Trends in Cognitive Sciences

CellPress

Review

Planning with Theory of Mind

Mark K. Ho, [1,4,*,@] Rebecca Saxe, [2] and Fiery Cushman [3]

Understanding Theory of Mind should begin with an analysis of the problems it solves. The traditional answer is that Theory of Mind is used for predicting others' thoughts and actions. However, the same Theory of Mind is also used for planning to change others' thoughts and actions. Planning requires that Theory of Mind consists of abstract structured causal representations and supports efficient search and selection from innumerable possible actions. Theory of Mind contrasts with less cognitively demanding alternatives: statistical predictive models of other people's actions, or model-free reinforcement of actions by their effects on other people. Theory of Mind is likely used to plan novel interventions and predict their effects, for example, in pedagogy, emotion regulation, and impression management.

Highlights

Theory of Mind research has traditionally emphasized its predictive function (e.g., predicting someone will be angry after being stuck in traffic). Prediction tasks have dominated decades of experimental and computational research.

Theory of Mind is also used to plan interventions on other minds (e.g., choosing how to cheer someone up who has been stuck in traffic) and representations used for planning will have different requirements from those only used for prediction.

Research on planning emphasizes the importance of abstract and structured causal models, like Theory of Mind.

Focusing on Theory of Mind for planning can illuminate a range of socio-cognitive phenomena, such as interpersonal affect regulation, impression management, pragmatic speech, and pedagogy.

What is Theory of Mind for?

Why have a '**Theory of Mind**' (see Glossary)? Curiously, ever since the concept was first proposed, Theory of Mind has been principally understood as a way to predict people's actions, by inferring their perceptions, beliefs, and desires. This predictive function motivates nearly every classic paper on the topic: '... the system can be used to make predictions, specifically about the behavior ...' [1]; 'The ability to make inferences about what other people believe to be the case in a given situation allows one to predict what they will do' [2]; 'one treats the system whose behavior is to be predicted as a rational agent ... and then predicts that it will act to further its goals in the light of its belief' [3], and so on.

But, this is like saying that a theory of cooking is useful mostly for guessing what is coming out of the kitchen. Theories are useful not only for passively predicting the future, but also for actively **planning** to change it. Just as we can plan and cook a meal, we can use Theory of Mind to plan to change people's actions and feelings, by intervening on their perceptions, beliefs, and desires.

Here we consider the implications of planning for why people have Theory of Mind and how it is structured. First, Theory of Mind is a single unified **causal model** that can be used to solve multiple problems, including both prediction and planning (Figure 1) [4]. In this respect, Theory of Mind contrasts with non-causal **predictive models** (which are not suited for action selection) and model-free methods of action selection (which are not suited for prediction). Second, the task of planning places important architectural constraints on Theory of Mind, requiring **abstract representation** and **structured representation**, and strategies for managing search and action selection. Finally, we consider current progress in three specific research topics, to identify common themes and future directions for understanding the representations and computations that underlie planning with Theory of Mind.

Predictions, plans, and habits

Predicting others' actions and reactions

The classic problems (and psychological tasks) used to study Theory of Mind require an observer to predict or explain another person's actions [2,5,6]. For example, consider Harold, who typically

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA
[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[3]Department of Psychology, Harvard University, Cambridge, MA, USA
[4]Department of Psychology, Princeton University, Princeton, NJ, USA

*Correspondence: mho@princeton.edu (M.K. Ho).
@Twitter: @Mark_Ho_

orders lunch from a Lebanese food truck [7]. Earlier this morning, Harold saw the food truck in the north lot, but since then it has moved to the south lot. Where will Harold go at lunch time? In this variant of the standard false belief task, Harold's friend Grace can predict that Harold will go to the north lot (and will be disappointed when he gets there [8]).

Using Theory of Mind, Grace could predict Harold's actions by invoking how Harold's perception and inference (he saw the Lebanese truck in the north lot, he infers it has not moved) cause his belief (the Lebanese truck is in the north lot), which combines with his desire (to get Lebanese food for lunch) to create a plan (go to the north lot).

Yet, a causal model is not necessary to predict Harold's action in this case. Good predictions of agents' actions in future situations can be derived from non-causal, statistical predictive models trained with extensive observations of actions in similar situations [9–11]. In this case, Grace could have learned from prior experience what will happen in situations like this and thus have a predictive model of the sequence of events. For example, 'When Harold wants Lebanese food, he goes to the last place he saw the Lebanese truck'. On this basis she could pass the false belief task without representing any causal relationships (Figure 1). Predictive models of this kind can also help us reason about perception, emotions, and so on.

The family of statistical predictive models is very diverse. As we use the term, their essential feature is that they encode how variables are correlated or temporally sequenced, rather than causal relations. The models may also be defined more in terms of observable states ('smiling') than unobservable ones ('happy') and they may make less use of structure and abstraction (i.e., failing to explicitly represent 'wanting ice cream' and 'wanting a cozy warm fire' as instances of the same type), but these features are more variable. Any system with the opposites of all of these features, one that encodes causal relations between abstract, structured, and often unobserved mental states, we would consider a model of 'Theory of Mind'.

There has been considerable debate about which of these types of models humans use and when. The prediction problems that demand flexible generalization to novel situations are most likely to benefit from a causal model. For example, previous experience with Harold might be sufficient for Grace to predict his reactions if he gets to the north lot and finds no Lebanese truck, but not if he finds a gorilla or his grandmother. Thus, human observers may use predictive models to anticipate previously encountered sequences, but switch to using Theory of Mind to predict how others will react in an infinite array of novel situations.

Here, however, we draw attention to another human capacity that requires Theory of Mind: planning to change others' actions, thoughts, and feelings.

Planning to evoke actions and reactions

In addition to predicting, people deliberately act to evoke desired actions and reactions in other people and to avoid undesired ones. For example, what if instead of predicting Harold's action, Grace wants to make Harold happy, by causing him to find the Lebanese truck at the south lot? For this purpose, a non-causal predictive model will not suffice. Grace does not need to know what predicts Harold going to the south lot (e.g., he's hungry, he saw the Lebanese truck in the south lot), but what actions she should select to make him go to the south lot.

One way to select actions is by planning. Planning involves considering the effects of different possible actions, in order to select the action that maximizes expected value [12–15]. To accomplish this, planning requires a causal model that specifies the asymmetric dependence between

Glossary

Abstract representation: a mental representation that categorizes and generalizes across instances with diverse particular features.
Causal model: a model of the structure of relations between variables that supports counterfactual and hypothetical reasoning (i.e., one that specifies the probability of one variable setting conditioned upon intervention on another variable).
Habit: actions chosen based on learned action–reward associations. Habits do not represent, and cannot be used to predict, the outcomes of actions.
Model-free reinforcement learning: learning the value of an action, in context, from repeated association of the action with subsequent reward.
Planning: a mechanism for action selection that relies on using a model to simulate the future consequences of potential interventions in order to choose one that maximizes expected value.
Predictive model: a model of statistical correlations or sequences of events that can be queried to generate an expectation for one variable or state, given an observation of another variable.
Prioritized memory: mechanisms and strategies for organizing memory that facilitate rapid retrieval of context- and goal-appropriate action representations in large state spaces.
Structured representation: a mental representation that is the composition of simpler representations and whose content derives from these constituents and how they are combined.
Theory of Mind: a causal model of the mind, specifying how mental states like perceptions, beliefs, and desires combine to cause actions and feelings.
Value-guided construal: top-down recombination of causal elements to construct a simplified causal model that is useful for planning in context.

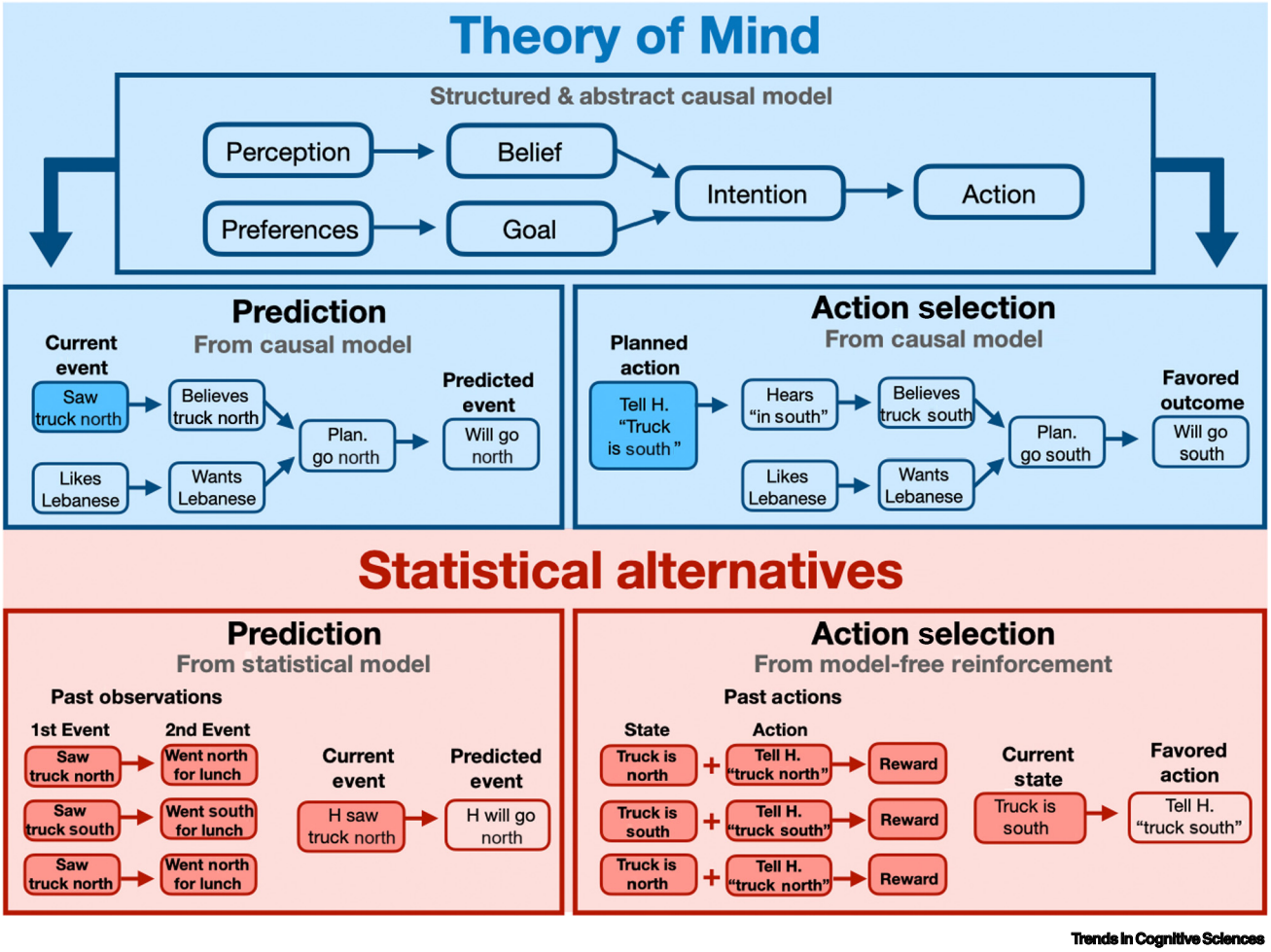

Figure 1. Contrasting Theory of Mind with statistical alternatives. Top: Theory of Mind is an abstract causal model, specifying how mental states like perceptions, beliefs, and desires combine to cause actions and feelings and so can be used for both prediction (e.g., given the target's inferred beliefs and desires, predicting their actions) and action selection via planning (e.g., given a desired action, selecting the best intervention on beliefs and desires). Here we illustrate how Theory of Mind could be used (left) to predict a person's actions, if they have a false belief (here, that a food truck is in the north), and (right) to intervene to cause the true belief (the truck is in the south). Below: statistical models generalize prediction and action selection from prior experience of similar states or sequences, without building a causal model. A non-causal predictive model cannot be used for action selection and a model-free action selection mechanism cannot be used for predicting events.

causes and their effects and so reveals what to expect from one variable, following a potential intervention on another variable. By contrast, non-causal predictive models reveal what to expect from one variable, given an observation of another variable [16].

Although both can be used for prediction, causal and predictive models are, as we use the terms here, conceptually and practically distinct. Even very strong predictive relationships may not be useful for planning, because an observation that predicts a variable might be spuriously associated (as the alarm clock predicts the newspaper delivery), or be an effect of the predicted variable (as smoke predicts fire).

The same is true for mental states and actions. For instance, suppose that Harold only uses the south door when the Lebanese truck is in the south lot. In this case, his using the south door predicts his going to the south lot. But, if Harold believes the Lebanese truck is in the north lot,

Grace cannot successfully intervene on his belief by urging him to use the south door. (As she knows, he would simply use the south door and then turn north again, his beliefs uncorrected). Instead, Grace can use her causal model to identify the belief and desire that are the core elements of his plan and consider interventions on these variables. For instance, she might plan to tell him to look out the south window, from which the south lot, and the Lebanese truck in it, are plainly visible. In this way, planners use a causal model to identify useful interventions and screen off spurious (yet strongly predictive) associations (Figure 2A).

Conversely, weak predictive relationships can nevertheless be crucial for planning interventions. Suppose Grace needs to make Harold find the Lebanese truck without tipping off their officemate Luke to its location. Grace might use a white lie, 'Harold, I saw that there are free t-shirts in the south lot'. The logic of her plan relies on Grace knowing that while both Harold and Luke will acquire a new belief, only Harold has the desire (to get a free t-shirt) that will combine with the belief to create the plan Grace wants to induce (to go to the south lot). Ordinarily, one does not make a person believe that a Lebanese truck is in the south lot by telling them that t-shirts are in the south lot. A non-causal predictive model would be unlikely to generate this sequence, assuming it has never been encountered before. Planning with Theory of Mind can reveal the value of an otherwise unlikely sequence of events (Figure 2B).

Evoking actions and reactions by habit

Planning is one way to select actions, but there are others. Planning is most often contrasted with **habit**. Whereas planning chooses an action based on a causal model of the action's consequences, habit chooses whatever action has tended to work best in the past. Habits guide action without making any prediction about, or using any model of, what will happen afterwards [17]. That is why habitual actions persist even when the actor no longer desires the outcome that the action predictably induces. This strategy is the essence of **model-free reinforcement learning** [12]. When given sufficient training, model-free mechanisms can capture many aspects of habit-learning in humans [14] and can be used to train deep neural networks to perform complex tasks [18].

Habit can explain many actions that people choose to evoke actions and reactions in other people. Habitual actions that influence another person's mind can be learned, reinforced, and used without ever explicitly modeling that person's mind [19]. For example, if Harold smiles every time Grace uses a certain turn of phrase, and she finds his smile rewarding, she might habitually use it more often around him.

Habits often arise from behaviors that were at one point planned, but have become automatized through repeated use [20,21]. Domain expertise consists substantially in automatized habits that are finely tuned to context [22]. If Grace needs to direct Harold to the Lebanese truck everyday, she can develop a habit and not spend time making a new plan each time.

So when might people plan using Theory of Mind, versus act by habit? The more the current situation resembles previous situations, in terms of both the environmental constraints and the expected value of possible actions, the more people can rely on habit [23]. In daily social life, habits plausibly underpin a large proportion of actions that influence others. Planning using a causal model is most useful when a novel goal is being pursued for the first time, or when new constraints arise. The problems that require planning with Theory of Mind are thus likely to be memorable and consequential.

Unified causal models versus statistical silos

So far, we have considered two problems and three solutions. The problems are to predict, and to intervene on, other's actions and reactions. A solution to prediction problems is to learn

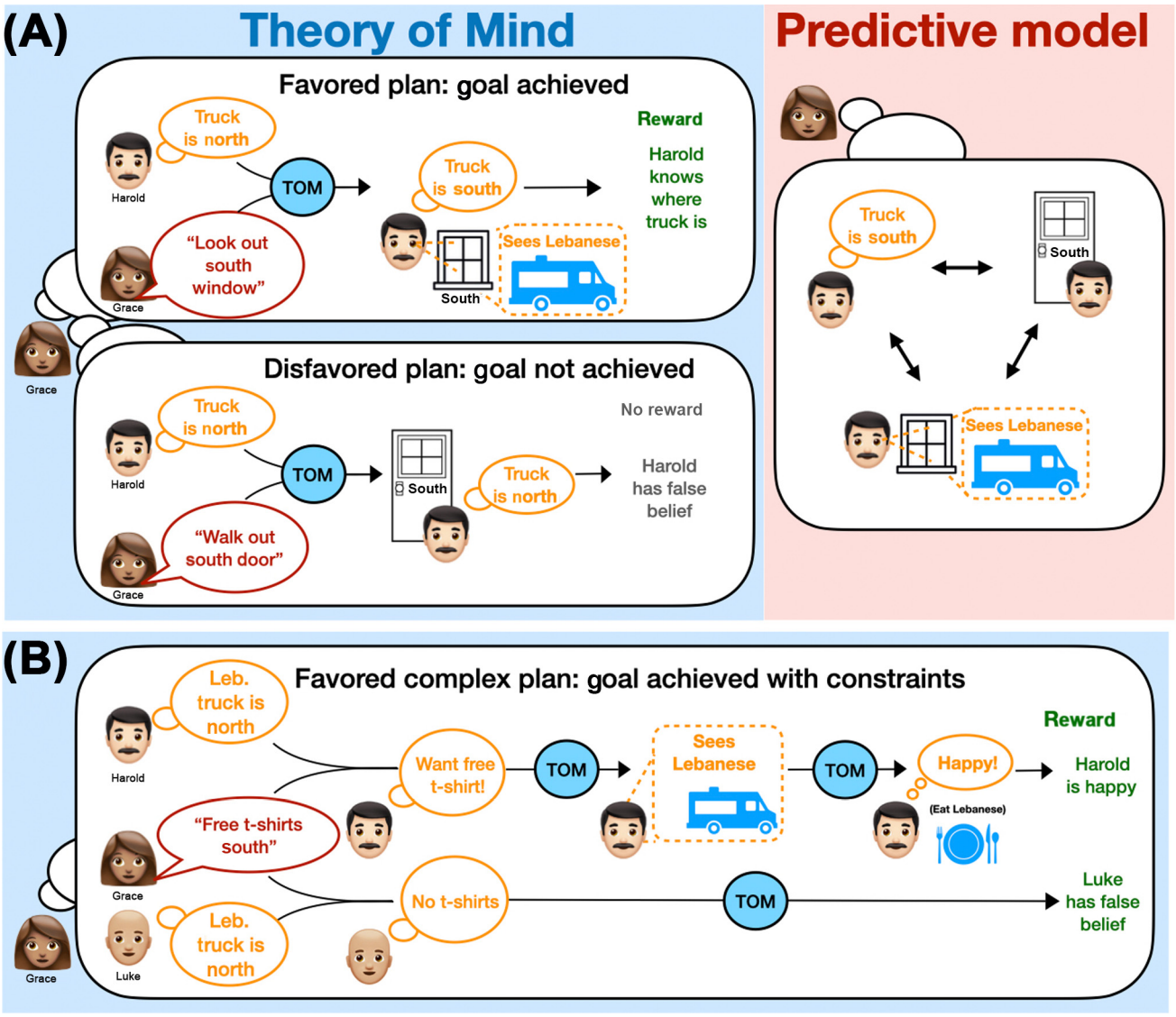

Figure 2. Examples of planning with Theory of Mind. (A) In the first example, Harold wants Lebanese food and falsely believes the Lebanese food truck is in the north lot. Harold's friend Grace wants to make him happy by changing his belief. Left panel: By using Theory of Mind to plan, Grace selects an action to achieve her epistemic goal for Harold by intervening on his perceptions (e.g., seeing the truck through the south window) and not by causing another state that predicts him going south (e.g., using the south door). Right panel: non-causal predictive models of mental states and behaviors do not distinguish between causal relations versus strongly predictive but spurious correlations. (B) In the second example, Grace's plan to inform Harold about the Lebanese truck is constrained by her desire not to tip-off their officemate Luke. To satisfy this constraint, Grace uses a white lie that there are free t-shirts in the south lot, reasoning that Harold, but not Luke, will want a free t-shirt, and so will see the Lebanese food truck. This example illustrates why planning with Theory of Mind is cognitively demanding: Grace's goal is unlikely to happen in the absence of intervention, and requires generating and comparing the expected value of many possible interventions. Abbreviation: TOM, Theory of Mind.

a non-causal predictive model that can capture correlations between events but is not suited to action selection because it does not encode causal relations. Conversely, a solution to intervention problems is to learn good habits, which enable action selection based on learned reward associations, but do not represent the likelihood of subsequent events. In other words, predictive models and habits are 'siloed'; each is well-tuned to its own problem, but cannot solve the other.

The third solution is to learn with and use Theory of Mind: a common causal model, integrating knowledge acquired while solving both prediction and intervention. In other words, Theory of Mind offers a unified solution to prediction and intervention problems, although at a greater computational cost. (For an example of another problem, see Box 1)

We hypothesize that humans use all three solutions. Statistical predictive models and model-free habits are sufficient for many everyday problems. For predicting or planning in novel circumstances, people likely rely on Theory of Mind. Planning, in particular, places architectural constraints on the unified representations composing Theory of Mind, that are not obvious from applications to prediction. We turn to these constraints next.

Architectural constraints imposed by planning

Planning places distinct architectural constraints on Theory of Mind, because planning in general is computationally hard [24,25] and planning in social domains is especially hard [26]. For prediction, one can efficiently determine what is likely to happen by fixing the present observation and sampling forward [16] from a predictive or causal model. By contrast, people typically plan in order to cause events that would otherwise be unlikely or rare [23]. So a planner typically must simulate more extensively in a causal model, to find actions that lead to infrequent but valuable outcomes. Planning also requires additional operations compared with prediction. Once the most likely distribution of future states is identified, the problem of prediction is solved. By contrast, planners must hold candidate plans and their expected values in mind, in order to subsequently evaluate, compare, and select among them.

These computational demands mean that planning typically relies on controlled processing as opposed to automatic processing. When people are making plans, they have slow reaction times, are sensitive to cognitive load, and activate frontal cortical regions, among other signatures of controlled processing [27–31] (for a discussion of the neural basis of planning with Theory of Mind, see Box 2).

Abstraction and structure

Computational research on planning in physical domains suggests that, to be useful for planning, causal models should be abstract and structured representations [32,33]. Abstract representations

Box 1. A third task: moral evaluation

For prediction and planning, solutions can be efficient for one problem, yet ill-suited for the other, while Theory of Mind offers a unified representational substrate that flexibly supports both tasks. The same is true of a third problem: making moral evaluations. People evaluate others' actions and reactions as morally good (or obligatory) or bad (or forbidden). Many such evaluations depend on inferences of the beliefs, desires, and plans that caused those actions or reactions [103,104]. For example, if Grace lies about free t-shirts (see Figure 2 in the main text), Harold's moral evaluation of Grace's action will depend on whether he accepts that her lie was intended to help him find the Lebanese truck, or was intended as a prank or inconvenience. Forgiving an accident, or understanding the positive intention behind a white lie, can be supported by Theory of Mind [105]. Indeed, using the structured, abstract representations in Theory of Mind, people can render complex yet consistent patters of moral judgments for highly unusual or contrived cases [106–108].

At the same time, people can and do make moral judgments using simpler mechanisms that do not depend on a causal model of the mind. Some moral evaluations seem to be supported by model-free value-based associations with actions [109,110] ('It could never be OK to push someone in front of a train'). At other times people fall back on inflexible, precompiled rules [111] ('Driving over 60mph on this road is prohibited'). Action–value associations and inflexible rules afford computationally cheap ways of making moral judgments. Yet these methods of moral evaluation are clearly siloed and are ill-suited to solve the problems of prediction and planning. Knowing that it is against the rules to drive over 65mph does not, by itself, help you predict when somebody will nevertheless do so, or tell you how to best intervene to prevent them.

Theory of Mind is only one solution to the problems of social life, but it offers a unified architecture that integrates representations for re-use in many problems, including predicting, planning, and morally evaluating others' actions and reactions.

Box 2. Neural basis of Theory of Mind

Based on cognitive and computational considerations, we suggest that Theory of Mind offers a unified basis for planning, prediction, and moral evaluation. Evidence that these different problems rely on a common mechanism comes from neuroimaging.

When people read stories, watch movies, or interpret cartoon vignettes, that require understanding characters' minds for the purpose of predicting and explaining actions, they show high activity in temporo-parietal junction (TPJ), medial prefrontal cortex (MPFC), and other regions [123]. Activity in these regions is particularly high when beliefs or desired are needed to predict otherwise unlikely, unusual, or novel actions or reactions [124,125] and when people make moral evaluations based on false beliefs, for example, when exonerating a character for an accidental harm [126]. Indeed, interfering with right TPJ function, using transcranial magnetic stimulation, selectively reduces the use of mental states in moral evaluation [105].

Regions in TPJ and MPFC are also recruited when people plan their own actions by considering the likely thoughts and goals of another person. Example situations include choosing words to describe a partly familiar character [127], picking words in a cooperative game [128], selecting an object from a collective pot [129], or generating deceptive clues in a zero-sum competition [130]. Deception is the most-studied example of choosing actions in order to manipulate others' state of knowledge. Spontaneous, opportunistic, rewarded deception recruits both regions associated with Theory of Mind (like TPJ and MPFC) as well as brain regions associated with effortful selection of response options (e.g., lateral prefrontal cortex), consistent with the idea that planning is cognitively demanding [131].

Nevertheless, the intervention tasks that have been studied typically afford a narrow range of goals and actions, to influence others' minds, and therefore do not demand efficient search and effective pruning. Also understudied is how activity, or connectivity, of these brain regions differs, in a direct comparison of prediction versus intervention tasks (see Outstanding questions in the main text).

are better for planning because they can prioritize just the information relevant for causal relations and successful action selection [34]. The best planning models are not those with the most precise, fine-grained representations. Work in artificial intelligence building video game-playing systems has shown that planning beyond a few steps is infeasible with an overly fine-grained representation of dynamics [35]. For example, successful model-based reinforcement learning algorithms (e.g., the MuZero system [36]) learn causal models in tandem with a planning algorithm. This allows them to learn abstractions that preserve just the relevant information for simulating the consequences of valuable actions [37].

Structured representations combine constituent elements, that partially inherit their causal relationships from their role in an overall theory, into a vast number of new possible actions or thoughts [32,33,38,39]. Structured representations thus allow modular search for interventions, manipulating one constituent element at a time [40]. The constituent pieces of structured representations are more independently controllable, which makes it possible for people to rapidly

Box 3. Planning interventions on one's own mental states

Many of the ideas we discuss can be applied to understanding how people plan interventions on their own mental states. For instance, people deliberately regulate their own emotional states to accomplish goals in the laboratory and in everyday life [112–114]. To intervene on their own emotions, people deploy strategies such as cognitive reappraisal, distraction, or emotional suppression [115]. Emotion regulation can be strategic and context-sensitive: people flexibly upregulate their experience of anger if they believe anger will improve performance on their current task, especially if performance is incentivized [116]. More generally, how people regulate their own emotions is influenced by whether they believe emotions are controllable and how much they value the control [117–119].

Interventions on one's own mental states may be a key mechanism by which people coordinate their thoughts and behaviors over time. People may plan to create desired, and prevent undesired, mental states in their future selves, for example, through scheduled reminders or commitment devices [120,121]. As in the case of planning interventions on other minds, the need to plan interventions on one's own future mind could bias the self-concept towards being organized around abstract and structured causal representations. Future research will need to explore these connections between the demands of coordinating behavior over time, planning interventions on one's own mind, and the structure of the self-concept [122].

and selectively construct **value-guided construals**: simplified, *ad hoc* causal models that are tailored to pursuing a particular goal [41].

Theory of Mind is both abstract and structured (Figure 1). The representation 'Harold wants Lebanese' is abstract, because it elides myriad degrees, characters, or reasons for wanting (e.g., Does he crave Lebanese or merely prefer it to the relevant alternatives? Does he want a specific dish or like the general cuisine? Does he value the flavor or the service at the Lebanese food truck? And so on). In exchange for lost precision, this representation achieves an efficient, compressed, and productive abstraction that preserves the most relevant information for considering possible interventions. Any action that would make Harold stop wanting Lebanese (e.g., telling him the Lebanese truck was poisoned, or feeding him pizza to satiety) would be an equivalent theory-guided intervention, to change his desire and thus his subsequent actions.

The representation 'Harold wants Lebanese' is structured, because it is composed of parts that inherit causal relationships. The causal relationships between 'Harold wants Lebanese', 'Harold believes the truck is in the north lot', and 'Harold plans to go to the north lot' are derived from the Theory of Mind's intuitive components of rational action, in which people have desires and beliefs and then choose an efficient plan of action to pursue their desires given those beliefs [42–46]. This structure allows planners to break down the final goal (make Harold happy) into intermediate goals (change his belief about the food truck) [47,48].

Thus, although abstract, structured, causal models can be used for predictions, they are pivotal for planning.

Generation and pruning

Even when operating with abstract, structured representations, planning can be derailed by intractably large search spaces. For instance, when Grace planned to send Harold to the Lebanese truck without tipping off Luke (Figure 2B), the set of actions she might plausibly search over is enormously large. There are all the actions that could change Harold's beliefs about the location of food trucks (write it in an email, show him a picture of the truck, put it on the evening news, etc.). There are all the actions that could increase Harold's desire to go to the south lot (offer to pay him, send his grandmother there, lie that a gorilla is there, etc.) or decrease his desire to go to the north lot (say the Lebanese truck went bankrupt, make the north lot dangerous, etc.). Even with the ability to abstractly represent many candidates, comparatively evaluating each candidate action would take substantial time and effort.

Recently, across a number of fields, considerable progress has been made in understanding how cognitive search in enormous spaces, like the possible interventions on another mind, can be accomplished. For our purposes, this work can be summarized under two broad themes. First, efficient search depends on exploiting precompiled representations that make good options easily 'available' (i.e., **prioritized memory**) [49]. Prior research suggests several ways this can be accomplished: for instance, by amortizing the outputs of prior computations [50–55], by using heuristics [56], by retrieving candidate actions as a function of their value or frequency of occurrence [49,57–59], or by drawing on precompiled semantic structures [60]. Second, once candidate actions are retrieved, people must consider the unique features of the current task or constraints, to eliminate contextually inappropriate candidates [61]. Thus, the specific constraint of not informing Luke excludes many otherwise high-value actions, like 'Tell Harold the truth'. Rather than perseverating on variants of this possibility, Grace should reject it from consideration in the hopes that a more circumstantially suitable option will come to mind (e.g., 'pass him a written note', or 'lie about the free t-shirts').

Thus, in order to enable efficient planning, Theory of Mind should be arranged to render certain kinds of interventions especially cognitively 'available' and to allow large sets of candidate interventions to be efficiently queried and pruned as circumstances dictate. This is a structural constraint that would not be identified if we considered only the predictive function of Theory of Mind and it is a promising area for further study (see Outstanding questions).

Examples of planning with Theory of Mind

Whereas the classic psychological investigations of Theory of Mind tended to focus on prediction, more recent research often focuses on planning with Theory of Mind. We briefly review three examples involving planning interventions on others' minds and in Box 3 discuss intervening on one's own mind.

Pedagogy and pragmatic speech

One of the most extensively studied forms of planning with Theory of Mind is teaching. When people teach intentionally (as opposed to habitually or instinctively), they must simulate potential pedagogical interventions on a learner's mental state. Teachers choose examples [62,63] or demonstrations [64–66] to cause a learner to draw correct inferences about the world.

Formal models of teaching call special attention to how structure and abstraction in Theory of Mind can explain human behavior. For example, using a 2D navigation paradigm, [66] examined how instrumental intentions to 'do' a task (e.g., avoid squares associated with a penalty) differ from communicative intentions to 'show' an aspect of a task (e.g., show a learner that purple squares have a penalty). Showing one aspect of a task could be explained in terms of planning over an augmented version of doing a task, in which the new task is a composition of the original task and a model of the learner's mind. The capacity to selectively combine Theory of Mind with a causal model of a task, depending on if one is attempting to simply do a task versus show a learner an aspect of the task, is a form of value-guided construal [41].

Pragmatic speech relies on a similar process of planning with Theory of Mind. For example, if Harold asks Grace whether she likes Lebanese food, and she does not, she may plan her words carefully to convey her preferences (which preclude calling it 'excellent') while soothing Harold's feelings (which may be hurt by calling it 'terrible'), and so chose the longer, indirect phrase 'not bad' [67]. Examples like these are well captured by the Rational Speech Act framework, which implements planning in recursive Theory of Mind [68–72].

It is worth noting that planning over recursive mental state representations directly leverages abstraction and structure in Theory of Mind. For instance, in our example (Figure 2), it would be easier for Grace to tip off Harold about the Lebanese truck without alerting Luke if Harold could guess at her goal. This requires Harold to have a second-order representation in which a representation of his own mental state is embedded inside of a representation of Grace's goals. Such embedding operations are a form of compositionality. Additionally, performing operations between distinct but related mental contents at different levels of recursion (e.g., if Grace wants to compare what Luke thinks Harold believes about the south lot with what Harold actually believes about the south lot) relies on Theory of Mind representations being abstract and structured.

Future research on pedagogy and language will need to examine when people engage in true planning with Theory of Mind versus computationally cheaper alternatives (see Outstanding questions). Clearly, not all linguistic utterances are selected by planning to induce a desired mental state in the listener. Some utterances are produced entirely by habit (e.g., reflexively saying 'Good morning' or 'How are you?') and almost all utterances are formed at least partly by

selecting predictable constituents and words, in other words, based on non-causal predictive models or habits. Moreover, consideration of the audience's inferences is highly practiced and so may also be automatized into habit. Effortful planning of utterances is most likely to occur when pursuing novel goals or communicating in a new context, like when repairing a breach of trust, avoiding a high-stakes misunderstanding, sharing sensitive personal information [73,74], or giving critical evaluative feedback [75]. In these social contexts, people may elaborately plan, and even rehearse, verbal utterances before delivering them. By contrast, existing computational models of pedagogy and pragmatics have tended to focus on highly simplified settings, where the space of possible goals and actions is small. Future research should investigate how humans scale up these computations through the use of heuristics, hierarchies, and abstractions.

Interpersonal affect regulation

One common goal of planning with Theory of Mind is to create specific emotional reactions in other people. Interpersonal affect regulation [76] occurs when people deliberately modulate others' emotions, such as when one person tries to help another feel better, calm down, or regain control, after a negative experience [77]. Yet not all plans are designed to make the target feel better. In some cases, people aim to worsen the target's affect, in order to help them [78], or make the target angry, to motivate them to accomplish prosocial goals [79]. Suffering may also be deliberately induced as a punishment [80].

As predicted for planning, these examples of interpersonal affect regulation depend on deliberate, controlled cognitive processes that select interventions to cause particular affective outcomes [76,81,82]. However, other people's emotional reactions, like smiles, are also directly rewarding, so action selection for interpersonal emotion regulation could be strongly influenced by model-free reinforcement learning.

Existing formal computational models of Theory of Mind can be adapted to include abstract causal models of emotions, defined by relating beliefs and desires to outcomes [8,83,84]. Such models can be used to predict others' emotional reactions (e.g., when Harold gets to the north lot, he will feel disappointed). Future research is needed to define benchmark tasks for planning in this domain (e.g., what should Grace do to make Harold feel relieved?), on which human and model performance can be directly compared.

Impression management

Surely one of the most common ways we want to change other's thoughts is to make them think better of us. People want to be perceived favorably and plan strategically to achieve this goal [85–87], by impression management. For instance, people frequently choose actions that other people see as valuable or that convey valued traits [88]. This motive can induce people to engage in costly acts of altruism, depending on whether they are being observed and by whom. For example, people engage in costly third-party punishment when it will make the punisher appear unselfish [89–93].

Impression management goals can be more specific than general approbation: people may wish to be perceived as competent rather than warm [94,95], or vice versa [96]. Even preschool children engage in strategic impression management by flexibly modifying their behavior as a function of their audience's knowledge [97,98].

Consistent with being a controlled process, impression management can be impaired by load [99,100]. People who have a deliberative cognitive style are especially sensitive to the

opportunities for impression management in unusual situations [101]. However, the need for impression management arises frequently, in similar ways, so some strategies are likely to be automatized into habit.

A formal computational model of recursive Theory of Mind, derived from the Rational Speech Act framework, has been used to capture human behavior on impression management tasks, including: how people balance the value of learning against the desire to appear competent [67,97,98], and how people balance the desire to appear impartial with the desire to reward effort with pay [102]. However, existing tasks remain highly simplified, considering only a narrow range of possible actions and valued outcomes. Thus, these tasks do not yet evoke the cognitive challenges that typify planning with Theory of Mind: the demand for broad and deep search, in a large space of abstract and structured representations, to quickly generate and then prune a large set of candidate actions, from which a single, often completely novel, action must be selected.

Concluding remarks

The traditional emphasis on problems of predicting others' actions has afforded a controversy. Do humans actually need, or use, Theory of Mind? A well-trained statistical predictive model is sufficient to solve many prediction problems and often easier to use for online computations. Here, we suggest that Theory of Mind is deployed not only for predicting and evaluating others' actions, but also for efficiently planning to change them. A unified solution to these diverse problems, particularly for generalizing to novel or unusual circumstances, requires a causal, abstract, structured model of other minds. Problems of planning thus offer unique insight into why people have Theory of Mind.

Outstanding questions

When a person successfully predicts someone else's actions, how can we diagnose whether they used Theory of Mind versus a well-trained predictive model?

When a person successfully intervenes on someone else's reactions, how can we diagnose whether they used Theory of Mind versus a habit?

Can computations in Theory of Mind be automatized, amortized, or compressed into value-guided task construals? If so, do prediction and planning still rely on the same causal, abstract, structured representations as controlled processing with Theory of Mind?

How are the mental states that afford effective and reliable intervention prioritized during the search process?

How are patterns of neural response, within or between brain regions, different when Theory of Mind is used for planning, versus prediction or moral evaluation?

How does the use of Theory of Mind for planning arise in development and vary across cultural contexts?

Acknowledgments

The authors would like to thank Xuechunzi Bai and Carlos Giovanni Correa for feedback on earlier versions of this manuscript. R.S. was supported by the Patrick J. McGovern foundation, the Guggenheim foundation, and the Paul and Lilah Newton Brain Science fund. F.C. was supported by grant 61061 from the John Templeton Foundation and grant N00014-19-1-2025 from the Office of Naval Research.

Declaration of interests

No interests are declared.

References

1. Premack, D. and Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526
2. Baron-Cohen, S. *et al.* (1985) Does the autistic child have a "theory of mind"? *Cognition* 21, 37–46
3. Dennett, D.C. (1988) Précis of the intentional stance. *Behav. Brain Sci.* 11, 495–505
4. Gerstenberg, T. and Tenenbaum, J.B. (2017) *Intuitive theories, Oxford handbook of causal reasoning*. pp. 515–548
5. Flavell, J.H. (1999) Cognitive development: Children's knowledge about the mind. *Annu. Rev. Psychol.* 50, 21–45
6. Wimmer, H. and Perner, J. (1983) Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128
7. Baker, C.L. *et al.* (2017) Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* 1, 1–10
8. Ong, D.C. *et al.* (2019) Computational models of emotion inference in theory of mind: A review and roadmap. *Top. Cogn. Sci.* 11, 338–357
9. Tamir, D.I. and Thornton, M.A. (2018) Modeling the predictive social mind. *Trends Cogn. Sci.* 22, 201–212
10. Thornton, M.A. and Tamir, D.I. (2021) People accurately predict the transition probabilities between actions, Science. *Advances* 7, eabd4995
11. Rabinowitz, N. *et al.* (2018) Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227, PMLR
12. Sutton, R.S. and Barto, A.G. (2018) *Reinforcement learning: An introduction*, MIT Press
13. Russell, S. and Norvig, P. (2009) *Artificial Intelligence: A Modern Approach* (3rd Edition), Prentice Hall Press, USA
14. Dayan, P. and Niv, Y. (2008) Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* 18, 185–196
15. Daw, N.D. *et al.* (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215
16. Pearl, J. (2009) *Causality*, Cambridge University Press
17. Wood, W. and Rünger, D. (2016) Psychology of habit. *Annu. Rev. Psychol.* 67, 289–314
18. Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature* 518, 529–533
19. Hackel, L.M. and Amodio, D.M. (2018) Computational neuroscience approaches to social cognition. *Curr. Opin. Psychol.* 24, 92–97
20. Dezfouli, A. and Balleine, B.W. (2012) Habits, action sequences and reinforcement learning. *Eur. J. Neurosci.* 35, 1036–1051
21. Miller, K.J. *et al.* (2019) Habits without values. *Psychol. Rev.* 126, 292
22. Chi, M.T.H. *et al.* (1982) Expertise in problem solving. In *Advances in the psychology of human intelligence* (Vol. 1) (Sternberg, R.J., ed.), pp. 7–75, Erlbaum, Hillsdale, NJ

23. Ouellette, J.A. and Wood, W. (1998) Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychol. Bull.* 124, 54
24. Papadimitriou, C.H. and Tsitsiklis, J.N. (1987) The complexity of markov decision processes. *Math. Oper. Res.* 441–450
25. Goldsmith, J. *et al.* (1997) The complexity of plan existence and evaluation in probabilistic domains. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence* (Vol. UAI-97), pp. 182–189, Morgan Kaufmann Publishers, San Francisco, CA
26. FeldmanHall, O. and Nassar, M.R. (2021) The computational challenge of social learning. *Trends Cogn. Sci.* 25, 1045–1057
27. Otto, A.R. *et al.* (2013) The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* 24, 751–761
28. McDannald, M.A. *et al.* (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* 31, 2700–2705
29. Daw, N.D. and Dayan, P. (2014) The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. B Biol. Sci.* 369
30. Solway, A. and Botvinick, M.M. (2015) Evidence integration in model-based tree search. *Proc. Natl. Acad. Sci.* 112, 11708–11713
31. Balaguer, J. *et al.* (2016) Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron* 90, 893–903
32. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40
33. Pouncy, T. *et al.* (2021) What is the model in model-based planning? *Cogn. Sci.* 45, e12928
34. Ho, M.K. *et al.* (2019) The value of abstraction. *Curr. Opin. Behav. Sci.* 29, 111–116
35. Oh, J. *et al.* (2015) Action-conditional video prediction using deep networks in atari games. *Adv. Neural Inf. Proces. Syst.* 28, 2863–2871
36. Schrittwieser, J. *et al.* (2020) Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588, 604–609
37. Grimm, C. *et al.* (2020) The value equivalence principle for model-based reinforcement learning. *Adv. Neural Inf. Proces. Syst.* 33
38. Chomsky, N. (1959) A review of BF Skinner's Verbal behavior. *Language* 35, 26–58
39. Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3–71
40. Boutilier, C. *et al.* (1999) Decision-theoretic planning: Structural assumptions and computational leverage. *J. Artif. Intell. Res.* 11, 1–94
41. Ho, M.K. *et al.* (2022) People construct simplified mental representations to plan. *Nature* https://doi.org/10.1038/s41586-022-04743-9
42. Baker, C.L. *et al.* (2009) Action understanding as inverse planning. *Cognition* 113, 329–349
43. Dennett, D.C. (1989) The intentional stance. *MIT Press*
44. Malle, B.F. (1999) How people explain behavior: A new theoretical framework. *Personal. Soc. Psychol. Rev.* 3, 23–48
45. Gopnik, A. and Wellman, H.M. (1992) *Why the child's theory of mind really is a theory*.
46. Saxe, R. *et al.* (2004) Understanding other minds: linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.* 55, 87–124
47. Correa, C.G. *et al.* (2020) Resource-rational task decomposition to minimize planning costs. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (Denison, S. *et al.*, eds), pp. 2974–2980, Cognitive Science Society
48. Tomov, M.S. *et al.* (2020) Discovery of hierarchical representations for efficient planning. *PLoS Comput. Biol.* 16
49. Morris, A. *et al.* (2021) Generating options and choosing between them depend on distinct forms of value representation. *Psychol. Sci.* 32, 1731–1746
50. Huys, Q.J.M. *et al.* (2015) Interplay of approximate planning strategies. *Proc. Natl. Acad. Sci.* 112, 3098–3103
51. Keramati, M. *et al.* (2016) Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proc. Natl. Acad. Sci.* 113, 12868–12873
52. Momennejad, I. *et al.* (2017) The successor representation in human reinforcement learning. *Nat. Hum. Behav.* 1, 680–692
53. Kool, W. *et al.* (2018) *Competition and cooperation between multiple reinforcement learning systems*. Goal-directed decision making pp. 153–178
54. Dasgupta, I. *et al.* (2018) Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition* 178, 67–81
55. Sezener, C.E. *et al.* (2019) Optimizing the depth and the direction of prospective planning using information values. *PLoS Comput. Biol.* 15, e1006827
56. van Opheusden, B. *et al.* (2021) *Revealing the impact of expertise on human planning with a two-player board game*. https://doi.org/10.31234/osf.io/rhq5j. URL psyarxiv.com/rhq5j
57. Mattar, M.G. and Daw, N.D. (2018) Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* 21, 1609–1617
58. Callaway, F. *et al.* (2022) Rational use of cognitive resources in human planning. *Nat. Hum. Behav.* 1–14
59. Cushman, F. and Morris, A. (2015) Habitual control of goal selection in humans. *Proc. Natl. Acad. Sci.* 112, 13817–13822
60. Zhang, Z. *et al.* (2021) Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proc. Natl. Acad. Sci.* 118
61. Huys, Q.J. *et al.* (2012) Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput. Biol.* 8, e1002410
62. Shafto, P. *et al.* (2014) A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cogn. Psychol.* 71, 55–89
63. Rafferty, A.N. *et al.* (2016) Faster teaching via pomdp planning. *Cogn. Sci.* 40, 1290–1332
64. Gweon, H. and Schulz, L. (2019) From exploration to instruction: Children learn from exploration and tailor their demonstrations to observers' goals and competence. *Child Dev.* 90, e148–e164
65. Bridgers, S. *et al.* (2020) Young children consider the expected utility of others' learning to decide what to teach. *Nat. Hum. Behav.* 4, 144–152
66. Ho, M.K. *et al.* (2021) Communication in action: planning and interpreting communicative demonstrations. *J. Exp. Psychol. Gen.* 150, 2246–2272
67. Yoon, E.J. *et al.* (2020) Polite speech emerges from competing social goals. *Open Mind* 4, 71–87
68. Goodman, N.D. and Frank, M.C. (2016) Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* 20, 818–829
69. Popp, P.J.O. and Gureckis, T.M. (2020) Ask or tell: Balancing questions and instructions in intuitive teaching. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (Denison, S. *et al.*, eds), pp. 1229–1235, Cognitive Science Society
70. Hawkins, R.D. *et al.* (2021) The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cogn. Sci.* 45, e12926
71. Kao, J.T. *et al.* (2014) Nonliteral understanding of number words. *Proc. Natl. Acad. Sci.* 111, 12002–12007
72. Sumers, T.R. *et al.* (2022) How to talk so your robot will learn: Instructions, descriptions, and pragmatics. *arXiv* preprint. arXiv:2206.07870
73. Gambetta, D. (2011) *Codes of the underworld, in: Codes of the Underworld*, Princeton University Press
74. Small, M.L. (2017) *Someone to talk to*, Oxford University Press
75. Heaphy, E. *et al.* (2022) Moved to speak up: How prosocial emotions influence the employee voice process. *Hum. Relat.* 75, 1113–1139
76. Niven, K. (2017) The four key characteristics of interpersonal emotion regulation. *Curr. Opin. Psychol.* 17, 89–93
77. Gummerum, M. and López-Pérez, B. (2020) "you shouldn't feel this way!" children's and adolescents' interpersonal emotion regulation of victims' and violators' feelings after social exclusion. *Cogn. Dev.* 54, 100874
78. López-Pérez, B. *et al.* (2017) Cruel to be kind: Factors underlying altruistic efforts to worsen another person's mood. *Psychol. Sci.* 28, 862–871

79. Niven, K. *et al.* (2019) Prosocial versus instrumental motives for interpersonal emotion regulation. *J. Theor. Soc. Psychol.* 3, 85–96
80. Rai, T.S. *et al.* (2017) Dehumanization increases instrumental violence, but not moral violence. *Proc. Natl. Acad. Sci.* 114, 8511–8516
81. Niven, K. *et al.* (2009) A classification of controlled interpersonal affect regulation strategies. *Emotion* 9, 498
82. Netzer, L. *et al.* (2015) Interpersonal instrumental emotion regulation. *J. Exp. Soc. Psychol.* 58, 124–135
83. Saxe, R. and Houlihan, S.D. (2017) Formalizing emotion concepts within a Bayesian model of theory of mind. *Curr. Opin. Psychol.* 17, 15–21
84. Wu, Y. *et al.* (2021) Emotion as information in early social learning. *Curr. Dir. Psychol. Sci.* 30, 468–475
85. Leary, M.R. and Kowalski, R.M. (1990) Impression management: A literature review and two-component model. *Psychol. Bull.* 107, 34
86. Schlenker, B.R. (2012) *Self-presentation*.
87. Kim, J. and Crockett, M.J. (2022) Narrating the" what" and" why" of our moral actions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44)
88. Sznycer, D. (2022) Value computation in humans. *Evol. Hum. Behav.*
89. Kurzban, R. *et al.* (2007) Audience effects on moralistic punishment. *Evol. Hum. Behav.* 28, 75–84
90. Jordan, J.J. *et al.* (2016) Third-party punishment as a costly signal of trustworthiness. *Nature* 530, 473–476
91. Raihani, N.J. and Bshary, R. (2015) The reputation of punishers. *Trends Ecol. Evol.* 30, 98–103
92. Rai, T.S. (2022) Material benefits crowd out moralistic punishment. *Psychol. Sci.* 0956797621054786
93. Radkani, S. *et al.* (2022) Modeling punishment as a rational communicative social action. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp. 44
94. Swencionis, J.K. and Fiske, S.T. (2016) Promote up, ingratiate down: Status comparisons drive warmth-competence tradeoffs in impression management. *J. Exp. Soc. Psychol.* 64, 27–34
95. Dupree, C.H. and Fiske, S.T. (2019) Self-presentation in interracial settings: The competence downshift by white liberals. *J. Pers. Soc. Psychol.* 117, 579
96. Holoien, D.S. and Fiske, S.T. (2013) Downplaying positive impressions: Compensation between warmth and competence in impression management. *J. Exp. Soc. Psychol.* 49, 33–41
97. Asaba, M. and Gweon, H. (2018) Look, i can do it! young children forego opportunities to teach others to demonstrate their own competence. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society, Cognitive Science Society* (Kalish, C. *et al.*, eds), pp. 106–111
98. Asaba, M. and Gweon, H. (2019) *Young children can rationally revise and maintain what others think of them*. https://doi.org/10.31234/osf.io/yxhv5. URL psyarxiv.com/yxhv5
99. Paulhus, D.L. *et al.* (1989) Attentional load increases the positivity of self-presentation. *Soc. Cogn.* 7, 389–400
100. Paulhus, D.L. and Levitt, K. (1987) Desirable responding triggered by affect: Automatic egotism? *J. Pers. Soc. Psychol.* 52, 245
101. Jordan, J.J. and Rand, D.G. (2020) Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *J. Pers. Soc. Psychol.* 118, 57
102. Kleiman-Weiner, M. *et al.* (2017) Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *CogSci*
103. Cushman, F. (2015) Deconstructing intent to reconstruct morality. *Curr. Opin. Psychol.* 6, 97–103
104. Sosa, F.A. *et al.* (2021) Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition* 217, 104890
105. Young, L. *et al.* (2010) Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc. Natl. Acad. Sci.* 107, 6753–6758
106. Mikhail, J. (2007) Universal moral grammar: Theory, evidence and the future. *Trends Cogn. Sci.* 11, 143–152
107. Cushman, F. and Greene, J. (2012) *The philosopher in the theater*.
108. Chakroff, A. and Young, L. (2015) How the mind matters for morality. *AJOB Neurosci.* 6, 43–48
109. Cushman, F. (2013) Action, outcome, and value: A dual-system framework for morality. *Personal. Soc. Psychol. Rev.* 17, 273–292
110. Crockett, M.J. (2013) Models of morality. *Trends Cogn. Sci.* 17, 363–366
111. Nichols, S. (2021) *Rational rules: Towards a theory of moral learning*, Oxford University Press
112. Tamir, M. (2016) Why do people regulate their emotions? a taxonomy of motives in emotion regulation. *Personal. Soc. Psychol. Rev.* 20, 199–222
113. Kalokerinos, E.K. *et al.* (2017) Instrumental motives in negative emotion regulation in daily life: Frequency, consistency, and predictors. *Emotion* 17, 648
114. Weidman, A.C. and Kross, E. (2020) Examining emotional tool use in daily life. *J. Pers. Soc. Psychol.* 120, 1344–1366
115. English, T. *et al.* (2017) Emotion regulation strategy selection in daily life: The role of social context and goals. *Motiv. Emot.* 41, 230–242
116. Tamir, M. *et al.* (2015) An expectancy-value model of emotion regulation: Implications for motivation, emotional experience, and decision making. *Emotion* 15, 90
117. Bigman, Y.E. *et al.* (2016) Yes i can: Expected success promotes actual success in emotion regulation. *Cognit. Emot.* 30, 1380–1387
118. Gutentag, T. *et al.* (2017) Successful emotion regulation requires both conviction and skill: beliefs about the controllability of emotions, reappraisal, and regulation success. *Cognit. Emot.* 31, 1225–1233
119. Ford, B.Q. and Gross, J.J. (2019) Why beliefs about emotion matter: An emotion-regulation perspective. *Curr. Dir. Psychol. Sci.* 28, 74–81
120. Gul, F. and Pesendorfer, W. (2001) Temptation and self-control. *Econometrica* 69, 1403–1435
121. Milyavskaya, M. *et al.* (2021) Self-control in daily life: Prevalence and effectiveness of diverse self-control strategies. *J. Pers.* 89, 634–651
122. Cushman, F. (2020) Rationalization is rational. *Behav. Brain Sci.* 43
123. Koster-Hale, J. and Saxe, R. (2013) *Functional neuroimaging of theory of mind*.
124. Saxe, R. and Wexler, A. (2005) Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43, 1391–1399
125. Cloutier, J. *et al.* (2011) An fmri study of violations of social expectations: when people are not who we expect them to be. *NeuroImage* 57, 583–588
126. Young, L. *et al.* (2007) The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci.* 104, 8235–8240
127. Achim, A.M. *et al.* (2021) The neural correlates of referential communication: Taking advantage of sparse-sampling fmri to study verbal communication with a real interaction partner. *Brain Cogn.* 154, 105801
128. Salazar, M. *et al.* (2021) You took the words right out of my mouth: Dual-fmri reveals intra-and inter-personal neural processes supporting verbal interaction. *NeuroImage* 228, 117697
129. Lemmers-Jansen, I.L. *et al.* (2018) Giving others the option of choice: An fmri study on low-cost cooperation. *Neuropsychologia* 109, 1–9
130. Shen, S.-S. *et al.* (2021) Collaborations and deceptions in strategic interactions revealed by hyperscanning fmri. *BioRxiv*
131. Lin, X.A. *et al.* (2021) Neural correlates of spontaneous deception in a non-competitive interpersonal scenario: A functional near-infrared spectroscopy (fnirs) study. *Brain Cogn.* 150, 105704