## Simulation, Predictive Coding, and the Shared World

## Robert M Gordon

Contribution to K. Ochsner and M. Gilead (Eds.). The Neural Basis of Mentalizing.  Springer Press.

**Acknowledgements**

The author is enormously grateful to Michael Gilead for guidance, encouragement, and feedback at several steps along the way. Many thanks to Karl Friston and Timothy Williamson for very welcome comments and suggestions for future development, and to Gualtiero Piccinini and Frank Van Overwalle for helpful early discussions of the topic. The idea of simulation as neural reuse owes much to the influence of Susan Hurley.

The debate between the "simulation" theory and the "theory" theory, initiated in the late 1980's, concerns the source of everyday human competence in predicting and explaining human behavior, including the capacity to ascribe mental states. This competence is approximately what the term *mentalizing* designates, when understood in its broadest sense. Since the 1960's it was widely assumed that the source of this competence is a body of implicit general knowledge or theory, commonly called “folk psychology” by philosophers and “theory of mind” by psychologists. This was usually understood to consist in a body of general information whose core stipulation is that intentional action is a causal product of the agents’ beliefs and desires.

The “simulation” theory locates the main source of mentalizing competence in a procedure or set of procedures called “simulation,” or “mental simulation.” Introduced by philosophers (Gordon 1986; Heal 1986; Goldman 1989), this account is usually thought to challenge the very assumption that mentalizing is an application of an implicit theory of mental states. The “theory versus simulation debate” soon became a topic of interest among developmental psychologists and later received attention in linguistics, social cognitive neuroscience, and social robotics.

One of the initial motivations for a simulation account of mentalizing was that it seemed to spare the brain the overhead costs predicated by the prevailing “theory” theory. These were the costs of acquiring, storing, and utilizing the theory. An important part of the simulationist response was to ask why a system would need to invest in a general theory or model of *systems like itself*. Wouldn’t it be more economical simply to use itself as a stand-in for these other, similar systems? Of course, “using itself as a stand-in” may be understood in different ways, and, as a consequence, various distinct simulationist approaches have developed.

The approach presented here aims to show how one’s own action planning system may serve as a stand-in for the action planning systems of other agents. Building on recent work on inverse planning, it explains how mentalizing by simulation can offer vastly greater economies than the mere elimination of the information-rich overhead required by a theory theory. The simulation approach as presented here is in fact much in line with the current view in psychology and neuroscience that neural systems tend to reduce metabolic and other expenses by conforming to a predictive coding strategy. This is a strategy of “guessing ahead.” Rather than waiting for the world to bombard us with new information, the system makes its latest best guess as to what will be coming in. This process of predicting input values minimizes the need for new information input, in that only discrepancies, or information that conflicts with the predicted values (prediction errors), need be encoded.

Indeed, simulation has been compared to compression schemes commonly used in the digital transmission and storage of video content (Gordon 1992). These schemes exploit the likelihood that video content will be redundant in a number of ways. Most important is temporal redundancy. Typically, little or no visual content changes in, say, the thirtieth of a second that separates one frame from the next; successive frames in a video sequence are nearly always very similar. Therefore it is an efficient strategy to treat each frame as “predicting” its successor. The default, or uncorrected, prediction, would yield a sequence of undifferentiated frames: essentially, a still picture. Any corrections, or departures from this default, are likely to be relatively small, requiring minimal resources to encode these differences.

Video compression was an early engineering application of predictive coding. A comparable simulation account should show how our mentalizing system exploits massive redundancies to achieve extreme code compression and resource parsimony. Simulation, as I understand it, does just that, I believe.

### Two kinds of projection

In broad view, simulation exemplifies a type of predictive strategy that begins with what is in effect a forward model – projection -- creating a default expectation. However, I should note that the term *projection* (or *self-*projection) is ambiguous in this context, yielding two distinct metaphors: “projection *onto”* and “projection *into*.” To project *onto* another person or entity is to push or impose (etymologically, *to throw*) one’s own image (or perspective or “way of seeing things”) onto the other, thereby assimilating the other to oneself.[[1]](#footnote-1) Projection *into*, on the other hand, is metaphorically a kind of travel, where it is we who move, rather than our image: we are transported into a perspective that is not currently our own. Understood in this second way, projection is not an imposition of one’s own perspective, but rather a shift to a different perspective: This may be the perspective of another human or other sentient being. Or it may involve mental time travel to a past perspective (as in episodic memory) or a future perspective (prospection). It might be travel to a counterfactual perspective, to an “alternative” past, present, or future. It has been suggested that the various forms of projection *into* may in fact be supported by a single specialized brain network (Buckner & Carroll 2007).

On the projection-and-correction account, simulation entails bothkinds of projection. It begins with projection *onto* another presumably sentient being, imposing one’s own perspective as an initial *a priori* prediction of the other’s; and it ends at, or at least aims at, being projected *into* the other’s perspective. Simulation proceeds from *onto* to *into* by a series of “corrections,” or *corrected projections onto.* These corrections may come from a comparison of predicted behavior with observed behavior, from internal resonances to observed behavior, from contextual evidence of various kinds, and other sources. In short, mentalizing by simulation begins with an uncorrected projection *onto* a target and then, in response to predictive errors, tests hypothetical modifications of this projection until a good enough projection *into* is achieved.

### An important problem not addressed here

The projection-and-correction account of simulation theory conforms to what Andy Clark calls “the core predictive coding strategy.” However, it does not entail a much more ambitious package, which Clark distinguishes as “hierarchical predictive processing” (Clark 2013). The latter not only analyzes neural inference as a process of prediction and correction; it also aims to specify the inference mechanism by which predictions are made and corrected. It posits a hierarchy of intermediate predictions and corrections, each of which operates by Bayesian inference. In this manner higher-level predictions are thought to pass stepwise down to lower levels, and lower-level corrections are in a similar fashion passed back up to higher levels. This has been a very influential idea (Rao & Ballard 1999; Friston 2005; Clark 2013). However, I believe the argument that follows does not require commitment to the hierarchical Bayesian account of prediction and correction. The focus here will be on the initial projections, or the a priori starting point, or what I will be calling the default condition for mentalizing.[[2]](#footnote-2)

##  Inverse planning

The holy grail for a theory of mentalizing is to account for our capacity to grasp the intentions behind observed behavior; and beyond the intentions, detect the goals and reasons motivating these intentions. The aim, one might say, is to illuminate the background that makes the observed behavior unsurprising.

To address the question of how the brain interprets the observed actions of others, it has been suggested that we adopt a predictive framework that has been particularly fruitful in studies of vision:

it is often said that "vision is inverse graphics" - the inversion of a causal physical process of scene formation (Baker, Tenenbaum, & Saxe 2006).

Just as the interpretation of a visual scene might involve, essentially, using in reverse the process of producing such a scene, so the interpretation of another’s behavior might be understood as a comparable inverse problem (Baker, Saxe, & Tenenbaum 2011; Baker, Saxe, & Tenenbaum 2009).

By analogy, our analysis of intentional reasoning might be called "inverse planning", where the observer infers an agent's intentions, given observations of the agent's behavior, by inverting a model of how intentions cause behavior. (Baker et al 2011)

The process is *inverted* in that, instead of proceeding forward from a given intention to its behavioral execution, it takes the behavior as the given and determines the intention most likely to have produced it. The planning process would thus be used as a mechanism for testing hypotheses about underlying intentions.

In the broadest terms, inverse planning exemplifies hypothesis-testing as unconscious inference, an idea introduced in the perceptual realm by Helmholtz (1856). The proposal bears some resemblance to “hypothetico-practical” inference (Gordon 1986), modeled on a traditional model of the scientific method, hypothetico-deductive inference. Instead of forming hypotheses and *deducing* consequences that match observations, hypothetico-*practical* inference would form hypotheses and then *­act on* them, *producing* consequences that match the observed behavior of the other agent. (Typically, the resulting action would be merely covert, inhibited from any outward expression that might be perceptible to others.) This was conceived as an inference requiring personal agency – as something *I* do, rather than as computational operations of a “subpersonal” neural system such as an action planning system. This chapter concerns such computational operations, building on a variant of the inverse planning proposal.

### Model or reuse?

The remainder of this chapter builds on a variant of the inverse planning thesis proposed by Baker et al 2009 and Baker et al 2011.

The term “inverse planning” appears to suggest that the very mechanism that is used to plan our own behavior may be reused as a platform for testing hypothetical explanations of the observed behavior of other agents. However, Baker et al 2011 actually proposes something more complicated. The authors speak of inverting a *model* or *theory* of the planning process. One might understand them to be saying that the action planning system itself can be used as a general model of an action planning system. But this is not their view. As they point out,

on first glance our work appears most consistent with the ‘‘theory-based” approach. Formalizing an intuitive theory of mind was in fact one of our original motivations…. On a theory-based interpretation, inverse planning consists of inverting a causal theory of rational action to arrive at a set of goals that could have generated the observed behavior, and inferring individual goals based on prior knowledge of the kinds of goals the observed agent prefers.

(A closely related view [Jara-Ettinger 2019] treats action understanding as inverse reinforcement learning: determining what model of the world and what positive and negative reinforcers would best explain an agent’s observed actions.)

The theory-based approach attributes to the brain a capacity for detachment: it *stands back from its own operations* and employs instead a general theory or model of these operations. As distinct from actual action-planning, the theory theorist proposal is that in mentalizing about others the brain engages in *plan-theorizing, theorizing about* the steps in the other’s planning process. The proposal assumes that we humans have an intuitive theory of mind and that our brains employ this theory not only in our explicit attributions of mental states but also in its unconscious subpersonal neural processing. I will call this *inverse plan-theorizing*. Thus understood, it does not make use of our capacity for planning: it is not inverse planning as such, i.e., an inverse reuse of one’s own action planning system.

Although Baker et al 2009 explicitly develops this plan-theorizing version of inverse planning, it acknowledges that a simulation-based account would cover the data just as well as their theory-based account. The simulation account would use, not a *model* of the planning process, but the planning process itself (running offline), as a mechanism for testing hypotheses about underlying intentions:

On a simulation account, goal inference is performed by inverting one's own planning process - the planning mechanism used in model-based reinforcement learning - to infer the goals most likely to have generated another agent's observed behavior.

If indeed such reuse of its own “first-person” planning system (what I will call *1p* inverse planning) would be sufficient for goal inference, the question arises, Why would the brain need to operate instead on a model of the planning process? Here again, using an existing system would avoid the overhead costs of storing and utilizing an information-rich theory or model.

Moreover, first person inverse planning would seem to be the proper analogue of the inverse graphics account of vision. As inverse graphics is "the inversion of a causal physical process of scene formation"(Baker et al 2011), so inverse planning should be the inversion of *a physical process* of action determination – *not* the inversion of a causal *theory of* a physical processof action determination. The "vision is inverse graphics" idea is generally understood to be an analysis-by-synthesis paradigm, and analysis by synthesis is not analysis by a *theory of* synthesis.

The perceptual system … is a mechanism for the hypothetical "synthesis" of natural images, in the style of computer graphics. Perception (or "analysis") is then the search for or inference to the best explanation of an observed image in terms of this synthesis. (Yildirim et al., 2015)

In other words, just as visual perception is thought to weigh alternative hypothetical ways of building a given scene, understanding action in terms of goals and intentions would be a search among alternative hypothetical ways of generating (planning) a given action, in an effort to find the most plausible simulation of the planning that might have generated the action. Action understanding as inverse planning would thus really be, in analogy to vision as inverse graphics, a case of analysis by synthesis.

In addition, inverse planning, as a reuse of one’s own planning system, would be in a position to exploit responses to the behavior of others that are themselves reuses of one’s own motor system. These would include various forms of motor resonance, including mirror neuron activation, motor mimicry, and suppressed action imitation. Rather than having to work with bare visual input, first person inverse planning could work on lower-level input already formatted in terms of first person motor planning. This suggests a considerable advantage not available to a model-based understanding of inverse planning.

### Use and reuse of action planning

I will suppose then that the human action planning system has, in addition to its primary use in generating one’s own actions, a reuse, or secondary use, in which the planning process is inverted in order to infer the goals and reasons that lie behind another agent's observed behavior. This dual use of the same system, I will argue, offers two major advantages. First, it would provide an important head start in understanding the basis of another’s behavior; and, second, it would make possible the most economical coding available to a mentalizing system.

It appears likely that the secondary use of the action planning system, namely, inverse reuse for explanatory purposes, runs concurrently with its primary use, for generating one’s own actions; otherwise, we would have to suspend our own actions in order to understand the actions of others. Thus the system is translating existing inputs into action and at the same time looking for hypothetical inputs that would explain the perceived actions of others. Concurrent processing for self-action and other-understanding would be consistent with evidence of “motor contagion,” or interference effects between observed and executed actions. First noted in the case of biological movements, it has been suggested that motor contagion may be “the first step in a more sophisticated predictive system that allows us to infer goals from the observation of actions” (Blakemore & Frith 2005). Indeed, recent research indicates that such interference is markedly increased when the observed movement is directed toward a visible goal (Bouquet, Shipley, Capa, & Marshall 2011). This interference suggests a competition for resources, and thus that the same, or strongly overlapping, neural resources are employed concurrently in goal-directed action planning and in interpreting the goal-directed actions of others.

Such concurrent double employment raises the question: What, if anything, must *change* as the planning system switches from primary use to reuse, and from self to other? Specifically, what happens to the existing inputs? When the system switches to inverse planning as it seeks to explain another’s behavior, does it clear the slate and approach the task with no a priori top-down commitments? More specifically, does it suspend the beliefs, desires, preferences, emotional valences, affordances, and other influences on one’s own action planning?

Consider three options:

1. The *suspend-all* option: Suspend all existing inputs and start with a blank slate. The mental states that lead to one’s own actions have no informational value for understanding the underlying causes of another agent’s behavior.
2. The *keep-all* option: Keep all existing inputs, add no others, and seek the best explanation of the other agent’s behavior strictly on the basis of one’s own mental states.
3. The *modify-as-needed* option: Keep existing inputs, but allow them to be suspended or modified as needed, and allow new inputs as needed.

Options 1 and 2 should be rejected out of hand. Consider this example: we see a puddle in someone’s path, and we expect the individual to deviate from a straight path. When they do, we readily surmise that they did so for a reason, namely, that there was a puddle in their path. More fully, they did so because stepping in a puddle gets you wet, and so they deviated in order to continue on their journey without getting wet.

Option 1, the suspend-all option, might make sense if it were useless, or at least a bad bet, to project onto the other our own perception of a puddle, as distinct from a patch of dry pavement or, for that matter, a manhole or a bed of geraniums. Likewise, if it were useless to project our own desire to avoid getting wet under similar circumstances. However, such projections, and the expectations and explanations based on them, are not bad bets in general, even if they are sometimes in need of correction.

Option 2, the keep-all option, lies at the opposite extreme. It locks all explanations of the behavior of others into our own mental mold, leaving us unable to accommodate differences between ourselves and others. It does not allow inverse planning to move beyond simple projection onto others.

Puddle-avoidance may seem to present a trivial problem of action understanding. It is common behavior, and it appears to be a matter of common sense. Instead of calling on our own desire to avoid puddles, we might simply apply the generalization, “People tend to avoid puddles.” Similar generalizations would apply to avoidance of snarling dogs, bears, and “shady-looking” people. Such generalizations do not necessarily compete with projection, however; in fact, it is at least plausible that they are themselves products of projection. I am fairly confident that my own acceptance of a generalization like, “People tend to avoid stepping in puddles,” is not based on extensive observation of people confronted with puddles. More likely, it is a projection of my own desire. Further evidence emerges in cases of conflicting desires. For example, will a runner in the hundred-meter run around a puddle on the track, losing time, or plow through it? There is no general rule I apply to this question, certainly not one based on observation. Rather, I project my own competing desires, letting the specific situation (e.g., is this a practice run or the real thing?) dictate the answer.

### Agent-neutral coding

Although such projection would provide inverse planning with an important head start, certainly better than to begin the process without any a priori predictive input, a capacity to modify the default projection would be advantageous as well. This leaves us with Option 3, the modify-as-needed option. As in Option 2, inverse-planning (i.e., hypothetically planning) of another’s actions starts with the existing inputs to planning our own actions. I will call this *agent-neutral* coding. The same undifferentiated coding would serve in two capacities, as our own desire to avoid puddles in our path and, within the context of inverse planning, as the other’s desire to avoid puddles in their path. However, per Option 3, the agent-neutral coding is subject to revision.

By *agent-neutral* coding I mean *identical*, *undifferentiated* coding, the same for self and other. This may seem problematic. Surely the brain must be able somehow to distinguish its own states from the represented states of others. Authors who speak of *shared representation*s, or *shared self-other representation*s, usually emphasize that coding for self and other is overlapping but not identical (Decety & Sommerville 2003). Jeannerod and Pacherie (2004) have argued that, when the actions of others are simulated in the brain by representations shared with similar actions of our own, they elicit agent-neutral or unattributed (“naked”) intentions.[[3]](#footnote-3) These neutrally coded intentions leave open the question, “Whose intention is this?” To determine who the author of the intention is requires collateral information. Because such information is fallible, misattribution of intentions is possible -- and is in fact often exhibited in people with schizophrenia.

One might think the same would be true for the beliefs, desires, and emotions that provide input to action planning. It is my own mental states that provide input to the forward planning of my own actions, and representations of the other’s mental states that feed into the inverse use of the planning system to explain the other’s behavior. It might be supposed that the system has to distinguish these in some way. But this is not so. Unlike intentions and motor plans, beliefs may remain happily undifferentiated, and failure to differentiate is not only not pathological, it is the norm. What the system needs to “know” is, simply, that there is a puddle in the path; it can deal with undifferentiated, impersonal “facts,” without marking them as facts-to-me, facts-to-you, or facts-to-another -- or, in other words, as facts *as I believe them to be,* or you, or another. Moreover, as will be argued, simple “factive” explanations, such as, “She stepped to the side because there was a puddle in the path,” are the preferred form of action explanation, in contrast to “because she believed ...” explanations. (Use of “because she believed ...” is taken to imply that there was reason not to use the simple factive form.)

Coding for beliefs would *begin* as agent-neutral, in the sense that any differentiation would be the result of intervention of some sort: identical coding for self and other would be the default. Would the same would be true of coding for desires and emotions? Regarding emotions, it is important to distinguish our own emotions from our empathetic responses to another's emotion, and maybe pathological not to. Nevertheless, it is common to think of elements of the environment as disgusting, frightening, and so forth, without specifying "to whom?" Likewise, the world may be seen as motivationally charged, or valenced. Objects may be seen as attractive or repulsive, without an implicit “*to* (somebody).” Even possible future states of the world may be regarded as emotionally and motivationally charged in this non-relative way. This would suggest that undifferentiated agent-neutral coding would present us with a shared world of facts and emotive and motivational valences -- the rich shared world that appears to us, I will argue, as a consequence of maximal code compression.

### Summary so far

The argument so far centers on three claims:

1. Understanding the intentions behind actions is accomplished by inverse planning.
2. The 1p (simulation) version of inverse planning is correct.
3. The default top-down inputs to 1p inverse planning are in agent-neutral coding.

 If A-C are right, then inverse planning gets a free head start, which can then be corrected as needed (Option 3 above).

If C is right, then inverse planning defaults to the greatest possible code compression. In the default condition, inverse planning requires no new input coding to explain the actions of others.

## Projection by default

If we understand simulation in terms of default agent-neutral coding, then we have to reject a well-known account of the simulation theory: that it requires introspective recognition of one’s own (actual or pretend) mental states (metacognition), followed by attribution of the same states to the other individual (Goldman 2006). Agent-neutral coding clearly would support a more economical account of simulation, one that requires neither metacognition nor self-other inference (Gordon 1995; Gordon 1996). It is simply by default that the inputs to inverse planning are the same as the inputs to forward self-planning; This carryover is not established through an inferentialleap from self to other, but rather, as I suggested, simply by omission: that is, crossing the self-other border without doing anything to *alter* the existing inputs.

It is often supposed that attributions of beliefs and desires constitute the very heart of our everyday understanding of the behavior of others. This questionable assumption is built into the use of terms such as *mentalizing*, *theory of mind,* *mindreading*, and *folk psychology.*  These expressions appear to suggest that our understanding of others is based on learning what is going on in their minds, particularly the mental states and processes that cause their behavior. I think this places undue emphasis on attributions to the individual, as opposed to attributions to the situation, or, more broadly, the world. Our everyday effort to make sense of the behavior of others is chiefly an attempt to discern the reasons for their actions; and to discern these reasons is, in general, to know *what it is about the world* that explains their actions. It is not, in general, to discern the *state of mind* behind the action. There are several reasons for asserting the primacy of the world in our understanding of others.

*First,* it makes evolutionary sense that people would prefer explanations of action and emotion that look to the world, rather than to mental states, such as beliefs about the world. It is often useful to identify items in the common environment, especially threats and rewards, and to explain behavior in terms of facts about them. We want to know what it is about the world that is making someone run: perhaps something behind them (from which we should run as well) or something ahead of them (to which we might want to run as well). The parent wants to know what it is about the environment that frightens or upsets the child; and, in social referencing, the child wants to know what the caregiver is responding to, so that it can copy and learn the response.

*Second,* linguistically, reference to the beliefs of the agent is generally treated as a fallback. “Why are you stepping off the path?” Ordinarily we wouldn’t respond, “Because I think (or: believe) there is a puddle.” Rather, we say, “Because there is a puddle.” Likewise, in the third person, “Why did he step off the path?” Mentioning what the agent thought or believed would imply that there is something wrong with a simpler explanation in terms of “the fact” that there is a puddle. For example, we worry that the agent may have been tricked by an illusion. Our explanations, whether of our own actions or the actions of others, default to the factive. (The relevant facts, I should add, may be facts *about* mental states: for example, “I’m calling the dentist because I have an awful toothache,” and, “I’m eating now because I’m hungry [or: bored].”)

*Third*, the most economical strategy for mentalizing, other things being equal, would be one that minimizes individuation, or information tagged to specific individuals. That is, it would minimize the need for explicit mentalizing, in the sense of judgments about mental states or processes. In the default case, with uncorrected agent-neutral coding, the actions of others would be interpreted in terms of a shared world – that is, to the world on the basis of which we ourselves act. Mentalizing, on this account, would be called on to complement or to correct what is passed along through agent-neutral coding. It would be reserved for cases in which a shared world proves inadequate to predict or explain the actions or emotions of particular individuals.

The economizing extends beyond the preference for facts over the beliefs of individuals. It is a feature of our phenomenology that we see objects and events as having, among other properties, *emotive* qualities: the qualities of being scary, repulsive, attractive, embarrassing, shameful, pleasing, and so forth. Such “externalizations” have the virtue of setting expectations and making the corresponding responses by others – being scared, repelled, attracted, embarrassed, etc. -- unsurprising. They can limit the need for coding to the exceptions, the surprising outliers. Similar considerations hold for the affordances of objects, making their standard uses unsurprising and reserving special coding for surprising, nonstandard uses, misuses, and non-uses. Phenomenologically, these properties of objects would be carried over as we slip seamlessly from the forward planning of our own actions to our hypothetical planning of the other's actions. It seems obvious that the more the brain is able to place the causes of actions and emotions in an objective world, the less coding it will need.

### Perspective-taking and positional correction

Probably the most familiar type of correction is spatial perspective-taking. For example, to a stranger observing the scene from a distance, the bear now approaching me is not likely to feel threatening, or in any case as threatening as it does to me. The threatening (or non-threatening) emotive quality of the bear may be seen as a function of one’s location relative to the bear -- or, the bear’s location and vector in egocentric space. With the ability to move mentally into another’s spatial perspective, individual differences become mere positional differences. That is, it is a good starting bet that (unless there is evidence to the contrary) any individual in the same position will see the bear as threatening. With the operation of “putting ourselves in the other’s place” by spatial perspective-taking, we are able to restore the economic advantages of a shared world. We allow the threatening quality to remain out there in the bear, or rather in the bear-from-a-point of view. We need not represent it as a function of individual mental makeup, even if some individuals may be found immune to the standard bear-approaching-me response.

Although it is spatial perspective-taking that gives us the general metaphor of “perspective-taking,” “adopting the other’s point of view,” and “putting ourselves in the other’s place,” many other kinds of corrections may be considered broadly perspectival, or positional. For example, differences in social or occupational role may be bridged by a kind of perspective shift: student/teacher, worker/manager, diner/waiter, patient/doctor, consumer/salesperson. In these cases, as in differences in spatial perspective, it may be sufficient to shift to a generic “point of view,” or, as we say, to understand where the other is “coming from,” to explain the other’s actions, without explicit mentalizing. That is, it may be a good starting assumption that a person in a given “position” will act in more or less the same “standard” way, an assumption that may underlie the notion of generic “scripts” of action sequences postulated by Schank & Abelson (1977). Such an assumption would exploit positional redundancies and limit new input to deviations from the standard.

### Caveats and Qualifications

(1) It seems unlikely that each instance of analysis by inverse planning would start with unrevised agent-neutral coding, and thus uncorrected projection onto the other. With experience and maturity, we should be able to make adjustments before observing the other's actual behavior. We should be able to adjust beforehand to the other's spatial perspective or, more generally, the other's epistemic situation. Likewise, known personal history, social and institutional roles, relationships, culture, may pre-adjust our projections away from uncorrected projection onto the other. Our expectations may conform to templates for particular individuals or classes of individuals, as I explain in a later section; these templates may be shaped by irrational biases as well as by evidence.

(2) Some mental states do not as a rule cross from self to other. Our pains, for example, should be left behind: generally, it would not be a good bet to project onto others the physical pain we feel; likewise, physical pleasure, hunger, and bodily sensations. These, of course, influence our own actions, but they are not generally allowed to motivate others in the same ways. Perhaps these are simply excluded from agent-neutral coding; or, as I think more plausible, infants, in the process of developing self-awareness and self-differentiation, acquire "export prohibitions" that, so to speak, prevent these inputs to action planning at the border between self and other.

## An evolutionary perspective

Agent-neutral coding may be expected to present problems for understanding the behavior of people of very different cultures inhabiting far-away lands. We may seem to share only the bare physical parameters of earthly human life: we breathe, we eat, we sleep, we procreate; the sky is blue (more or less) and grass is green (more or less). Undiscriminating projection onto such people might seem so wide of the mark that, without extensive correction, we could neither predict nor explain much of their behavior.

And yet it is easy to forget that until the very recent past, nearly all human social encounters would have occurred among people in close cultural as well as physical proximity. For much of the history of our species, people would have had little need to depart from the simple strategy of looking to the shared world, with its its facts, emotive qualities, affordances, attractions and repulsions, for the causes of observed behavior. With like-minded individuals in close spatiotemporal proximity, people would have gotten by with few corrections beyond the export prohibitions and positional adjustments such as those mentioned earlier.

The small social groups in which early homo sapiens lived would have shared a local environment and probably formed similar mental maps of that environment. They would have had a shared understanding of the elements of their environment and of the causal properties of these elements, as well as their affordances and emotive qualities. There would have been wide agreement on which elements were salient, menacing, frightening, attractive, or disgusting.

Of course, even in the distant past there would have been cracks in the vault of this shared world. Adjustments would have to be made to differences in temperament, in sensory and cognitive capacities, in knowledge, acculturation, and in goals. Such differences would of course have been salient and noteworthy against the more or less fixed and predictable shared background. But they would have been relatively rare in social groups with strictly limited horizons.

We should note that a social predictive system doesn't just exploit redundancies; it reinforces them and also adds new redundancies. To benefit from the redundancies within our small group, it would help to have grown up within the group. For much of the redundancy within the group is likely to have been a product of earlier corrections. This is especially true of infants and young children, who tend to fill in or replace their own view of the world with those of their adult caregivers. For example, in social referencing. the child observes the adult’s response to *x* (a person, object, or situation), and then copies the response. For example, if the adult appears frightened by *x,* the child will be frightened by *x.* This not only modifies any prior expectation the child may have had of the adult’s response to *x*; it also modifies the child’s future responses to *x*. The child’s response now conforms to the adult’s. This increases redundancy within the group and makes future predictions easier.

### Ignorance and false belief

I argued earlier that our explanations, whether of our own actions or the actions of others, default to the factive. Mentioning what an agent thought or believed would imply that that an explanation in terms of the corresponding “fact” would in some way be defective. For example, we observe someone in broad daylight walking nonchalantly into a deep puddle (or: a lamppost). Why? What accounts for his aberrant behavior? Answer: he was looking at his cell phone, oblivious. We could have predicted it, and now we can explain it.

We make his behavior unsurprising by disconnecting or “decoupling” the fact that there was a puddle in his path from the input to inverse planning. Decoupling a fact from inverse planning is a way of marking ignorance of fact. Ignorance in turn may engender false belief: because he was ignorant of the fact that there is a puddle – out of touch with the facts concerning his current environment -- he continued operating on the false default assumption of a puddle-free path.

In the classic false belief experiments (Wimmer & Perner, 1983), the task is one of *predicting* behavior, rather than interpreting or explaining a given action. One’s action planning system may be recruited for this task, but it would be through a vicarious *forward* use of the system, rather than an inverse use. As before, you would decouple the system from the events that transpired in the other’s absence. Absence from a scene creates a blind patch, a scotoma of ignorance. Given ignorance, the forward use of the planning system would predict that other would not do the well-informed or “correct” thing. (See also Perner & Roessler 2010)

Agent-neutral coding and the possibility of toggling between knowledge and ignorance would give us the neural underpinnings for two theses long held by Josef Perner: first, that well before they have an explicit grasp of belief attribution, young children are quite capable of explaining action in terms of the external situation; and second, that older children and adults use the same type of explanation young children use, except in the occasional cases where it proves inadequate; then they must fall back on explanations that mention the mental states, especially the beliefs, of the agent. Young children and, where possible, older children and adults

make sense of intentional actions in terms of justifying reasons provided by "worldly" facts (not by mental states). (Perner & Roessler 2010)

The young child's conception is all we usually call upon, because it is typically all we need. This comes to saying that explaining and predicting actions in terms of actual situations or facts is our default mode of explanation and prediction, the mode we employ unless we find some reason not to. Only where this appears inadequate do we invoke beliefs in our explanation.

## Anchoring and adjustment

The projection-and-correction understanding of simulation bears a close kinship to recent work based on the "anchoring and adjustment heuristic" originally proposed by Tversky & Kahneman (1974).(See Epley et al 2004, Tamir & Mitchell 2010, Tamir & Mitchell 2010). The idea is succinctly expressed in Epley et al (2004):

people adopt others' perspectives by initially anchoring on their own perspective and only subsequently, serially, and effortfully accounting for differences between themselves and others until a plausible estimate is reached.

The serial adjustments are conceived as a process of moving out from one's own -perspective, as necessary, to more "distant" perspectives, with hypothesis-testing at each stop along the way - a process that should take longer, the farther out one goes.

Tamir and Mitchell 2010 cites evidence that

the MPFC [medial prefrontal cortex] subserves the use of self-projective simulation as one route to understanding other minds.

They suggest that

subregions of the MPFC not only may use the self as an anchor point from which to understand others but also may actively allow perceivers to adjust their inferences about another person.

That is, the MPFC may make possible, not only self-projective simulation, but also the corrections or adjustments that enable perceivers to accommodate minds different from their own.

According to the view developed in this paper, the default to an "anchor" is a consequence of predictive coding at the neural level, specifically, agent-neutral coding in self-planning and inverse other-planning. Indeed, Koster-Hale & Saxe 2013 finds that anchoring-and-adjustment to be implemented by predictive coding in the medial prefrontal cortex (MPFC):

When specific information about a person's reputation or traits is unavailable, we may predict others' preferences by assuming that they will share our own preferences (Krueger and Clement, 1994; Ross et al., 1977). In one series of studies (Tamir and Mitchell, 2010), participants judged the likely preferences of strangers (e.g., is this person likely to ''fear speaking in public'' or ''enjoy winter sports''?) about whom they had almost no background information. Under those circumstances, the response of the MPFC was predicted by the discrepancy between the attributions to the target and the participant's own preference for the same items: the more another person was perceived as different from the self, for a specific item, the larger the response in MPFC.

This suggests that in at least one area of mentalizing, the default is uncorrected projection onto a target, with adjustments in response to contrary information. Essentially, subjects responds initially as if the question were about themselves, and then make adjustments for differences.

The only caveat I would raise concerns the explanation of the MPFC responses: that they reflect the subject’s “assuming that they will share our own preferences.” This imposes on the data the unwarranted assumption that the default to an anchor is an optional belief-based heuristic, a shortcut we use because it seems like a good idea. If the argument of this chapter is correct, the default to an anchor is a structural feature of 1p inverse planning, a consequence of doing nothing to alter the input to self-planning -- including, of course, our preferences.

There is evidence that the adjustments made in these anchoring-and-adjustment experiments do not consist in theorizing about personalities different from one’s own, but rather in impersonating them. The subject becomes a shapeshifter, modifying her brain to respond in character. The amygdala, in particular, is an active participant in this impersonation. It plays a role when, as in several of the anchoring and adjustment studies, we predict the responses of a (hypothetical) person with an emotional disposition that differs from our own (Gilead et al 2016). It is also activated when we respond empathetically to another’s emotional suffering (Bruneau et al 2015). I think it is reasonable to speculate that similar changes occur when we try to anticipate the behavior of familiar individuals or try to understand the intentions behind their actions. Our actual mentalizing in such cases might begin with a priori templates, comparable to the ready-to-use set of transformations of an actress getting into an accustomed role: her Lady Macbeth template, her Blanche DuBois template, and so forth. For many actors and actresses such transformations become instantaneous and automatic. Something comparable may be true of our simulation of particular individuals or classes of individuals. We project with a fixed set of adjustments, and these become *a priori* expectations.

## Conclusion

This chapter develops the idea of simulation as a predictive strategy for mentalizing. The predictive aspect consists in an initial projection onto the other, which is then corrected and revised as needed. Such a projection is implemented by allowing our own mental states to govern, not just our own behavior, but also our hypothetical interpretations of the observed behavior of other agents. The mechanism proposed for this is a variant of the inverse planning mechanism put forward by Baker, Saxe and Tenenbaum (2009). According to what I call *first person* (or *1p*) inverse planning, our own action planning system is recruited as a hypothesis-testing device. In the default condition, the top-down inputs to interpreting the behavior of other agents would simply be the same as the inputs to planning our own actions. This agent-neutral coding would be modified as needed to generate intentional actions that come close enough to matching bottom-up perceptual-motor input from observing the behavior of others. In creating and evaluating alternative reconstructions of the processes that gave rise to the other’s behavior, the system performs an analysis by synthesis, like the inverse graphics account of visual perception.

In reconstructing the processes behind the other’s action, inverse planning locates the agent’s reason or reasons for acting, as far as possible, within a shared world of facts, affordances, emotive and motivational valences, and other “objective” properties. Where this is problematic, as in the case of ignorance, false belief, and non-standard emotional responses, inverse planning locates the causes within the mental states of the individual agent. Shared world explanations have a number of advantages over those requiring explicit mentalizing: they can identify environmental threats and rewards, they are conceptually and linguistically less demanding, and they achieve greater code compression. If this is correct, then we must reject the common assumption that explicit mentalizing, or mental state attribution, is the paramount explanatory aim of the procedures we lump under the term *mentalizing.* The aim is rather to interpret behavior in terms of a shared world where this is possible and to diagnose cases where it is not.

The notion of a shared world may seem quaint today, when few people belong to small, culturally and geographically isolated groups. Living successfully in the world of social media, especially, would seem to demand constant sophisticated and resource-hungry mentalizing to accommodate disparate voices. And yet, notoriously, we seem to manage by stepping into echo chambers of like-minded individuals creating the semblance of a shared world. These “pocket worlds” may reflect a more general human tendency to form limited “social niches” with mutual expectations based on shared “cultural affordances” constituting a shared world. (Veissiere, Constant, Ramstead, & Friston 2019).

###  The shared world comes first

I said at the beginning that the simulation theory stipulates that the main source of everyday human competence in predicting and explaining human behavior consists in a certain procedure or set of procedures. As commonly understood, a simulation procedure allows us to put ourselves in the other’s place when we are not, in the relevant sense, already in the other’s place.[[4]](#footnote-4) The term is also used more broadly to cover the inverse use of the action planning system for testing hypothetical explanations of another's action (or hypothetico-practical inference).

However, in the light of what I have argued here, I would have to say that a key contributor to our capacity for explaining and predicting action is not a procedure at all. It is simply the carryover of top-down inputs as our mechanisms for decision-making and planning switch to inverse use for testing hypothetical explanations of the actions of others. This identity, or agent-neutral coding, of inputs has the effect of projecting onto others a shared world within which we act and interact. Strictly speaking, the brain doesn’t have to do anything to accomplish this; rather, what is required is inaction: simply leaving the inputs unchanged as it switches from self to other. The shared world is the default.

The fact that such a system of action interpretation, starting with no additional expenditure of resources and adding only as needed, would be as cost-effective as a system can be, is reason for confidence that it is the system we have.

## References

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. In Advances in neural information processing systems (Vol. 18, pp. 99–106).

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

Baker, C.L., Saxe, R.R., and Tenenbaum, J.B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In Proceedings of the thirty-second annual conference of the cognitive science society (pp. 2469–2474).

 Baker, C.L., Jara-Ettinger, J., Saxe, R., and Tenenbaum , J.B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour 1: 1-10.

Blakemore, S-J, and Frith, C. (2005) The role of motor contagion in the prediction of action. Neuropsychologia 43: 260–267

Bouquet, C.A, Shipley, T.F., Capa, R.L., & Marshall, P.J. (2011). Motor contagion: goal-directed actions are more contagious than non-goal-directed actions. Exp Psychol. 58(1): 71-8

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36, 181–253.

Decety, J. and Sommerville, J.A. (2003). Shared representations between self and other: a social cognitive neuroscience view. Trends in Cognitive Sciences 7, 527-533.

Friston, K. J. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B-Biological Sciences, 360 (1456), 815-836.

Goldman, A. I. (1989). Interpretation psychologized. Mind & Language, 4, 161–185.

\_\_\_ (2006). Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford: Oxford University Press.

Gordon, R.M. (1986) "Folk Psychology as Simulation," Mind and Language, v. 1 (pp. 158-171. Reprinted in M. Davies & T. Stone, eds., Mental Simulation: Evaluations and Applications. Oxford: Blackwell, 1995). Reprinted in W. Lycan, ed., Mind and Cognition: An Anthology, 2nd edition (Blackwell Publishing, 1998)

\_\_\_ (1992). Reply to Stich and Nichols. Mind and Language 7, 1992; reprinted in M. Davies & T. Stone, eds., Folk Psychology: The Theory of Mind Debate. Oxford: Blackwell, 1995

\_\_\_ (1995). Simulation Without Introspection or Inference From Me to You. In M. Davies & T. Stone, eds., Mental Simulation: Evaluations and Applications. Oxford: Blackwell

\_\_\_ (1996). Radical' Simulationism. In P. Carruthers and P. Smith, eds., Theories of Theories of Mind. Cambridge: Cambridge University Press.

Heal, J. (1986). Replication and functionalism. In J. Butterfield (ed.), Language, Mind and Logic. Cambridge: Cambridge University Press, 135-50.

Helmholtz, H. v. (1856). Treatise of physiological optics: Concerning the perceptions in general. Classics in psychology, 79-127.

Jara-Ettinger, J. (2019) Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences, 29:105–110

Jeannerod, M., and Pacherie, E. (2004). Agency, Simulation and Self-identification. Mind & Language, 19: 113–146.

Koster-Hale, J., and Saxe, R. (2013). Theory of mind: a neural prediction problem. Neuron 79: 836-848.

Perner, J., and Roessler, J. (2010). Teleology and causal reasoning in children's theory of mind. In J. Aguilar & A. Buckareff (Eds.), Causing human action: New perspectives on the causal theory of action (pp. 199-228). Cambridge, MA: Bradford Book, The MIT Press.

Perry, J. (1979). The problem of the essential indexical. Nous 13: 3-12.

Schank, Rodger C.; Abelson, Robert P. (1977). [Scripts, Plans, Goals and Understanding](https://content.taylorfrancis.com/books/download?dac=C2007-0-06021-9&isbn=9781134919666&format=googlePreviewPdf). Hillsdale, NJ: Erlbaum.

Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. Science 185:1124–1131.

Veissière SPL, Constant A, Ramstead MJD, Friston KJ, Kirmayer LJ. (2019) Thinking Through Other Minds: A Variational Approach to Cognition and Culture. Behav Brain Sci. May 30:1-97.

Wimmer, H. and J. Perner 1983. “Beliefs About Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children’s Understanding of Deception,” Cognition, 13, 103-128

1. To the Freudians, it is a defence mechanism by which one deviously assigns to someone else a mental state or trait one can't accept in oneself. [↑](#footnote-ref-1)
2. The temporal terms “initial,” and “starting point” should not be taken to imply that each instance of analysis by inverse planning begins with uncorrected agent-neutral coding. [↑](#footnote-ref-2)
3. Joëlle Proust and Shaun Gallagher independently called my attention to similarities between my agent-neutral coding and the agent-neutrality of intentions posited by Jeannerod and Pacherie (2004). [↑](#footnote-ref-3)
4. Daniel Dennett once pointed out to me that if simulation were our ordinary way of understanding others, we wouldn’t have to *tell* people to put themselves in the other’s place. He was right, regarding “putting in place” simulation. [↑](#footnote-ref-4)