiorev.2023.105088

Hintzman, D. L. (1991). Why are formal models useful in psychology? In S. Lewandowsky (Ed.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hubel, D. H., & Wiesel, T. N. (1977). Ferrier lecture—Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *198*(1130), 1–59. https://doi.org/10.1098/rspb.1977.0085

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, *99*, 317–340.

Klein, J. T., Shepherd, S. V., & Platt, M. L. (2009). Social attention and the brain. *Current Biology*, *19*(20), R958–R962. https://doi.org/10.1016/j.cub.2009.08.010

Kramer, R. S. S., Mulgrew, J., Anderson, N. C., Vasilyev, D., Kingstone, A., Reynolds, M. G., & Ward, R. (2020). Physically attractive faces attract us physically. *Cognition*, *198*, 104193. https://doi.org/10.1016/j.cognition.2020.104193

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50–59. https://doi.org/10.1016/S1364-6613(99)01436-9

Lockwood, P. L., & Klein-Flügge, M. (2021). Computational modelling of social cognition and behaviour—A reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, *16*(8), 761-771. https://doi.org/10.1093/scan/nsaa040

Markman, A. B., & Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychological Science*, *16*(1), 6–10. https://doi.org/10.1111/j.0956-7976.2005.00772.x

McClelland, J. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, *4*, 503. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00503

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science*, *38*(6), 1139–1189. https://doi.org/10.1111/cogs.12146

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Michael, J., & D'Ausilio, A. (2015). Domain-specific and domain-general processes in social perception—A complementary approach. *Consciousness and Cognition*, *36* (Supplement C), 434–437. https://doi.org/10.1016/j.concog.2014.12.009

Moore, T., & Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annual Review of Psychology*, *68*(1), 47–72. https://doi.org/10.1146/annurev-psych-122414-033400

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. https://doi.org/10.1038/s41562-016-0021

Mundy, P., & Bullen, J. (2022). The bidirectional social-cognitive mechanisms of the social-attention symptoms of autism. *Frontiers in Psychiatry*, *12*, 752274. https://www.frontiersin.org/articles/10.3389/fpsyt.2021.752274

Mundy, P., & Newell, L. (2007). Attention, Joint attention, and social cognition. *Current Directions in Psychological Science*, *16*(5), 269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1016/S0364-0213(80)80015-2

Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, *13*(3), 135–143.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Pelphrey, K. A., Shultz, S., Hudac, C. M., & Vander Wyk, B. C. (2011). Research review: Constraining heterogeneity: The social brain and its development in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *52*(6), 631–644. https://doi.org/10.1111/j.1469-7610.2010.02349.x

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, *47*, 329–342.

Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, *16*(4),671–681. https://doi.org/10.1177/17456916211017098

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384–396. https://doi.org/10.1177/1745691619896607

Ramsey, R. (2018). What are reaction time indices of automatic imitation measuring? *Consciousness and Cognition*, *65*, 240–254. https://doi.org/10.1016/j.concog.2018.08.006

Ramsey, R. (2021). A call for greater modesty in psychology and cognitive neuroscience. *Collabra: Psychology*, *7*(1), 24091. https://doi.org/10.1525/collabra.24091

Ramsey, R., & Ward, R. (2020). Putting the nonsocial into social neuroscience: A role for domain-general priority maps during social interactions. *Perspectives on Psychological Science*, *15*, 1076–1094. https://doi.org/10.1177/1745691620904972

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185. https://doi.org/10.1016/j.neuron.2009.01.002

Riddoch, M. J., Edwards, M. G., Humphreys, G. W., West, R., & Heafield, T. (1998). Visual affordances direct action: Neuropsychological evidence from manual interference. *Cognitive Neuropsychology*, *15*(6-8), 645–683. https://doi.org/10.1080/026432998381041

Robertson, C. E., & Baron-Cohen, S. (2017). Sensory perception in autism. *Nature Reviews Neuroscience*, *18*(11), 11. https://doi.org/10.1038/nrn.2017.112

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart, J. L. McClelland, University of California, San Diego PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure, Vol. 2: Psychological and biological models* (pp. 7–57). Cambridge, MA: MIT Press.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology*, *59*(4), 239–245. https://doi.org/10.1037/h0047274

Sowden, S., & Shah, P. (2014). Self-other control: A candidate mechanism for social cognitive function. *Frontiers in Human Neuroscience*, *8*, 789. https://doi.org/10.3389/fnhum.2014.00789

Spengler, S., von Cramon, D. Y., & Brass, M. (2009). Control of shared representations relies on key processes involved in mental state attribution. *Human Brain Mapping*, *30*(11), 3704–3718. https://doi.org/10.1002/hbm.20800

Spunt, R. P., & Adolphs, R. (2017). A new look at domain specificity: Insights from social neuroscience. *Nat Rev Neurosci*, *18*(9), 559–567. https://doi.org/10.1038/nrn.2017.76

Squire, R. F., Noudoost, B., Schafer, R. J., & Moore, T. (2013). Prefrontal contributions to visual selective attention. *Annual Review of Neuroscience*, *36*(1), 451–466. https://doi.org/10.1146/annurev-neuro-062111-150439

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654. https://doi.org/10.1126/science.153.3736.652

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

van Elk, M., van Schie, H., & Bekkering, H. (2014). Action semantics: A unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge. *Physics of Life Reviews*, *11*(2), 220–250. https://doi.org/10.1016/j.plrev.2013.11.005

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628.

Wang, A., Qian, Q., Zhao, C., Tang, X., & Zhang, M. (2022). Modal-based attention modulates attentional blink. *Attention, Perception, & Psychophysics*, *84*(2), 372–382. https://doi.org/10.3758/s13414-021-02413-y

Wang, Y., & Hamilton, A. F. de. C. (2012). Social Top-down response modulation (STORM): A model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience*, *6*, 153. https://doi.org/10.3389/fnhum.2012.00153

Ward, R. (1999). Interaction between perception and action systems: A model for selective attention. In G. W. Humphreys, J. Duncan, A. Treisman, Royal Society (Great Britain), Novartis Foundation for Gerontological Research (Eds.), *Attention, space, and action: Studies in cognitive neuroscience* (pp. 311–322). Oxford, England: Oxford University Press.

Ward, R., & Ward, R. (2008). Selective attention and control of action: Comparative psychology of an artificial, evolved agent and people. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1165–1182. https://doi.org/10.1037/0096-1523.34.5.1165

Ward, R., & Ward, R. (2009). Representation in dynamical agents. *Neural Networks*, *22*(3), 258–266. https://doi.org/10.1016/j.neunet.2009.03.002

Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? *Annual Review of Vision Science*, *6*(1), 539–562. https://doi.org/10.1146/annurev-vision-091718-015048

Woodman, G. F., & Luck, S. J. (2003). Serial deployment of attention during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 121–138. https://doi.org/10.1037/0096-1523.29.1.121

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685