TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 11 (2019) 281–298

ISSN: 1756-8765 online
DOI: 10.1111/tops.12424

Editors' Introduction: Computational Approaches to Social Cognition

Fiery Cushman, Samuel Gershman

Department of Psychology, Harvard University

Received 21 February 2019; received in revised form 19 March 2019; accepted 22 March 2019

Abstract

What place should formal or computational methods occupy in social psychology? We consider this question in historical perspective, survey the current state of the field, introduce the several new contributions to this special issue, and reflect on the future.

Keywords: Computational social science; Social cognition; Computational cognitive science

1. Introduction

Social psychology was founded on two great ideas, one widely known and another mostly forgotten. The first idea is what we ought to study: ourselves. Social psychologists do not study isolated systems (memory, language, or vision), isolated behaviors (saccades or lever-presses), or isolated physical mechanisms (neurons or networks). Rather, social psychologists study complete individuals in context—"human persons" (Allport, 1961); "persons as wholes" (Lewin, 1936)—and their interactions with each other. The basic questions that animate social psychology—who we love, despise, or neglect; how we make sense of others, and ourselves; why we strive for goodness, and why we often fall short—cannot be expressed without reference to persons: I, you, we, and they. Rooted in Gestalt theory, this holistic approach has branched into a wide canopy of inquiry covering every aspect of human thought, feeling, and action.

The second idea is how we ought to study ourselves. Social psychology's founders aspired to organize the astonishing complexity of human thought around new formal concepts and theories. In other words, their new field was intended not only to make the topics of psychological inquiry broader, but also to make its theoretical underpinnings

Correspondence should be sent to Fiery Cushman, Harvard University, 1480 William James Hall, 33 Kirkland St., Cambridge, MA 02138. E-mail: cushman@fas.harvard.edu

deeper. Kurt Lewin (1936) lamented the "mere piling up facts" practiced by his contemporaries, which "can only lead to a chaotic and unproductive situation." He strained against the temptations of our naive, prescientific psychological concepts—what today we would call "folk psychology." He perceived, instead, that scientific insight often depends on new concepts and conceptual relations. He imagined that formal methods, such as mathematics, could help us to discover and specify them.

This vision captivated a fledgling field. *The Psychology of Interpersonal Relations* (1958), Heider's masterpiece that inspired attribution theory, echoed Lewin's call: "[W]e shall not attain a conceptual framework by collecting more experimental results. Rather, conceptual clarification is prerequisite for efficient experimentation. . . . Such systematization is an important feature of any science and reveals relationships among highly diverse events." Likewise, Festinger's *A Theory of Cognitive Dissonance* (1957) does not begin with a lengthy discussion of some surprising new fact, but rather the quick observation of a well-worn one: "It has frequently been implied, and sometimes even pointed out, that the individual strives towards consistency within himself." Festinger assumed his project was not to document something new and surprising, but rather to explain an obvious thing more deeply. Festinger promised to provide ". . . a more formal exposition of the theory of dissonance. I will attempt to state the theory in as precise and unambiguous terms as possible."

Put simply, the founders of social psychology did not envision a new field that would principally contribute new "effects" (experiments and their discoveries). Such phenomena are usually apparent enough: We live them! Rather, they envisioned a new field whose most valuable contributions would be novel concepts, theories, and formal structures. These would be broad, general principles of mental and social organization of the kind necessary to structure every branch of psychological science, including cognitive, developmental, and clinical research. They asked us, in short, to discover the hidden logic of familiar things.

How are we doing? As a topic, social psychology is flourishing. The field itself has never been more vibrant, or its inquiries more varied. It also runs a brisk export business: Social psychology's motivating questions are ascendant in every corner of psychological science, from cognitive and comparative to developmental and clinical.

Strangely, however, the approach intended by the founders of social psychology is in decline. Contemporary research emphasizes facts: experiments and their results. The ideal experiment delivers a large effect due to a small, seemingly innocuous manipulation. The ideal result is counterintuitive, perhaps from a scientific standpoint but certainly according to lay theory. The ideal theory does not require new and unfamiliar concepts, but instead offers a striking composition of familiar concepts; it should be instantly accessible to a college undergraduate. These ideals are often implicit but sometimes explicit (e.g., Gray & Wegner, 2013). In short, while social psychology was founded in order to discover the hidden logic of familiar things, today it often strives for transparent descriptions of counterintuitive things.

It is hard to say why. Perhaps, the cumulative effect of world events—the second world war, the holocaust, and the civil rights movement—made practical applications more urgent than theoretical progress. Or perhaps, fault lies with the founders themselves, whose actual theoretical contributions fell well short of their ambitions. Lewin, for instance, is rightly celebrated for observing that a person's behavior is jointly determined by aspects of the person and also of the environment. It is not clear, however, what was gained when he stated this formally as $B = f(P, E)$. Here, it seems, the idea itself gains nothing from formalization. Similarly, Heider's baroque notational scheme for folk psychology never caught on, and Festinger's attempt to formalize some concepts underlying cognitive dissonance went unused, even as the phenomenon became indispensable. These early efforts seemed not to make important ideas accessible but, ironically, to obscure them. For one reason or another, a general feeling emerged that, for social psychology, formal concepts and methods are at best unnecessary and at worst an impediment.

Of course, nobody denies that formal tools are indispensable to other fields of scientific research. Newton's, Boyle's, and Mendel's laws are fundamental not merely because they summarize some body of facts, nor because they make quantitative predictions, but rather because their formal structure allows us to simplify and organize our understanding of complex processes. Formal theories give order to disorderly things, from avalanches to thunderstorms to heritability. In order to use the theories, you must master new concepts and operations. The return on this investment, however, is to perceive the hidden logic of familiar things—even things that once seemed to defy any form of logic. The concern, then, is not that there is something generally useless about formal approaches in science.

Rather, the concern is that there is something specifically useless about formal approaches in social psychology. When theories of social psychology used plain English and ordinary folk concepts, they flourished; when theories constructed new concepts, used formal tools, and aimed for broad, general, and abstract theories, they floundered. Perhaps, social psychology is just different than other fields of study. Perhaps, avalanches, thunderstorms, and inheritance have the kind of hidden logic that demands new formal concepts and theories. Perhaps, individuals and their interactions—"whole persons"—are best understood by rearranging familiar concepts in new ways to explain counterintuitive things.

2. Three successful frameworks

During the decades when social psychology mostly turned away from formal approaches, however, neighboring fields began to train their talents on Lewin's goal: establishing an abstract and formal language to organize the study of human individuals and their interactions. Three areas of research deserve special attention: inference (how we form beliefs), choice (how we assign values), and strategic interaction (how our thoughts and actions influence each other). The contributions to this specific issue, among many others, illustrate how formalisms from these three areas can support our understanding of social cognition.

2.1. Inference: Bayesian cognitive models

Suppose that you are in a cafe and overhear a young woman say, “I’m leaving you,” to which her young male companion replies, “Who is he?”. From this faint sketch, you can form a remarkably detailed image of their relationship: past, present, and future. How is this accomplished?

One approach to this problem is some form of deductive reasoning (i.e., the application of rules: “If a woman says, ‘I’m leaving you’, it’s a romantic breakup.”). Deductive rules assume that the conclusion is logically entailed by the premises, but people make many inferences that do not satisfy such a strong requirement. A woman saying “I’m leaving you” does not, after all logically entail a romantic breakup; she might simply be leaving the cafe. Nonetheless, the latter conclusion seems implausible, and inductive reasoning offers a way of formalizing the degree of support for the conclusion. The most well-studied form of inductive reasoning is the probability calculus, according to which the degree of support for a conclusion corresponds to the conditional (posterior) probability of the conclusion given the premises.

Viewed as a psychological theory (Griffiths, Kemp, & Tenenbaum, 2008), the probabilistic framework makes three intertwined claims about human inductive reasoning:

1. We construct generative models of the world (see also Friston, 2010; Rao & Ballard, 1999). In other words, these models describe probabilistic (possibly causal) relations, and we can use them to derive the hypothetical consequences of various states of the world that are not directly observable. For instance, we might have the model: “Infidelity causes breakups.” If you input a representation of infidelity, the model “generates” predicted data: a breakup. Or, if you input the visual percept of a ball flying toward a window, the model generates predicted data: a shattering crash.
2. These models are probabilistic. That is, we represent probability distributions over the possible states of the world and the transitions between them. For instance, we might represent that breakups are rare, infidelity is rare, but breakups commonly follow from infidelity.
3. Inference involves inversion of the generative model, which is accomplished by application of Bayes’ rule. Generically, suppose we are attempting to infer the probability of some hypothesis given data. This desired posterior P(hypothesis | data) is calculated by multiplying the prior probability of the hypothesis P(hypothesis) and the likelihood that the hypothesis would generate the data P(data | hypothesis), then normalizing so the resulting probabilities sum to 1. In other words, you ask: “Which state of the world is most likely to have generated the actual data I observed?” When you observe a breakup, you ask: What are the most likely states of the world that would have caused such a breakup to occur? Infidelity is a likely cause; an ill-considered April fool’s joke is less likely. Bayes’ rule parses this belief into a function of the prior probability (“How common is infidelity vs. April fool’s jokes?”) and the likelihood (“How consistent is this conversation with infidelity vs. an April fool’s joke?”).

> These elements enable a powerful set of cognitive operations. We are able to predict future states of the world by applying our causal models, plan actions to maximize reward and gain information, and learn causal models from experience.

Consider the application of Bayesian methods to a classic social–psychological finding in the trait inference literature. Early studies of trait inference identified an apparent "negativity bias": If you are told that a person did a bad thing (e.g., stealing candy from a child), you infer that they are a very bad person, whereas if they did a good thing (giving candy to a child), you do not make as strong an inference that they are a very good person (Anderson, 1965; Birnbaum, 1972). Yet subsequent studies uncovered contexts in which the reverse bias seemed to apply. For instance, if you are told a person did a smart thing (e.g., programmed an iPhone app), you infer that she is a very smart person, whereas if she did a dumb thing (e.g., threw out her own wallet), you do not make as strong an inference that she is a dumb person (Martijn, Spears, Van der Pligt, & Jakobs, 1992).

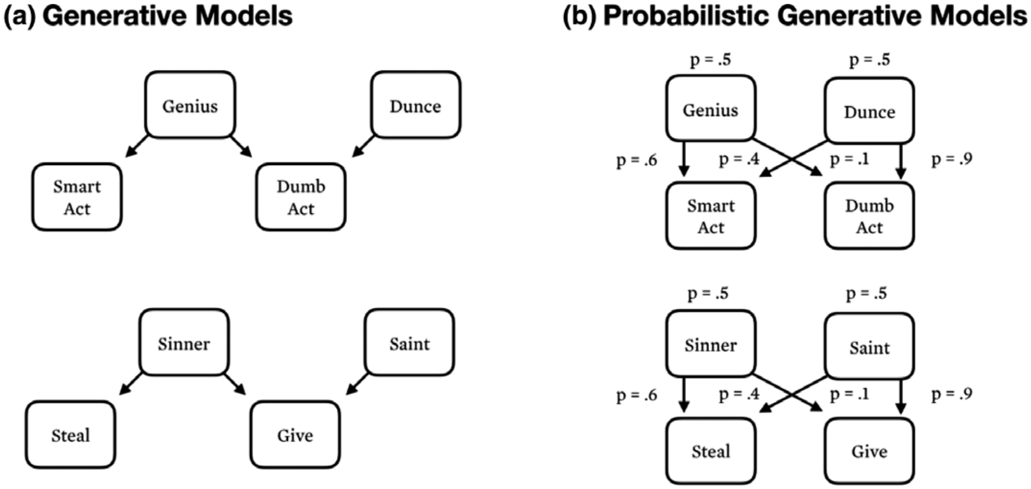
The prevailing explanation for such findings centers on the idea of "diagnosticity" (Reeder & Brewer, 1979; Skowronski & Carlston, 1989). Doing a bad thing is highly diagnostic of being a bad person, whereas doing a good thing is only weakly diagnostic of being a good person. Yet doing a smart thing is highly diagnostic of being a good person, while doing a dumb thing is only weakly diagnostic of being a dumb person.

The Bayesian approach formalizes this intuition. To begin with, we assume that people have generative models that specify how different behaviors (good, bad, smart, and dumb) are caused by different types of people (Fig. 1a). These are specified probabilistically (Fig. 1b). For instance, smart people often do both smart and dumb things, but dumb people rarely do smart things. Meanwhile, mean people often do nice and mean things, but nice people rarely do mean things. Given these causal models, it follows from the logic of Bayesian inversion that smart acts are more "diagnostic" than dumb ones, and mean acts are more diagnostic than nice ones (Fig. 1c). This is because observing an act only tells us what type of person somebody is if that act is very often performed by one type of person, and very rarely by another.

Although the application of Bayes' rule to this problem allows us to be "quantitatively precise"—that is, to assign numerical probabilities to things—this is not its main selling point. Nobody has a precise measure of the relevant inputs (what is the exact prior probability of a person being a "sinner" or a "saint"?), and nobody has a use for precise outputs anyway (what hangs on the question of whether p(sinner | steal) = .84 or .92?).

Rather, the strength of the Bayesian approach is that it lays bare the mechanics of representation and reasoning. Representations are probabilistic and causal; reasoning involves taking observed things (theft, or charity) and then systematically asking which unobserved properties could best explain those (saintliness, or sin). It explains how we update old beliefs in light of new information; how predication relates to explanation; how sparse observations can sometimes license strong inferences.

Most of all, the Bayesian program fulfills Lewin's vision of a general theory—one that unifies a large body of facts accumulated across diverse subdisciplines: cognitive



(c) Bayesian Inversion for inference

$$P(type \mid act) = \frac{P(act \mid type)P(type)}{p(act)} = \frac{P(act \mid type)P(type)}{\sum_{type} p(act \mid type)p(type)}$$

$$P(\text{sinner} \mid \text{steal}) = \frac{.6 \times .5}{(.6 \times .5) + (.1 \times .5)} = 0.86 \qquad P(\text{saint} \mid \text{give}) = \frac{.6 \times .5}{(.6 \times .5) + (.1 \times .5)} = 0.69$$

$$P(\text{saint} \mid \text{steal}) = \frac{.1 \times .5}{(.6 \times .5) + (.1 \times .5)} = 0.14 \qquad P(\text{sinner} \mid \text{give}) = \frac{.1 \times .5}{(.6 \times .5) + (.1 \times .5)} = 0.31$$

Fig. 1. An example of inference by Bayes' rule. Given that a person steals, one can draw a relatively strong inference that he is a thief, because only thieves steal with appreciable probability. Given that a person gives to charity, one cannot conclude as strongly that he is not a thief, because even thieves occasionally give to charity. (a) The claims about how thieves (and non-thieves, "saints") behave are encoded in a generative model. (b) These generative models can be specified probabilistically. (c) In this case, Bayes' rule supports an inference to the best explanation for the data, taking into account the prior probability of those explanations as well as the likelihood of the data conditional upon them.

psychology, social psychology, developmental psychology, and clinical psychology. The Bayesian approach is fundamental to current models of how we see (Knill & Richards, 1996), how we remember (Hemmer & Steyvers, 2009), how we communicate (Goodman & Frank, 2016), how we infer others' thoughts (Baker, Saxe, & Tenenbaum, 2009), how we predict (Vul, Goodman, Griffiths, & Tenenbaum, 2014), and much more (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). As diverse as these kinds of thinking are, Bayesian methods show how each of them shares some structure in common with the others.

Bayesian models have also been especially transformative in our understanding of social learning (Acemoglu, Dahleh, Lobel, & Ozdaglar, 2011; Griffiths & Kalish, 2007; Perreault, Moya, & Boyd, 2012; Shafto, Goodman, & Griffiths, 2014). In this issue, Vélez and Gweon investigate the case of social learning from imperfect teachers. In their experiments, the participants are playing a simple gambling game involving picking among several cards. Both the participant and a "teacher" have imperfect information about the value of the cards, and the teacher gives the participant advice. Consistent with some prior research (Biele, Rieskamp, & Gonzalez, 2009; Toelch, Bach, & Dolan, 2013), Velez and Gweon find that they can model participant behavior using a simple heuristic assignment of "accuracy" to the teacher's recommendations (i.e., by summarizing the proportion of the time that the teacher makes a "good" recommendation). Moving beyond this prior research, however, they show that a superior fit to the data can be obtained by modeling the participant's joint inference over the value of the cards and the teacher's beliefs about the value of the cards. Contemporary Bayesian methods allow the authors to formalize this model in a clear way that integrates seamlessly with a larger body of research on learning and mentalizing.

Several of these core themes are echoed in another contribution to this issue by Yang, Vong Vu, and Shafto. They leverage formal tools to draw out an important and surprisingly overlooked connection between teaching and active learning. In episodes of teaching, a knowledgeable individual structures information and experiences for a naive learner in order to give her new, true beliefs. In order to do this effectively, the teacher may use theory of mind to reason about what the learner currently believes and how different kinds of information and experience are likely to influence those beliefs. Prior work, much of it by the same authors, casts this problem as a form of recursive Bayesian mental state inference ("I infer that if I do X, then you will infer Y"). In their contribution to this issue, Yang and colleagues show that the same framework can be adapted to model "active learning": the process by which a naive individual structures her *own* experiences in order to attempt to gain new, true beliefs. Their use of formal methods allows them to illustrate deep connections between teaching and active learning, and to identify the ways in which this novel account of active learning differs from prior approaches.

Also in this issue, Ong and colleagues review a spate of recent work integrating emotion into current Bayesian models of mentalizing. Early attempts to frame theory of mind in Bayesian terms focused specifically on the theory of rational action (e.g., Baker et al., 2009). According to this model, a person acts to efficiently achieve a set of desires conditional on his beliefs about the environment. The basic function of Bayesian inference in this domain, then, is to identify the "hidden" beliefs and desires that best explain the observed actions that a person takes. Now, there is no doubt that many human decisions are governed by this kind of reasoning; we review some of the relevant evidence in the next section. But there is equally little doubt that humans sometimes make decisions in other ways; for instance, by acting out of habit, instinct, or emotion. Similarly, humans' mental states are obviously not exhausted by beliefs and desires, but also include various emotional experiences and reactions. It would be remarkable, then, if people's intuitive theories of mind were blind to these kinds of thoughts; that unlikely possibility is undermined immediately by the presence of the words "habit," "instinct," and "emotion" in the lay vocabulary. Ong and colleagues summarize and systematize a variety of proposals that enrich the spare principle of rational action with a variety of representations and causal relations that capture the place of emotions in folk theory of mind. They then show

how the tools of Bayesian inference could allow people to infer and reason about the emotional states of others, even though these cannot be directly observed.

These contributions to the issue are notable because they integrate the methods of Bayesian inference with the capacity for decision making by humans. And, indeed, there is a large and productive literature on Bayesian decision theory. Nevertheless, a distinct lineage of formal tools has been relatively more central in psychological theories of value-guided decision making, and we consider these next.

2.2. Choice: Value-guided decision making

The most basic formal model of decision making is expected utility theory. Described by Bernoulli in 1738, it states that a person should choose an action by considering the value of the outcomes it might produce, weighted by the likelihood of each of those actions. Bernoulli understood this as a normative model of decision making: one that explains how we ought to decide. Construed instead as a descriptive model of decision making, however, it may be the first true formal model in the history of psychological theory.

It may also be the most maligned. In the second half of the 20th century, scholars in social psychology, anthropology, cognitive psychology, and economics began treating expected utility theory as a straw man on a shooting range, knocking off its axiomatic commitments in experiment after experiment—initially with surprise, later with ease, and finally with a touch of sadism.

During that very same period, however, the key elements of expected utility theory were rearranged into several families of new formal tools that have revolutionized our understanding of decision making, with far-reaching consequences across economics, psychology, neuroscience, and computer science. In short, the basic formal tools of expected utility, rational choice, and reinforcement learning may be descriptively inaccurate, but they are also conceptually indispensable.

2.2.1. Utility

The first key legacy of expected utility theory is, simply, that humans represent something like "utility": A subjective assignment of value (or disvalue) to certain actions, states, objects, etc., that guides choice. (In reinforcement learning, this is called "reward," although it could be positive or negative.) This did not have to be so. It is perfectly possible to produce organized, adaptive behavior without representing utility; your laptop computer does it all the time. Similarly, many early theories in neuroscience were organized roughly as "reflexes" (reviewed in Glimcher, 2004): direct mappings from stimulus to action via a neural Rube Goldberg machine. There is no doubt that some aspects of these theories are correct, but we now know they are incomplete: Throughout the brain, neurons explicitly represent utility, and many of our choices are guided by these representations.

2.2.2. Expected utility

It is useful for people to represent utility because it helps them make decisions. The formal concept of expected utility offers one vision of how, precisely, utility and choice ought to be related. A person should begin by enumerating every possible action and then, for each action, enumerate its possible outcomes, their utilities, and their probabilities of occurring. Finally, they should choose the action with the highest expected utility —that is, the sum, for all possible outcomes, of the product of their probability and their utility.

People do not do this, exactly. People overweight unlikely events in decision making; they encode value against certain "reference points"; they respond to losses and gains asymmetrically. Theories that seek to explain these deviations, most famously Prospect Theory (Tversky & Kahneman, 1992), are often best described as psychologically motivated modifications of expected utility theory. In other words, the concept of expected utility is not useful because it predicts choice with perfect precision, but rather because it helps us to see a crucial dimension of the problem that actual choice algorithms are designed to solve, and because it shows us the abstract simplicity of its optimal solution (Marr, 1982).

Economists and psychologists have long known that the maximization of expected utility helps to organize theories of human choice. One of the more remarkable discoveries of the past 30 years, however, is how elegantly it organizes theories of neural representation (Rangel, Camerer, & Montague, 2008; Schultz, 2006). Across the brain, neural populations encode variables corresponding to the rewards associated with events, their probabilities of occurrence and, ultimately, their expected utility. These representations have been repeatedly demonstrated to structure choice behaviors ranging from where to look (Platt & Glimcher, 1999) to what to buy (Knutson, 2007).

2.2.3. Reinforcement learning

The computation of expected utility asks us to evaluate actions according to their likely outcomes and associated utilities. Sometimes, however, this is very hard to do. Many human goals require extended and complex sequences of actions: earning a bachelor's degree; visiting the Louvre; catching a salmon. If you are contemplating opening moves for a game of chess, for instance, it is not especially illuminating to be advised: "Just choose whichever move has the highest expected utility." You knew that; the problem is to discover what the expected utility is and to coordinate across the many sequential actions you will have to take.

Reinforcement learning methods aim to solve the problem of learning or estimating values in order to make adaptive sequences of decisions. Chess, for instance, can be formalized as a Markov decision process (MDP) in which each arrangement of the board is a "state," the movements of pieces are "actions," and checkmate is the "reward." The goal of reinforcement learning is to assign expected values to things other than checkmate, in order to simplify decision making. For instance, you might assign value to certain advantageous states of the board, or to certain advantageous actions.

A set of formal tools developed in the late 1950s (e.g., Bellman, 1957) helped researchers to understand what, in principle, it means to make such value assignments, and how they could be derived under ideal conditions. Later, in the 1980s, researchers in the computer sciences began to develop efficient ways to approximate such values

(Sutton & Barto, 2018). This line of research eventually led to machine learning algorithms exceeding expert human play in games such as backgammon (Tesauro, 1995), Atari video games (Mnih et al., 2015), and, most recently, go (Silver et al., 2016, 2017). Indeed, it is integral to the current state-of-the-art solutions to many diverse machine learning problems. The same formal tools are also central to our understanding of human decision making. Several computational hallmarks of reinforcement learning have been identified in the dopamine reward system of humans and non-human animals (e.g., Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997).

One especially useful idea formalized by reinforcement learning (RL) captures the essence of "dual process" models of decision making (Dolan & Dayan, 2013). At least since Thorndike (1927)—arguably, since Plato (*Phaedrus*, sections 246a–254e)—behavioral scientists have understood that humans choose actions by a variety of means, each suited to different circumstances. When acting quickly or thoughtlessly, we tend to simply repeat actions that have served us well in similar past circumstances. Habits are the most extreme example of this basic form. In contrast, when acting slowly and deliberately, we choose actions by considering their likely consequences. Goal-directed planning is the most extreme example of this basic form.

Reinforcement learning models refine our understanding of the essential difference between these processes. "Model-based" RL captures key aspects of goal-directed planning. All model-based methods share the feature that the agent maintains a subjective representation of the relevant task—a "world model." It evaluates actions by assessing their likely outcomes and then chooses among them by maximizing expected value. In contrast, "model-free" RL captures key aspects of habitual action. All model-free methods share the feature that the agent learns and stores summary representations of the values of actions, but without ever representing a world model. One simple way to do this is to reinforce actions proportional to their history of reward. (This approach becomes especially powerful when "reward" refers not just to the immediate rewards but also to the prospect of future reward encoded in the expected value of the next action, as in temporal difference learning methods). Evidence suggests that humans make use of both model-free and model-based methods of value estimation (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), often in competition but also sometimes in productive combination (Kool, Cushman, & Gershman, 2018).

2.2.4. Applications to topics in social psychology

These foundational insights about the structure of value-based decision making are now having a transformative effect in research on classic social–psychological topics (Hackel & Amodio, 2018). These include prejudice and stereotyping (Kurdi, Satcher, Gershman, and Banaji, in prep), morality (Crockett, 2013; Cushman, 2013), norms (Ruff & Fehr, 2014), cognitive dissonance (Izuma et al. 2010; Sharot, De Martino, & Dolan, 2009), attraction and relationships (Walum & Young, 2018), social valuation (Krienen, Tu, & Buckner, 2010), prosociality (Zaki & Mitchell, 2011), trust (Behrens, Hunt, Woolrich, & Rushworth, 2008), conformity (Izuma, Saito & Sadato 2010, Zaki, Schirmer, & Mitchell, 2011), and more.

Several contributions to the present issue compellingly demonstrate how computational models of value-guided decision making can make unique contributions to our understanding of social interactions.

In this issue, Le Mens and colleagues consider the question of whether, all else being equal, people should observe that popular things are better than unpopular things, or instead that popular things are worse than unpopular things. By modeling this question, they uncover a form of "self-fulfilling prophecy." People who tend to believe that popular things are good will explore many popular things, discovering that the best of these are much better than they expected. In contrast, people who tend to avoid popular things will fail to discover those very best-of-the-popular and persist in believing that they were mediocre options. For this reason, even the absence of any "true" correlation between popularity and quality, the propensity to choose popular things (or not) can create an illusory correlation (or anticorrelation).

Also in this issue, Krafft pursues the application of a value-based decision-making framework to a social setting (or what is often called a "multiagent" setting in the computer science tradition). Specifically, he studies the conditions that give rise to "collective intelligence"—the ability of a group of individuals to jointly optimize their individual payoffs across diverse tasks. By formulating this problem as a variety of MDPs (one that is multiagent and partially observable), he shows that a narrow set of conditions guarantees optimal performance: perfect alignment of value and beliefs, and perfectly coordinated action. In the real world, such conditions are presumably rarely met, however, and so their main contribution is to define a notion of "general collective intelligence," akin to IQ but for groups. Generally, groups will have higher collective intelligence to the extent that they are aligned in their beliefs and their preferences, and when they are also coordinated in their actions.

2.3. Strategic interaction: Game theory

Not surprisingly, value-based theories of choice do a good job of explaining how people value each other, and how they value different actions or outcomes in a social setting. Reinforcement learning methods, in particular, have also been fruitfully applied to develop good policies for strategic play in games such as backgammon, chess, and go. They are not especially well suited, however, to the *efficient* description of such strategic interactions—the outcome of a negotiation, the bargain over a price, the conduct of war, or the commitments of a romantic relationship.

Game theory is a formal framework for understanding the interaction of multiple agents motivated by rewards that depend on each other's behavior—for instance, people playing games. Although the foundations of game theory are historically deep, its most pivotal period of development began in the late 1940s and 1950s, around the dawn of social psychology's golden era.

Game theory formalizes interactions between agents, each of whom has one or more actions available to it. A set of rewards—typically called "payoffs" in game theory—are at least partially dependent on the actions of the various players. The players are assumed to be motivated to pursue payoffs. The players may act simultaneously or sequentially; once or many times; with perfect or imperfect information about their environment; rationally or with variable degrees of irrationality. The framework has been productively used to ask questions about how humans ought to make decisions ("What should a rational actor do?"), to predict how humans do make decisions individually or through bargaining ("What are people likely to do?"), and also to model evolutionary dynamics ("What are organisms likely to evolve to do?").

Consider, for instance, the game of "chicken." In this game, adolescents drive cars directly at each other in order to find out who (the "chicken") will swerve first. If neither driver swerves, both suffer a major loss (death); if either swerves, they suffer a minor loss (honor); if one stays the course while the other swerves, they win. This vividly illustrates the property of games that one player's payoffs can depend on another player's choices.

There is, obviously, no single "right thing to do" in all games of chicken. Rather, the right thing to do depends on what your partner will do. If you know that she will drive on, then swerve; if you know she will serve, then drive on. And, as game theory shows, for every game of chicken there is some pair of "stochastic" policies (i.e., "drive straight with probability p, swerve with probability 1-p"), such that neither player can unilaterally improve upon flipping a p-weighted coin. A related result in evolutionary biology shows that games of chicken can induce a form of "balancing selection" in which a population of players comprises proportion p of "straight" genes and proportion 1-p of "swerve" genes, and any drift in gene frequencies away from these proportions is rebalanced by natural selection.

As this example suggests, one of the key features of game theory is that it allows us to reason about the kinds of "equilibria" that can be achieved among a set of strategies adopted by different players. For example, the celebrated Nash equilibrium identifies sets of strategies for which no individual player can improve their payoff by unilaterally changing their strategy. While all players might be better off if they collectively and simultaneously changed strategies, this requirement of collective action may pose a practical barrier to strategy change. In some contexts, then, we may expect to find individuals persistently settled into one Nash equilibrium, even when every one of them would prefer another. The formal tools of game theory are certainly not necessary to help predict these kinds of situations (among many others), but they make it especially easy to identify and reason about them.

For this reason, game theory has been used to analyze a staggering array of human interactions. For instance, influential theories of human communication (e.g., the meanings of words) depend on the analysis of "coordination" games in which two players are striving to arrive at a mutually convenient set of policies ("what I mean by "BAGEL" is the same as what you mean by "BAGEL"). An extension of these ideas explains, for instance, why people bother with "indirect speech" ("officer, it's a shame there isn't some other way we could settle this ticket…"; Pinker, Nowak, & Lee, 2008).

Another major branch of game theory analyzes "social dilemmas," such as the Prisoner's Dilemma or Public Goods Game. These capture situations in which self-interest conflicts with the public interest. A major area of research aims to understand how cooperation is or could be achieved in such cases (Nowak, 2006). Potential solutions include contingent reward or punishment ("reciprocity"), reputational consequences ("indirect reciprocity"), intergroup competition ("group selection"), and family relationships ("kinship").

Game theory has also been used to model commitment, and the ways in which we attempt to signal it. A classic example, arising from the possibility of strategic nuclear war, is the doctrine of mutually assured destruction. In order to deter a nuclear first strike, it can be rational for a nation to "precommit" itself to a retaliatory strike. The retaliatory strike is apparently paradoxical; by the time a nation launches it, it is already assured of annihilation. The commitment to a massive retaliatory strike is advantageous not in the event that it occurs, but rather because of the precommitment—it is designed to credibly and successfully deter any preemptive strike in the first place. This logic has been applied to model the structure of the human revenge motive (Frank, 1988).

It is obvious, however, that "precommitments" of this form can only be effective if they are clearly and credibly signaled. In other words, you will not deter your enemy from a preemptive strike if they are unaware that you have precommitted to retaliation, or if they do not believe you. This insight has led to a large and productive literature on honest signaling.

For instance, one recent analysis explores the ways in which people can signal commitment by deliberately restricting their own access to information (Hoffman, Yoeli, & Nowak, 2015; Jordan, Hoffman, Nowak, & Rand, 2016). Suppose, for instance, that a friend asks whether you could help him move apartments. If you ask, "How much stuff do you have?", this indicates to your friend that you lack a strong commitment to helping him; if you offer assistance before knowing how much stuff he has, this signals a stronger commitment. Similar logic has been proposed to explain why romantic partners are strongly committed to each other—in other words, disinterested even in non-partner mates that, from some immediate standpoint of evolutionary fitness, are superior (Frank, 1988; Hoffman et al., 2015).

In this issue, Bear and Rand take that suggestion as their point of departure. Their work illustrates an important consequence of formal argument in the cognitive sciences. They verify that the logic of the original suggestion holds for a certain parameter space. Specifically, if the costs of being left by your partner are very high (i.e., it would be very hard for you to find a new partner at all, or to find one nearly as good), then the optimal strategy is to avoid considering new partners (even at the cost of finding better ones), and to leave your current partner if you sense that they are considering alternatives themselves. But the parameter space in which this result obtains is rather small. Beyond those narrow conditions, it instead makes sense to continuously evaluate your partner options and to tolerate similar behavior by your partner. This suggests that many cases of "blind commitment" to romantic partners may require a different kind of explanation.

3. Conclusion

Social psychology matured together and yet strangely apart from the computational methods we have described, which together comprise a foundation for computational social science. These fields grew up at the same time, in many of the same places, engaged with a similar set of questions about the world. Yet, like twins adopted into different families, they grew up in isolation from each other. Did this isolation give social psychology the room and freedom that it needed to grow, unfettered by formalities? Or, was it instead deprived of a natural ally and companion?

We know how Lewin, Heider, and Festinger would have felt: as sorrowful as a parent who learns that their children never met. But they are also old and gone; today, the field is younger, and perhaps wiser. We conclude by considering some of these younger voices: contributors to this issue who, in their own way, deliver a call to arms by and for the current generation of social psychologists.

Yu, Siegel, and Crockett offer a compelling case study of how computational methods can enrich social psychology. They review a program of research that integrates two topics: how people make moral decisions, and how people perceive others' moral decisions. Obviously, these topics are related: People make moral decisions by weighing self-interest against their concern for others; observers take note of this fact and attempt to infer just how much a person's motives are selfish versus selfless. In the absence of formal tools, however, it has been challenging to move beyond such general statements. How, precisely, are decisions made; how are they perceived and inferred; and, to what extent is inference biased or noisy, and to what extent does it accurately capture "ground truth"? Focusing on an experimental paradigm in which participants allocate electrical shocks to others at a self-profit (or make judgments of others doing so), Yu and colleagues (a) discuss a utility-based formal model of decision making, (b) specify a Bayesian inference procedure that can recover key parameters of this model, allowing observers to model the motivations of actors, and (c) describe a process by which such inferred variables can generate moral judgments of the actors.

This case study draws out one of the key virtues of computational methods: They attain simplicity through abstraction. A concept is abstract when it generalizes across the peculiar features of many individual cases. It is a successful abstraction when the generalization retains the core, essential features that make those cases importantly similar. Part of the particular value of formal tools is that they naturally afford useful abstractions of this kind. For instance, the process of making a moral decision and the process of judging another person's moral decision are quite different. (Moreover, each one of these processes is instantiated in very different ways in different contexts.) Yet the formal tools afforded by theories of value-based decision making and Bayesian inference allow us to see certain abstract similarities—a common set of core concepts and conceptual relations that influence how we act, how we judge, and whom we trust. In order to understand the formal models, we must learn new concepts ("value," "generativity," "Bayesian inversion"), often new symbols, and certainly new ways of thinking. But the return on that investment is, in fact, to achieve a greater simplicity of thought: We may clearly apprehend the form of a spare abstraction, uncluttered by the extraneous details of any particular case.

Jolly and Chang's contribution to this issue provides, however, a wholly different perspective on the value of formal methods in the social sciences. They identify value not in reducing complex phenomena to simple theories, but rather in building theories sufficiently complex to mirror the intended phenomena. They borrow a metaphor from Abbot's (1884) novella *Flatland*, which describes people who interpret a three-dimensional world through a two-dimensional conceptual framework. The central conceit of Abbot's fiction should, they argue, feel uncomfortably close to home to psychologists. Many prominent theories are organized around dichotomies (for instance, "automatic" vs. "controlled") that surely obscure a much more complex set of facts. These facts, they argue, require more "dimensions" than just two in order to be properly understood. The difficulty with high-dimensional theories, however, is that the domain-general mental systems we generally use for scientific inference and communication are not capable of representing or reasoning about them. In short, the world is too big, and our brains are too small. Here, Jolly and Chang identify a crucial virtue of computational approaches: Formal tools such as mathematics allow us to both represent and reason about complex theories symbolically, on a piece of paper or the screen of a computer. They augment low-dimensional minds in a way adequate to the high-dimensional theories that science often requires.

We seem, then, to have two diametrically opposed visions of the utility of formal theories in psychological research. According to one vision, formal methods are useful because they facilitate abstractions of great simplicity; according to another vision, they are useful because they facilitate a high-dimensional representation of great complexity.

In fact, however, these perspectives are remarkably aligned in their essence. First, they agree that progress in psychological research will require new concepts that go beyond our prescientific "folk psychology," and new conceptual relations that go beyond mere descriptions of facts. Formal methods give us a way of constructing new concepts and describing new relations among them. This is equally true whether the relations are abstracted and simple, or high-dimensional and complex. Second, they agree that a goal of psychological research should be to develop theories that draw disparate phenomena under common frameworks. Formal methods provide a common, compositional language in which disparate data can be aggregated into a larger theory, and disparate theories into a larger framework. Indeed, this issue provides many excellent examples in which quite different formal methods are integrated—for instance, combining theories of value-based decision making and Bayesian inference to develop a more comprehensive theory of morality.

Of course, these virtues align elegantly with the animating vision behind social psychology: For social psychology to move beyond the "piling up of facts" by developing new conceptual tools, and for these tools to systematize and unify theories across every corner of psychological research—to explain, finally, "whole persons" and their interactions. Formal tools provide a lingua franca in which we may state, finally, the hidden logic of familiar things.

References

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, *78*(4), 1201–1236.

Allport, G. W. (1961). Pattern and growth in personality. Oxford, England: Holt, Reinhart & Winston.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113* (3), 329–349.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245.

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press. Republished 2003: Dover.

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science*, *33*(2), 206–242.

Birnbaum, M. H. (1972). Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, *93*(1), 35.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken, NJ: John Wiley & Sons.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Hoboken, NJ: John Wiley & Sons.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York: WW Norton & Co.

Glimcher, P. W. (2004). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Gray, K., & Wegner, D. M. (2013). Six guidelines for interesting research. *Perspectives on Psychological Science*, *8*(5), 549–553.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). New York: Cambridge University Press.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

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

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*(6), 1727–1732.

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of the National Academy of Sciences*, *107*(51), 22014–22019.

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience*, *22*(4), 621–631.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127.

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience*, *30*(41), 13906–13915.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. In R. Morris, A. Bornstein & A. Shenhav (Eds.), *Goal-directed decision making* (pp. 153–178). Cambridge, MA: Academic Press.

Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007). Neural predictors of purchases. *Neuron*, *53*(1), 147–156.

Lewin, K. (1936). *Principles of topological psychology* (F. Heider & G. M. Heider, Trans.). New York: McGraw-Hill.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Martijn, C., Spears, R., Van der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology*, *22*(5), 453–463.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563.

Perreault, C., Moya, C., & Boyd, R. (2012). A Bayesian approach to the evolution of social learning. *Evolution and Human Behavior*, *33*(5), 449–459.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400* (6741), 233.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549.

Sharot, T., De Martino, B., & Dolan, R. J. (2009). How choice reveals and shapes expected hedonic outcome. *Journal of Neuroscience*, *29*(12), 3760–3765.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275* (5306), 1593–1599.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, *57*, 87–115.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, *38*(3), 58–68.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212–222.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Toelch, U., Bach, D. R., & Dolan, R. J. (2013). The neural underpinnings of an optimal exploitation of social information under uncertainty. *Social Cognitive and Affective Neuroscience*, *9*(11), 1746–1753.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Walum, H., & Young, L. J. (2018). The neural mechanisms and circuitry of the pair bond. *Nature Reviews Neuroscience*, *19*, 643–654.

Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences*, *108*(49), 19761–19766.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science*, *22*(7), 894–900.

Papers in this topic

Bear, A., & Rand, D. (2019). Can strategic ignorance explain the evolution of love? *Topics in Cognitive Science*, *11*(2), 393–408.

Jolly, E., & Chang, L. (2019). The flatland fallacy: Moving beyond low-dimensional thinking. *Topics in Cognitive Science*, *11*(2), 433–454.

Krafft, P. (2019). A simple computational theory of general collective intelligence. *Topics in Cognitive Science,*, *11*(2), 374–392.

Le Mens, G., Denrell, J., Kovacs, B., & Karaman, H. d. (2019). Information sampling, judgment and the environment: Application to the effect of popularity on evaluations. *Topics in Cognitive Science*, *11*(2), 358–373.

Ong, D., Zaki, J., & Goodman, N. (2019). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in Cognitive Science*, *11*(2), 338–357.

Vélez, N., & Gweon, H. (2019). Putting two heads together: Integrating incomplete information with imperfect advice. *Topics in Cognitive Science*, *11*(2), 299–315.

Yang, S., Vong, W. K., Yu, Y., & Shafto, P. (2019). A unifying computational framework for teaching and active learning. *Topics in Cognitive Science*, *11*(2), 316–337.

Yu, H., Siegel, J. Z., & Crockett, M. J. (2019). Modeling morality in 3-D: Decision-making, judgment and inference. *Topics in Cognitive Science*, *11*(2), 409–432.