# Notes on Group Agency[1]

Brian Epstein
Tufts University, Medford

Michael D. Ryall
University of Toronto

July 9, 2021

[1]

Individuals and organizations alike navigate the world and accomplish their ends. Like individuals, organizations can be regarded as agents, possessing conceptual systems that help them accomplish their ends over time as they move through a changing environment, and interact, cooperate, and compete with other agents along the way. Individuals and organizations are also bounded in a variety of ways: bounded in the information they have about the world, bounded in their ability to calculate, bounded in their memories, bounded in the amount of time they have to make decisions and think things through.

In certain contexts, it is useful to think of agents as ideal, or at least as designed in the image of a perfect and unbounded reasoner, with simple and general-purpose cognitive systems. Game theory models situations in which a set of forward-looking agents make interactive decisions. The building blocks of game theory are general and parsimonious. They enable game theory to model a broad range of phenomena with elegant mathematical structures.

Yet to a certain extent, it makes sense that nature or human design has set up these cognitive systems specifically so as to be effective in commonly encountered environments despite being severely bounded.

While simple and general models of cognition–such as those that models agents as bundles of beliefs and desires ranging over possible states of the world–have been enormously fruitful, many theorists in recent years have worked to modify these models to accommodate human bounds. It is clear that there are lots of ways for systems to navigate environments, and to do practical reasoning in the service of ends. A parsimonious model is beneficial for certain purposes, but there is reason to broaden the way we think about the components of cognitive systems.

Many ways of accommodating bounds involve changing features of the cognitive systems, such as modifying the calculation algorithm (as with different forms of discounting or modifying the sort of probabilistic reasoning), or else by introducing different systems or cognitive pathways, such as the automatic or rule-of-thumb systems discussed by dual-process theorists. A different way of thinking about modeling cognition involves modifying the repertoire of cognitive mechanisms employed by a cognitive system.

Our aim in this paper is to contribute to this work by bringing together recent work in game theory and the philosophy of action. Both fields have sophisticated ways for thinking about forward-thinking agents, reasoning about what to do at a given time based on their aims and beliefs regarding the future. But there has been little interchange on this topic between the fields. The challenges for

bounded systems are particularly prominent when we consider their operation through time, when they need to be forward-looking, and also when they live in dynamically changing environments. We regard this as a general feature of real-world contexts which cognitive systems navigate. We aim to build on the insights of both fields to develop models of bounded agents acting with an eye to the future.

One feature that is present in the lion's share of game theory work is its implicit belief-desire approach to modeling mental states. The mental attitudes agents possess, in these models, are bundles of beliefs and desires that range over possible states of the world. Because those attitudes can range over states into the indefinite future, for sets of agents of any size, they provide a natural framework for thinking about diachronic action.

The belief-desire approach to cognition, however, is widely challenged in action theory. Many philosophers argue that more elaborate models of cognition and action are essential for explaining the practical reasoning and behaviors of individuals. Intention is the most widely discussed mental state argued to contribute to practical activity of agents. (fn: Intention is only one of many, including plans, reliance and so on.) Some of this work has been incorporated in computational models of interacting agents, but little has been adopted into game theory proper.

and because the models are designed to

Game theoretic models

have proven enormous has been difficult–in part because of the elegance and fecundity of simple and general models that regard agents as bundles of beliefs and desires that range over possible states of the world–to incorporate Models of cognition that date back to the enlightenment and that are still most widely used in modeling reasoning and tell me in such does economics tend to favor simple and powerful general purpose systems, and when bounds or limitations are encountered the general purpose system is provided with an adaptation or mechanism for taking account of that specific limitation.

The basic idea is to contrast what we're doing with commitment with the way other people have been thinking about commitment – e.g., strategic precommitment (like to prevent entry, mutually assured destruction, etc.); precommitment as a counter to impulsivity or weakness of will, and commitment to a strategy (like tit for tat).

Basically the idea is just to say a bit about these so as to help communicate how we're regarding agent commitment as a way of increasing payoffs for certain kinds of bounded agents. (It also will

be helpful I think to contrast that a little bit with how people think about reasoning for bounded agents in more typical ways, that is, as heuristics etc. for modifying choice approaches, as opposed to diachronic commitment.

In social ontology, even reductive accounts of group agency are typically built from individuals with a wide range of cognitive states, not just beliefs and desires. And both reductive and non-reductive accounts theorize about features of groups that appear to be distinct from features that aggregates of coordinating agents would possess—features such as group-level attitudes.

The project of which this paper is a part aims to approach game theory with a more elaborate ontology of mental states, informed by work in action theory and social ontology. Rather than reducing multi-scale social phenomena to groups of coordinating agents or representing the epistemic limitations via ad-hoc modifications to a belief-desire model, we accept a richer set of building blocks at the outset. With these, we treat the cognitive lives of limited individuals and groups in greater detail, and hope to provide more systematic explanations of the sort of game theoretic challenges described above. We also clarify and correct claims in action theory about relations among cognitive states of individuals, as well as between individuals and social groups. In this paper, we sketch the foundations of a mathematical framework in which intentions perform two broad functions. First, as is widely discussed, intentions improve the efficiency of agents with epistemic limitations. Second, individual intentions perform a social function: pairs or groups of agents who possess individual intentions may induce behaviors that are unavailable to belief-desire agents. We also begin to extend this to a formal theory of groups, including group formation and persistence, as well as aspects of group cognition. Our framework lends itself to addressing issues that the formal literature on agency (such as the BDI literature in logic and computer science) has not been able to treat, such as plan revision in a social context and in response to the plans of others.

There are two desiderata that we are attempting to meet: to retain mathematical tractability and certainly to take advantage of the ability of game-theory to be forward-looking; and yet to add some cognitive complexity to the model to account for realistic adaptations built into human cognitive systems that allow us to be more successful in a general purpose way when operating as bounded individuals or groups.

We are aiming to balance these considerations by adapting a modified cognitive system, incorporating awareness and unawareness. This machinery recently developed gives us resources to

augment the cognitive mechanisms, and yet already has a reasonably well explored mathematical structure. There is one prominent difference between the way awareness structures are typically deployed and how we will treat them: usually the lack of awareness reflects a deficiency in an agent, and the agent who becomes aware of some feature of the world is better positioned to act than one who is unaware. We, however, will argue for the utility not of awareness but of unawareness. We will model intention in particular in terms of agents fruitfully becoming unaware of things that they had previously been aware of. We will model the focusing of attention, and the taking on of commitments, as involving a kind of "intentional unawareness" which can enhance outcomes for a bounded agent.

The paper consists of the following sections:

1. Some motivating examples in which we may be more successful by engaging certain cognitive mechanisms, and in particular by limiting our attention
2. Setting up a particular toy case (literally a toy case!)
3. Introducing the awareness framework and setting up the machinery for awareness; introducing the idea of "intentional unawareness"; the machinery in the formal case
4. How the awareness framework helps in the toy case
5. Formalizing awareness in the dynamic case
6. More general treatment of belief and intention; an overall model
7. Application to some other cases
8. Suggesting how this can be applied to groups

# 1 Motivating examples

[[These need to be thought through and filled out. Problems that are difficult to model using traditional methods (even if they can be modeled traditionally).]]

a. Dynamically changing environments; where commitment is necessary

b. Organizations, where dissemination and coordinating people to common action is more important than optimizing the decision (This comes out as a long lag between decision and implementation, in a changing environment)

c. Long term projects for individual lives

The claim is not that current models are incapable of addressing these. But rather, that there are unifying features of cognitive systems that allow them to be understood in a more systematic way.

## 2 Toy case setup

We consider the problem of Brian's Toddler, whom we refer to as individual $i$. At time $t = 0$, the $i$ finds herself in a room with two toys, $A$ and $B$, situated in front of her. At this point, $i$ needs to figure out what to do. Her goal is to obtain the best toy.

The problem she encounters is this: the toys in the playroom are, annoyingly, designed to attract the toddler's attention. So even when she is not playing with them, they ally ring or light up or say things to the toddler, drawing her attention. Her preferences change over time, but they are not inconsistent: she favors the toy that is making the greatest clamor at any given time. The problem is that it takes time to crawl over and pick up any given toy. Even as she is crawling toward one toy, another will do something to attract her.

At 10 months, the toddler crawls in circles, chasing the toy that is the latest to ring, only to change direction while she is on her way, and thus crawls around and around the playroom, never reaching a toy.

At 12 months, things have changed. Despite the dynamic clamor in the playroom, something has changed in the toddler's cognition. Somehow she fixes on a particular toy, she heads in that direction and gets it, even though other toys are ringing around her.

We will consider a parsimonious game theoretic treatment of this latter situation, and then a different treatment using awareness structures.

### 2.1 Parsimonious game theoretic treatment

The parsimonious game theoretic treatment is shown in Fig. 1. In this diagram, the toddler successfully chooses one toy over the other by being aware of everything, including the consequences of her actions into the future. With beliefs about the appeal of each of the toys into the future, she is able to form a belief as to which toy is likely to be best, and to head to that toy even when the other toy lights up.

The uncertainty of individual 1 with respect to which toy is actually best is represented by including an initial move by Nature at time $t = 1$ – i.e., the two possible states are A-Best and B-Best, one of which is true and the other of which is counterfactual. The bold lines indicate the choices of the players. Individual 1 is then faced with a choice as to whether to get A or get B. As illustrated by the dashed line connecting 1's decision nodes, 1 does not know with certainty which is the true state but, as indicated, believes it is most likely that A is best. Therefore, 1 chooses to get A in $t = 2$. As a result, 1 obtains A in $t = 3$.

The features to note are: 1) the world simply presents 1 with a decision; 2) although 1 is uncertain about which toy is best, she is aware of the counterfactual possibilities – indeed, 1 knows everything about the game, including what will happen as a consequence of her actions; 3) all of 1's cognitive processes associated with the decision are compressed into the act of making a decision. The decision could be elaborated as one involving probabilistic beliefs on the part of 1, but this is not necessary. For whatever reason, at the time of her decision, 1 believes (with some measure of uncertainty) that A is best. Given these beliefs, and a desire for possessing the best toy, 1 chooses to get A.

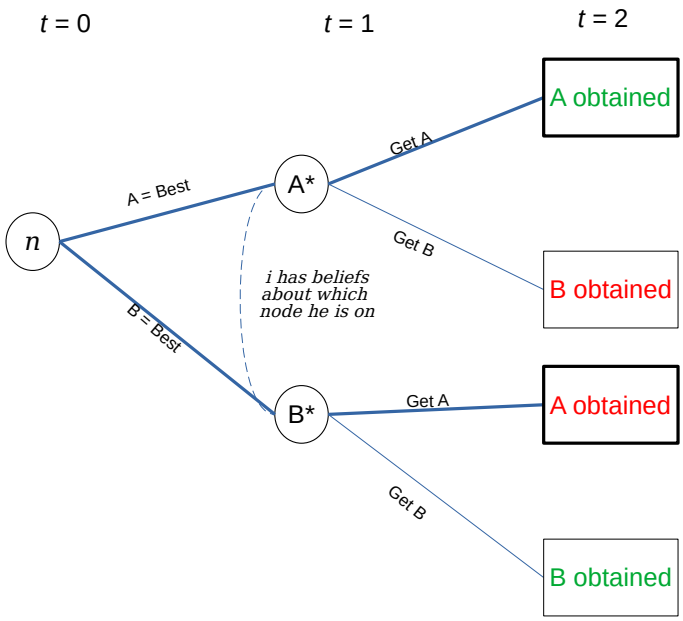


Figure 1: Brian's Toddler, parsimonious game-theoretic treatment

## 2.2 Boundedness and decision paralysis

While the parsimonious game theoretic treatment correctly arrives at the outcome that the toddler chooses a toy, it has three serious shortcomings. First, as is often observed, it is unrealistically idealized. The toddler does not have the information that the parsimonious treatment ascribes to her, nor is it plausible that she projects forward in the way it models her.

Second, while the model does have her choose a toy, it is not at all clear that the choice behavior has her choose one as "best" given the information she has, as opposed, say, to her choosing the one that is momentarily the most salient.

Third, it seems to misdiagnose what is going on in her transition from decision paralysis at 10 months to the successful choice at 12 months. She may be better at assessing consequences than she was two months earlier. Still, it does not seem likely that her "phase transition" in choice-making is a product of her suddenly having become an assessor of consequences, or of her having acquired calm acceptance of the fact that either toy is better than none, so that satisficing is suddenly adequate. Instead, toddlers are characteristically willful, arbitrary, and tyrannical.

We suggest that we should not model the success of the 12-month-old's choice via a decrease in her boundedness as a rational agent. But instead, that we model her as being able to activate or make use of a cognitive mechanism that generates resolution even in her dynamic environment, while she remains subject to roughly the same bounds as before.

We should clarify that this remains a "toy" example: though we are discussing a hypothetical transition for a toddler, we are not making empirical claims in developmental or cognitive psychology. Rather, this is meant to illustrate how certain cognitive mechanisms can serve as general purpose solutions to problems that even severely bounded agents face. These mechanisms continue to be of use even for less bounded agents, and when we turn to organizations, we will find that while in certain ways they are forward-thinking and powerful reasoners, they may in other ways have even more limitations as rational actors than toddlers do.

## 2.3 Presenting intentional unawareness via the toy example

In the next sections, we introduce the notation and machinery of awareness structures, using this toy example as a running illustration. For the basic machinery of awareness, we follow [[Schipper and others]].

We begin by introducing the machinery of awareness, following [[Schipper and others]]. We

present the mathematical formalism, and contrast our use of awareness with prevailing uses in the literature.

The initial formalism we present is synchronic: we start with the machinery of agents that differ in their awareness of states at a moment in time. We go on to formalize the toy example with this machinery, and show that we also need to add a dynamic feature to the treatment of awareness, in order to capture how choices are made in the described changing environment.

Subsequently, we discuss other less toy cases, and formalize dynamic awareness structures.

# 3 Modeling awareness

## 3.1 Notational conventions

### 3.1.1 General

Capital letters ($G$, $N$, etc.) refer to sets. Small Arabic and Greek letters refer variously to elements of sets (e.g., $i \in N$) and functions (e.g., $\sigma : N \to \mathcal{N}$). Terms are *italicized* at the point of definition. A *profile* is a placeholder for a list of elements. We denote these in boldface: e.g., $\mathbf{x}$ where $\mathbf{x} \equiv (x_1, \ldots, x_n)$. The "$\equiv$" symbol indicates the definition of a mathematical object. If $X$ is a set, then $2^X$ denotes the set of all subsets of $X$. Calligraphic letters refer to sets of sets (e.g. $\mathcal{X} \equiv 2^X$). Curly parentheses indicate sets, typically in defining them (e.g. $X \equiv \{x | x \text{ is an even integer}\}$). The notation "$|\cdot|$" indicates set cardinality (e.g., if $X \equiv \{a, b, c\}$, then $|X| = 3$). If $X$ is a set and $Y \subset X$, then $X \setminus Y$ is the set $X$ minus $Y$; i.e., the set of elements of $X$ that remain when the elements of $Y$ are removed. All sets are assumed to be finite unless otherwise indicated.[1]

[1] In almost all cases, our results extend to uncountably infinite sets (e.g., the domains and ranges of continuous variables). However, extending the analysis to include these would involve bulking up the discussion with technical material that would add little, if anything, to the conceptual content of the paper.

### 3.1.2 Notation Reference

Table 1 elaborates all the mathematical objects used in the paper.

| Object | Description | Comments |
|---|---|---|
| $N \equiv \{1, \ldots, n\}$ | The set of $n$ individuals | |
| $i \in N$ | An arbitrary individual | $i = 0$ is Nature |
| $S^i$ | States of the world of which $i \in N$ is aware | $S^0$ contains all possible states |
| $s \in S^i$ | An arbitrary state | |
| $S^\emptyset \equiv \{\emptyset\}$ | Space representing complete unawareness | |
| $\mathcal{S} \equiv \{S^\emptyset, S^0, \ldots, S^n\}$ | Lattice of awareness spaces | Maximum is $S^0$ and minimum is $S^\emptyset$ |
| $S^i \succeq S^j$ | Individual $i$ is at least as aware as $j$ | |
| $\Sigma \equiv \bigcup_{i \in N} S^i$ | Union of the individual spaces | |
| $r^{i \to j}(s^i)$ | Impoverished version of $s^i$ perceived by $j$ | Only defined if $S^i \succeq S^j$ |
| $B^\downarrow$ | A synchronic event $B^\downarrow \subseteq \Sigma$ | $B^\downarrow = \bigcup_{S^j \in \mathcal{S}} \left(r^{0 \to j}\right)^{-1}(B)$, $B \subseteq S^0$ |
| $A^i(s)$ | $i$'s feasible acts in state $s \in S^0$ | |
| $a^i \in A^i(s)$ | An arbitrary feasible act of $i$ | |
| $\mathbf{a}$ | A profile of acts, $\mathbf{a} \equiv (a^0, \ldots, a^n)$ | |
| $\mathbf{A}(s)$ | All act profiles in $s$ | $\mathbf{A}(s) \equiv \times_{i=0}^n A^i(s)$ |
| $\mathbf{A}_t$ | All act profiles at time $t$ | $\mathbf{A}_t \equiv \cup_{s \in S_t^0} \mathbf{A}(s)$ |
| $\mathbf{A}$ | All possible act profiles | $\mathbf{A} \equiv \cup_{s \in S^0} \mathbf{A}(s)$ |
| $\omega(\mathbf{a}_t, s_t^0)$ | State actualized from $s_t^0$ given $\mathbf{a}_t$ | e.g., $s_{t+1}^0 = \omega(\mathbf{a}_t, s_t^0)$ |
| $\mathbf{h}(s_t^0)$ | History at $s_t^0$ | a profile, $\mathbf{h}(s_t^0) = (s_0^0 \ldots, s_t^0)$ |
| $\mathbf{H}_t$ | Set of all feasible histories at $t$ | |
| $\mathbf{h}_t \in \mathbf{H}_t$ | An arbitrary feasible history at $t$ | |
| $\mathbf{H}_T$ | The set of all histories through $T$ | |
| $\mathcal{H}_t$ | Set of all subsets of histories at $t$ | |

Table 1: Notation Reference (WILL NEED TO BE UPDATED)

## 3.2 Awareness in game theoretic models

### 3.2.1 The idea of awareness

The idea of the awareness formalism is to address an important limitation of the standard models of how agent attitudes relate to the world, in both game theory and formal epistemology. Standard models regard agents as having attitudes toward all states of the world. Regarding any particular state of the world, agents may not know whether or not it obtains, but attitudes are nonetheless defined over every state.

Unawareness has long been a tricky problem for decision theorists. A decision maker can only choose between acts of which he or she is aware which, typically, does not include all the truly feasible acts at that moment. Moreover, the decision problem is further compounded by unawareness of future possibilities associated with one's acts. It is easy enough to represent a static decision problem which is constrained by the decision maker's awareness of possible acts by simply defining the "feasible" acts as those corresponding to his or her awareness. The problem is how to model what happens in a dynamic setting in which the decision maker suddenly faces an unexpected consequence. For example, in a standard Bayesian decision problem, unawareness of certain consequences can be modeled as zero-probability states according to the decision maker's subjective beliefs. However, such decision makers will be confounded should a subjectively impossible state occur. Added to this is the problem of representing decision makers of differing awareness when decision problems are interactive.

**?** demonstrate that standard state-space approaches cannot model unawareness. **?** surveys various alternatives to modeling unawareness, including approaches from AI, logic, and game theory. We adopt a version of the framework used in **?** which itself builds on previous work developed in **?** (also see **??**, for related extensions). This approach solves the problems mentioned above by creating multiple state spaces, each one associated with the awareness of a particular individual. This allows different agents to have different perceptions of the the true state of the world as well as the future states that might obtain in the future.

### 3.2.2 Our use of awareness

The literature on awareness generally addresses the shortcomings of decision-making under conditions of unawareness. For instance, unawareness structures are helpful for modeling contracts

written by parties who differ in their awareness of features of the world that affect possible contract outcomes.

We will be using the formalism for a somewhat different purpose. In a sense, we flip the consequences of unawareness, showing that unawareness can be a virtue for bounded agents.

In particular, we proceed from the uncontroversial claim that, at any given moment, an individual can only attend to some finite number of conscious concerns. We say that an individual is *aware* of the matters toward which his or her attention is directed. The idea is as follows. To the extent some share of the mind's resources are occupied in solving a problem (e.g., deciding what kind of car to buy), those resources are not available for other conscious operations, such as solving other problems, constructing a feasible plan by which to acquire a car, or actualizing that plan by driving to the car dealer and making the transaction. We conjecture that an individual's finite stock of cognitive resources almost always acts as a hard constraint on his or her decision- and act-making capability.

Under these ordinary conditions, it can be a cognitive virtue to be selectively unaware even of states of the world that might be pertinent to ones plans. A cognizer with some sort of mechanism for triggering unawareness may be more successful than one without that mechanism.

Many philosophers of action treat intention as a distinct cognitive attitude, not reducible to beliefs and desires. Intention has long been recognized as playing a key role in distinguishing actions from mere behaviors (contrast a twitch of the eye with a wink), and in recent years a good deal of work has been done on the role of intention in future-directed action and planning. **?** hypothesizes that the cognitive role of intention is tied to the boundedness of of cognizers.

We aim to models agents with intentions, as distinct from beliefs and desires, using the awareness model. We argue that under constrained Under constrained awareness, intentions take on an important role that is distinct from beliefs and desires.

In our model, intentions serve as the pivot from goal choice assessment to goal acquisition planning and implementation. The formation of an intention moves an individual from a state of reckoning about what goal to pursue to one in which that choice becomes a commitment accompanied by *plan* by which to attain it.[2] Thus, forming an intention frees up the mental resources required to determine which goal to pursue and how to pursue it. When events arise consistent

[2] A more elaborate treatment might separate each step by an act of intention: first, the move from goal assessment to plan selection; then, from plan selection to plan implementation, etc. For now, we bundle these steps into one.

with the plan, the individual can proceed accordingly – without engaging the mental machinery required to reassess goals and plans. Because deciding to focus attention on some new problem can, itself, be an intentional goal, one's awareness is dynamic and, to some extent, influenced by one's own intentions. As we will see, there are also social implications as individuals become aware of the intentions of others.

Beliefs and desires will operate in a familiar way. The distinction here is that they are restricted to those matters about which an individual is aware. As we show below, because beliefs cannot account for awareness and because intentions shift awareness, a belief-desire model cannot do the work of an awareness-belief-desire-intention model.

## 3.3 Formalizing synchronic awareness

We break this section into two subsections. The first develops the mathematical machinery to discuss and analyze actual and potential states of the world at a moment in time, elaborated in their fullest detail. The second extends this basic setup to account for each individual's awareness of these fully elaborated states.

### 3.3.1 The individuals

Begin with a *population of individuals*, indexed by the set $N \equiv \{0, \ldots, n\}$ with typical element $i \in N$. For now, we focus on an individual actor. Later, we consider groups. The evolution of the world through time is driven by the actions of individuals as well as of the onset of natural phenomena. We account for natural phenomena as the "actions" of Nature which we assign to population index 0.

### 3.3.2 Nature: all possible states in all detail

A *state*, denoted $s$, is a snapshot of the world at a moment in time. Each state represents a set of possible features of the world at a moment. This includes not only "mind-independent" features of the world, but also the attitudes of individuals in the world as well as features of the world that involve those attitudes. Plausibly, it would require an uncountably infinite number of states to elaborate everything about the actual world in a given moment (much less all the potential features that *could* be actualized). However, our discussion will always focus upon a finite set of individuals who are interacting within some specific domain of interest.

With this in mind, let $S^0$ denote the set of all *possible* states of the world. The "0" superscript indicates that we might think of this as the state descriptions of all of reality, not restricted to the subsets of which agents may be aware. Any given $s \in S^0$, that is, is a complete description of the world as it could possibly be.

### 3.3.3 Individual awareness structures

There are two conditions that must be met for an individual to be aware of some feature of the world. First, the feature must be accessible to the individual for active consideration. The sources of accessible features include contemporaneous sense data, active imagination, and knowledge – essentially, anything an individual can call to mind. Second, the feature must be actively brought to mind. For example, an airline pilot may be able to call to mind how to navigate a jetliner but not a container ship. That same pilot may not be actively considering how to navigate a jetliner while driving his car down the freeway. We cannot bring to mind things we do not know or cannot imagine. Of the things we know or can imagine, we are constrained in the collection to which we can actively attend.

The set of states that are discernable to individual $i$ given her awareness of reality is denoted $S^i$. Conceptually, $s^i \in S^i$ includes all the features of reality that individual $i$ can bring to mind and discuss should it be actualized. Given the information encoded in a state, individuals may also be aware of what they know, what they believe, what they intend, and so on. Importantly, awareness may extend to the mental states of others. Assume $\mathcal{S} \equiv \{S^\emptyset, S^0, S^1, \ldots, S^n\}$ along with $\succeq$, a partial order on $\mathcal{S}$, is a complete lattice in which $S^0$ is the maximum (a complete expression of reality) and $S^\emptyset$ is the minimum. $S^\emptyset \equiv \{s^\emptyset\}$ is the space consisting of a single, "null" element in which nothing about the world is distinguished. Then, $S^i \succeq S^j$ means that individual $i$ is able to distinguish at least as much about the world as individual $j$ in that moment. Because $\succeq$ is a partial order, not all state spaces are comparable; i.e., individuals $i$ and $j$ may be aware of different things in a given moment. Let $\Sigma \equiv \bigcup_{i \in N} S^i$ denote the union of the states of all the individual spaces.

We wish to keep track of how the different state spaces relate to reality ($S^0$) and, when possible, to each other. Therefore, define the surjective *projection* $r^{i \to j} : S^i \to S^j$, which is only defined if $S^i \succeq S^j$. Then, $s^j = r^{i \to j}(s^i)$ is the impoverished version of reality $j$ perceives relative to the awareness of $i$. By assumption, for all $i \neq 0$, $S^0 \succeq S^i$. Assume the projections are transitive: if $S^i \succeq S^j \succeq S^k$, then $r^{i \to k} = r^{j \to k} \circ r^{i \to j}$.

Here, the projections are functions that are pulling the awareness levels of all the individuals out from Nature's states. Thus, $s^j = r^{0\to j}(s^0)$ identifies the features of reality $s^0$ of which individual $j$ is aware when state $s^0$ is actualized. The awareness of $j$ is a feature of $s^0$. The transitivity rule ensures that Nature's states do not contain internal inconsistencies across awareness levels. By assuming the projections are onto functions, we impose an overarching consistency on the individual awareness structures across the states of Nature: $S^j$ contains all possible awareness states of $j$ – everything $j$ would find herself aware of in every state of Nature that could arise. Note that, in this paper, we assume that individuals have a limited-but-true awareness of reality. That is, Bob may be aware that it is raining outside the window while being unaware of the weather conditions in other geographic locations. However, Bob is not hallucinating rain when it is really sunny outside.

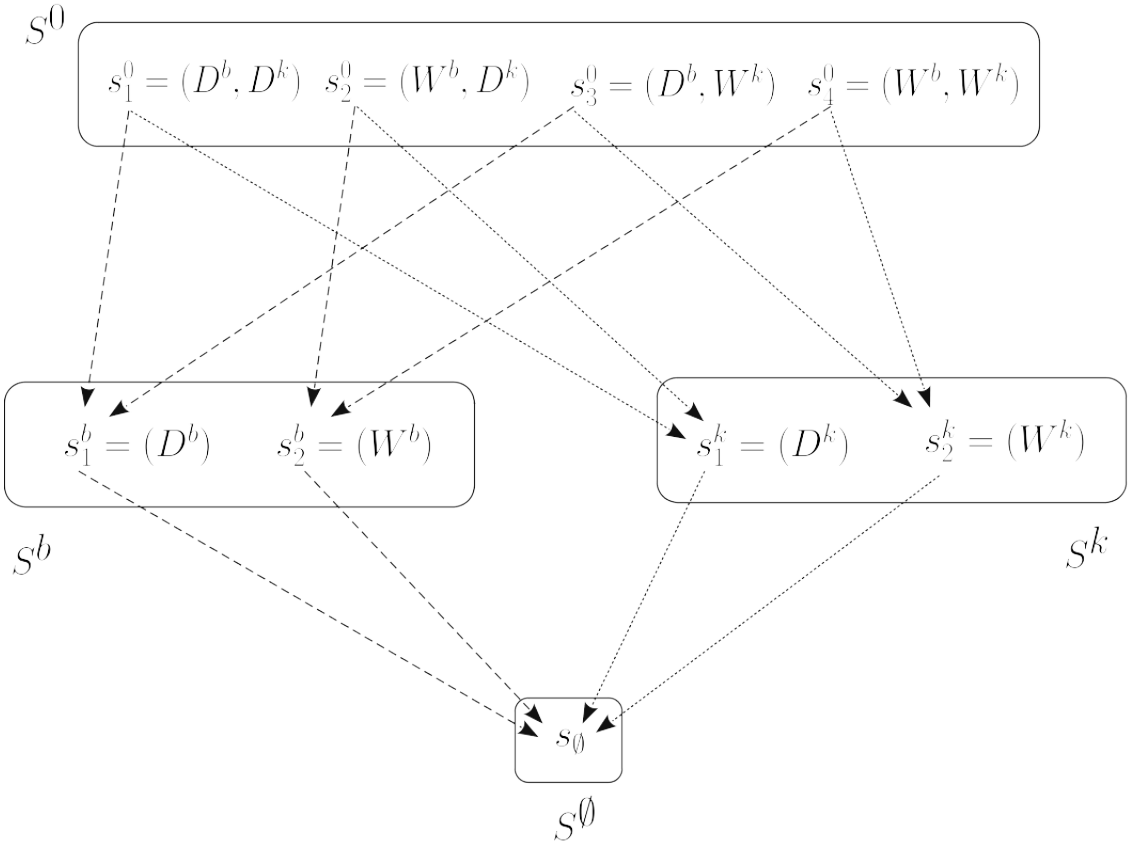


Figure 2: Awareness of Bob and Kate

To see a simple example of the setup, consider a situation in which Bob ($b$) and Kate ($k$) live in different cities and are looking at their front lawns. The lawns are either dry ($D$) or wet ($W$).

Let $D^b$ and $W^b$ indicate that Bob's lawn is either dry or wet, respectively, and similarly for Kate. Then, in this simple world, Nature's state space is $S^0 = \{s_1^0, s_2^0, s_3^0, s_4^0\}$, which are defined as shown in Figure 2 Suppose Bob and Kate are aware of the status of their own lawns but not of each other's. Then, Bob's state space is $S^b$, and Kate's is $S^k$. The projections from $S^0$ to $S^b$ and $S^k$ are shown as are the ones from the latter two space to $S^\emptyset$. (Not shown are the projections from $S^0$ to $S^\emptyset$.)

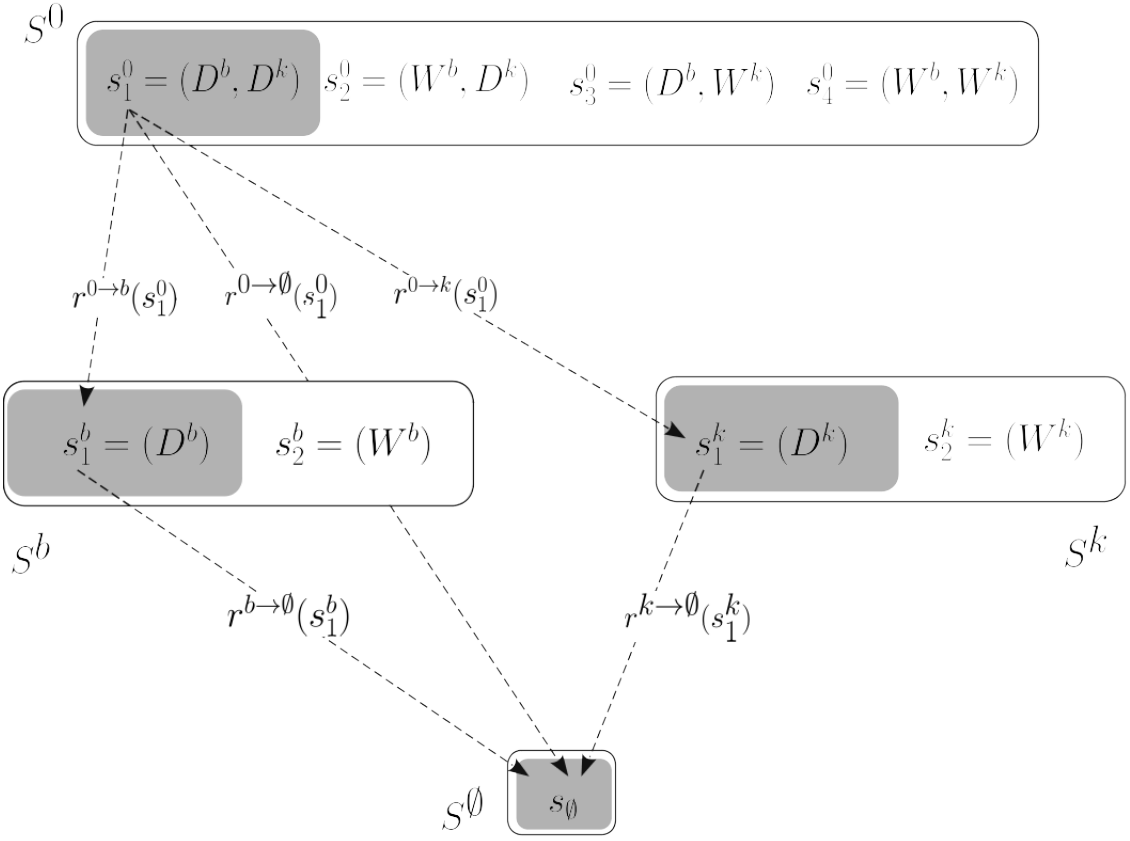


Figure 3: Awareness of state $S_1^0$ illustrated by the awareness diagram

Suppose state $s_1^0$ is actualized. Then, the projection functions map from this state to the awareness of each individual as well as to that of individuals of coarser awareness. Fig. 3 illustrates the actualized state and all the projection sequences mapping from it. For example, $r^{0\to b}(s_1^0)$ maps $s_1^0$ in Nature's space maps to $s_1^b$ in Bob's space (as shown): when the state of the world is that both lawns are wet, Bob is only aware of the wetness of his own lawn. In like fashion, $s_1^0$ maps to $s_1^k$ in Kate's space (as shown). All states in Bob's and Kate's spaces map to $s^\emptyset$ in $S_\emptyset$. Note that,

while $S^0 \succeq S^b \succeq S^\emptyset$ and $S^0 \succeq S^k \succeq S^\emptyset$, $S^b$ and $S^k$ are neither richer nor poorer than the other – they are not comparable. We might imagine an individual living in a third city who is unaware of the states of both Bob and Kate's grass. In this simple world, that individual would have complete unawareness – i.e., would have a state space equal to $S^\emptyset$. Keep in mind that the greyed areas in Fig. 3 showing what everyone is aware of is extracted from the information contained in Nature's actualized state, $s_1^0$.ig. 3.

Alternatively, we can imagine a world in which Bob is the only individual. Assume he can be in one of two places: Dallas ($D$) or Miami ($M$). He is aware of the wetness of the ground only in the geographic location in which he is present. He is also aware of his location. The awareness structure consistent with this setup is illustrated in Figure 4 (state and projection labels have been removed to reduce clutter).

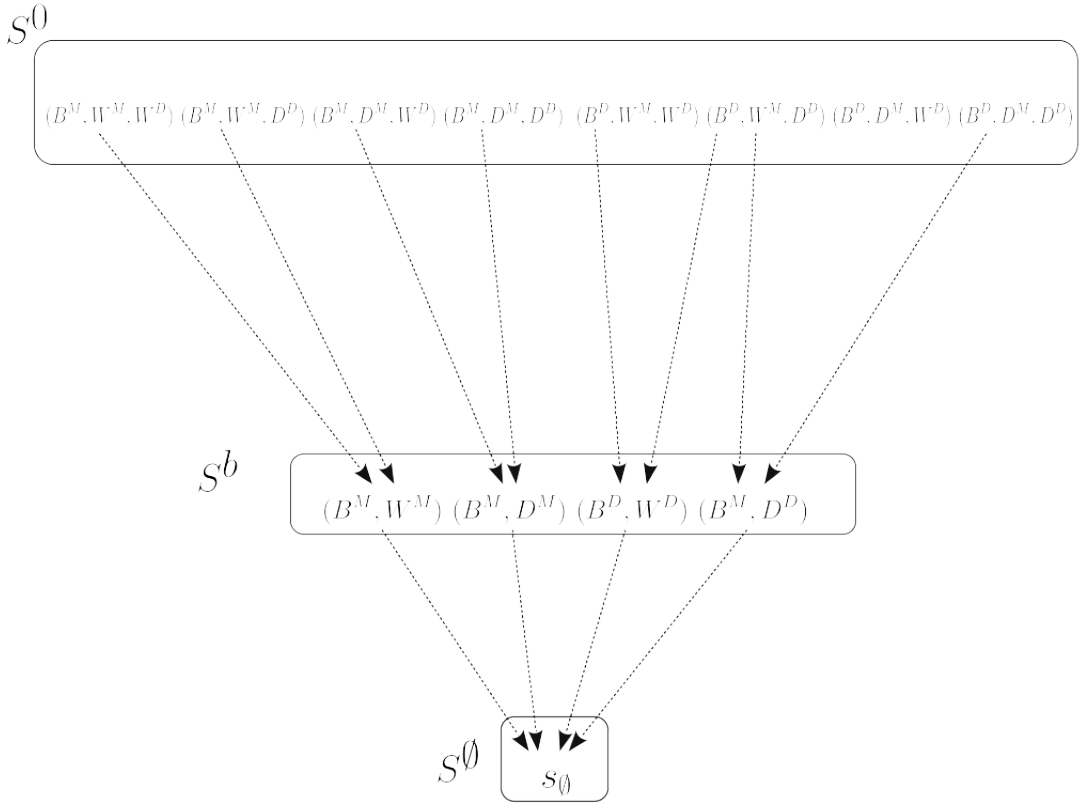


Figure 4: Bob's awareness is contingent on his location

### 3.3.4 Formalizing events (or properties)

The term 'event' is used differently in philosophy than it is in probability theory. Since we are writing to audiences familiar with one or the other, it is important to clarify this difference. In probability theory, 'event' is used similarly to the term 'property' in philosophy, where properties are understood intensionally. Philosophers typically use 'event' to mean a spatiotemporal particular extended over time. We refer to events associated with states at a moment in time (the game theory useage) as *synchronic events*, and those associated with states unfolding through time (the philosophy usage) as *diachronic events*. Below, we define the former. We wait to define the latter until Section 6.

In probability theory, events are subsets of state spaces. For example, the event "Mike intends to get a cup of coffee includes *all* states in which getting a cup of coffee is the intention of Mike. In philosophical terminology, this is equivalent to the property *being in a state in which Mike intends to get a cup of coffee,* where the intension of the property is all the states of the world in which the world exemplifies that property. Because each individual is associated with a state space that elaborates states according to the features of the world of which that individual could be aware in a given moment, the events of which he or she could be aware are subsets of that space. For example, $B = \{s \in S^0 | s \Rightarrow \text{Bob has a cup of coffee}\}$ is the event that collects all the states in Nature's space in which Bob has a cup of coffee.

Because individual state spaces may be related to one another and, in any case, are all related to reality fully elaborated ($S^0$), it will be helpful to associate events in $S^0$ with the awareness of the individuals of them. For an event $B \subseteq S^0$, let $B^{\downarrow} = \bigcup_{S^j \in \mathcal{S}} \left(r^{0 \to j}\right)^{-1}(B)$ be the extension of $B$ to include all states in the individuals' state spaces consistent with the projection of $B$ into them. Then, $E \subseteq \Sigma$ is a *synchronic event* if it is of the form $B^{\downarrow}$ for some $B \subseteq S^0$. We refer to the state space event, $B$, as the *basis* of the sychronic event $E = B^{\downarrow}$.

By this definition, not every subset of $\Sigma$ is a synchronic event. If $B \subseteq S^0$, define the negation of the synchronic event $B^{\downarrow}$, denoted $\neg B^{\downarrow}$, as $(S^0 \setminus B)^{\downarrow}$, a subset of $\Sigma \setminus B^{\downarrow}$. Typically, $B \cap \neg B$ is a strict subset of $\Sigma$; unlike the standard probability space setup, $2^{\Sigma}$ is not the event space. Nevertheless, by our definition, $\neg\neg B^{\downarrow} = B^{\downarrow}$.[3]

[3]This is not true in **?**. The difference is our extended events are generated by events in a specific space (in their terminology, Nature's space is always the "base space"), which is an added restriction that assures this nice technical feature.

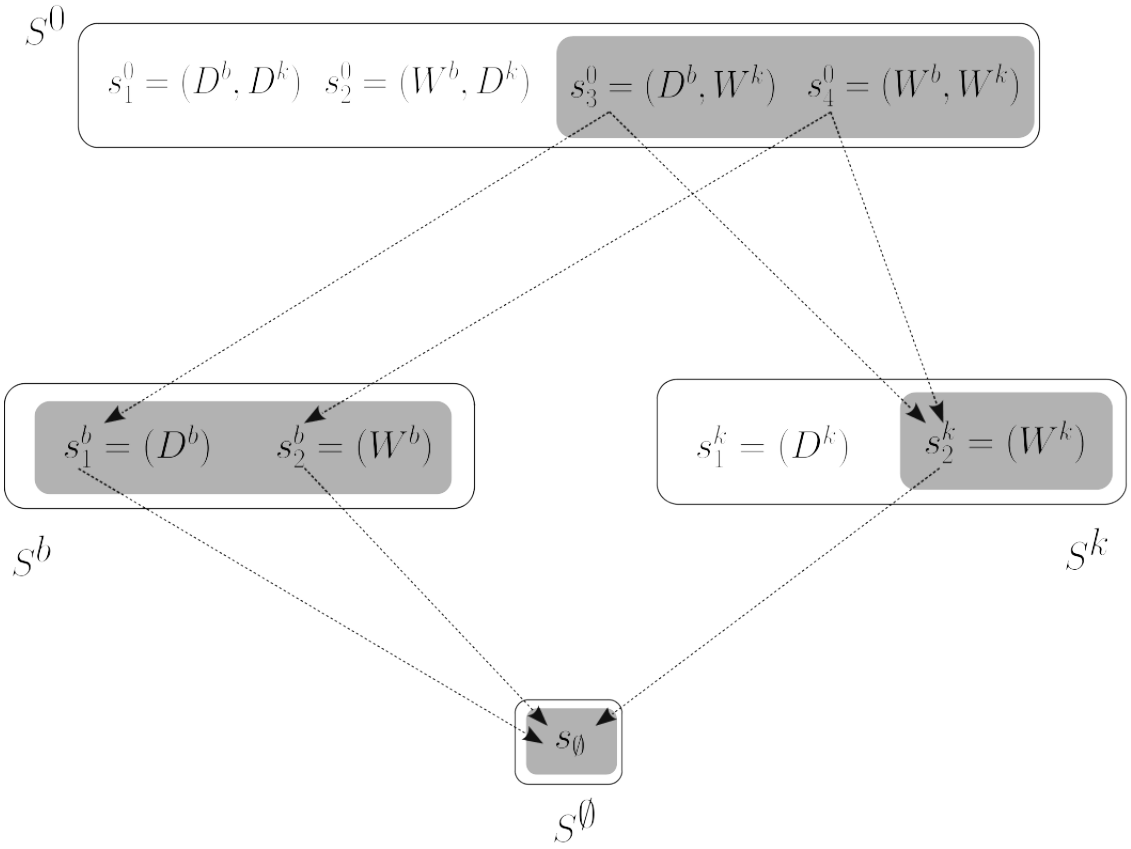


Figure 5: The synchronic event "Kate's grass is wet" with corresponding projections shown.

Returning to our example with Bob and Kate, consider the event "Kate's grass is wet" in $S^0$: $B = \{s_3^0, s_4^0\}$. $B$ is the basis of $B^\downarrow = \{s_3^0, s_4^0, s_1^b, s_2^b, s_2^k, s^\emptyset\}$. Notice that $\neg B^\downarrow = \{s_1^0, s_2^0, s_1^b, s_2^b, s_1^k, s^\emptyset\}$. Thus, $B^\downarrow \cup \neg B^\downarrow = \Sigma$. Here, the impossible event "Kate's grass is both wet and dry" corresponds to $\emptyset^{S^k}$.

# 4 Awareness structures in the toy case

## 4.1 The toddler's field of awareness

A *state* captures the world at a moment of time. The state space in a past or present period contains the state that actualized at that time and all the counterfactuals (states that could have occurred instead at that time). The state space for a future period contains all the states that could occur at that time.

In this example, the simplest state space is $S_0^n = \{A^*, B^*\}$. The $n$ superscript indicates that the state space is associated with "Nature" (i.e., these states are the ones associated with the world as it really is). Here, either Toy $A$ or Toy $B$ is the best toy for $i$. Which is truly the best is determined by Nature at the beginning of time.

The toddler lands in the situation at $t = 0$ in one state or the other. States elaborate everything about the world at a given time, including the cognitive status of everyone involved. Here, we wish to expand the parsimonious representation in Fig. 1 to dive more deeply into $i$'s cognitive status as it evolves through the decision process.

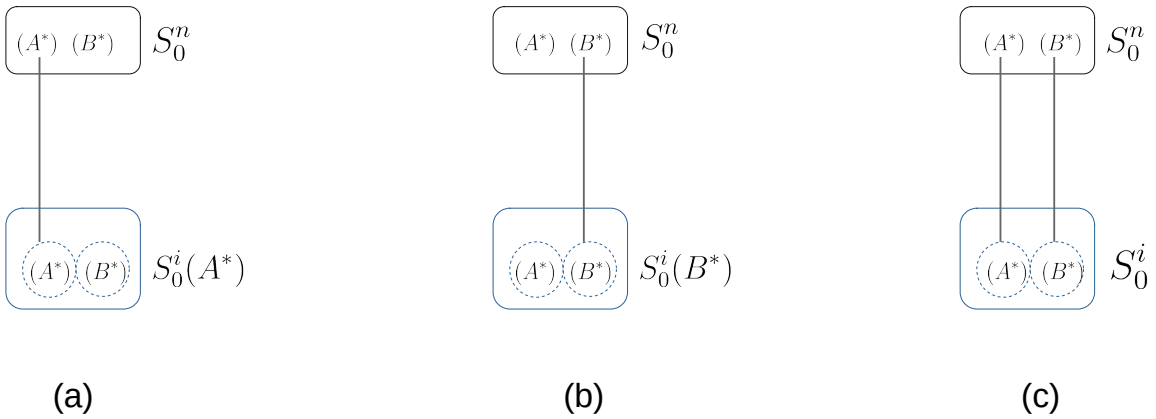


Figure 6: Toddler's FOA in (a) state $A^*$, (b) state $A^*$, and (c) shorthand, combined diagram

We are interested in describing what $i$ knows (or would have known) about the actualized (or counterfactual) states of the world contained in $S_0^n$. We do this by equipping $i$ with a description of the world as he is aware of it. We refer to this description as $i$'s *field of awareness* (FOA). Because $i$'s cognitive status is determined by the actualized state, we write, e.g., $S_0^i(A^*)$ to indicate the state space describing $i$ awareness in state $A^*$. If $i$ is fully aware, then $S_0^i(A^*) = S_0^n$. Specifically, the FOA includes everything about the actualized state about which $i$ can reason, including counterfactual states.

In addition, the state elaborates $i$'s beliefs about which state is true. For example, it may be true that $A$ is the best toy (state $A^*$) and that $i$ knows it. Alternatively, it may be true that $A$ is best, but $i$ is uncertain about it. Fig. 6 panel (a), illustrates the case in which $A$ is the best toy (shown by brackets around $A^*$ in space $S_0^n$). Below $S_0^n$ we depict $S_0^i(A^*)$. The grey line shows that, when $A^*$ is true, the FOA for $i$ is $S_0^i(A^*)$. The toddler is aware of both states. The

dashed lines around the states indicate what she knows. Here, she knows that the true state is $A^*$. Moreover, she can reason about counterfactual state $B^*$. The blue dashed line around $B^*$ means that $i$ believes that, had $B^*$ been the case, she would know it.

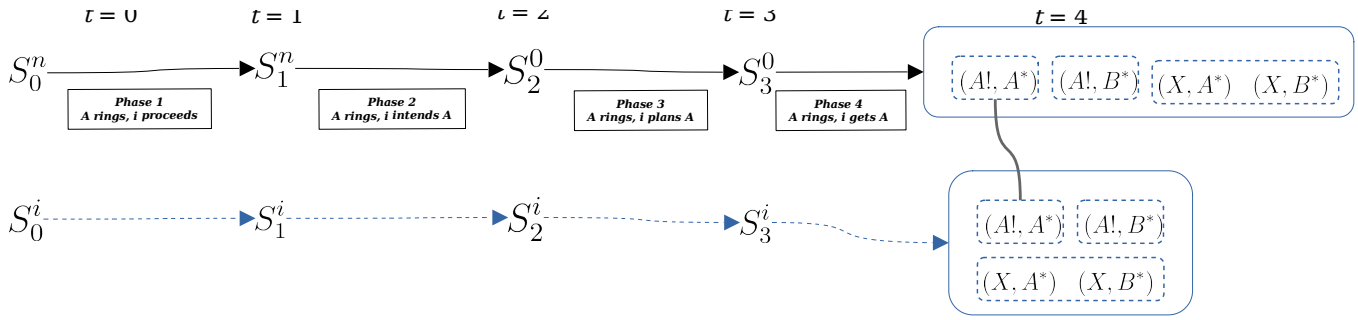


Figure 7: Toddler's FOA in (a) state $A^*$ with uncertainty about which state is true

Panel (b) illustrates the FOA that arises when $B^*$ is the true state of the world. When awareness and beliefs are consistent across states, diagrams can be combined without danger of ambiguity, as shown in Panel (c): when $A^*$ is true, $i$ knows it and, moreover believes that, had $B^*$ been the case, he would have known that ... and the converse when $B^*$ is true. Illustrated this way, the implication is that beliefs about which toy is best are the same in both state. In other words, Nature determines which toy is best, $i$ sees the two toys and forms beliefs about which is best. These are *the* beliefs regardless of which toy is truly best. Note that this diagram corresponds to the information set (the nodes connected by the dashed line) illustrated in the parsimonious game shown in Fig. 1.

Fig. 13 Panel (b), illustrates a situation in which $i$ is unaware that Toy B could be the best. In this case, $i$'s FOA includes a single state: $A^*$. The diagram shows both states projecting into it. That is, whichever state is true, $i$ is only thinking about Toy A. Perhaps Toy B is under the couch or, even if it is in $i$'s field of vision, he simply isn't thinking about it. Not that, in the situation depicted in Fig. 13 Panel (a), $i$ could believe that $B^*$ is impossible – i.e., a probability-zero occurrence. In Fig. 13 Panel (b), the possibility of $B^*$ is not even in her mind – she is completely unaware of it. As we show elsewhere, these two cases (zero-probability and unawareness) are *not* equivalent.

Finally, Fig. 13 Panel (c) illustrates the flexibility of the setup. Here, in state $A^*$ the toddler

is uncertain whether $A^*$ or $B^*$. However, in state $B^*$, he is certain that $B^*$ is true and reasons (incorrectly) that she would similarly be certain that $A^*$ had that state occurred. Given the simplicity of the story, it is hard to provide an explanation of why this sort of inconsistency would arise. Nevertheless, the framework is sufficiently general to accommodate it.

### 4.2 A four-phase decision process

Our goal is to expand the parsimonious treatment of individual $i$'s cognitive process to include some of the elements discussed in the philosophy literature. We begin by elaborating what $i$ is aware of about the state of the world, how he thinks about his decision at a given moment in time, how the thinks about his decision in light of possible future states and how this evolves dynamically over time.

First, what is the decision process? One useful disaggregation for our context is one that involves four-phases. At $t = 0$, individual $i$ finds himself in a state of the world in which Nature presents him with the potential to make a decision about something. For example, our toddler finds himself faced with a potential welter of information along with an opportunity to begin the process of achieving some goal. For example, he is set down on the with the two toys in front of him. Toy $A$ is making a ringing sound. There are other people in the room doing interesting things, the TV is on, and the pet dog is barking in the next room. The thought occurs to him that one end he might consider is obtaining one of the toys. The toddler chooses to think about obtaining the best toy. In this setting, he naturally organizes the available information for the next phase – he considers the condition of the toys, realizing that what the other people are doing, what is happening on the TV, and what the dog is doing are all irrelevant to an analysis of which toy is best.

Importantly, *states* indicate what is, could have been, or could be true about the world at a moment in time. Alternatively, *phases* represent the flow of the world between states. In this framework, phases are the periods within which acts occur. Acts include the acts of Nature, physical actions by individuals, and non-physical acts by individuals (e.g., pondering a choice). Acts are the events that cause the world to evolve from one state to the next over time (where "Nature's acts" are all the things that co-determine the next state in conjunction with the acts of individuals). It is also important to note that the acts associated with higher phases require the completion of the lower phases. For example, one cannot begin analysis until one chooses to do so and organizes the information required for it. In game-theoretic language, the acts associated with Phase $k$ are not

*feasible* until Phases 1 through $k-1$ are completed.

At the conclusion of Phase 1, the decision maker arrives in $t = 1$. The actualized state in which he finds himself is determined by the state actualized in $t = 0$, the acts of Nature, of $i$, and of others during Phase 1. At this point, $i$ has either settled upon opting to continue observing the situation further (which begins another iteration of Phase 1) or to move forward and analyze whether to form the intention of interest based upon the information organized in Phase 1 (which begins Phase 2).

In **Phase 2**, $i$ analyzes the information organized in Phase 1 and determines whether forming an intention is, indeed, warranted and, if so, which one specifically. Intentions are commitments to achieve some future state of the world (i.e., in which the desired end is obtained). For example, in Phase 1, the toddler chose to think further about forming an intention to obtain the best toy. As part of that choice, he also organized the available information so as to support that analysis (the condition of the toys versus other features of his environment). Suppose that, at the beginning of Phase 2 ($t = 1$) Toy $A$ is still making the ringing sound and that this leads the toddler to conclude that $A$ is best. As Phase 2 ends (at $t = 2$), the analysis is concluded and an intention is either formed (move to on Phase 3), the analysis is chosen to be extended (reiterate Phase 2), or the movement toward the focal end is abandoned (return to Phase 1).

**Phase 3** begins at $t = 2$. During this phase, $i$ formulates a plan to complete the intention. A plan provides a decision maker with a menu of state-contingent acts over time designed to achieve the intention. In a given state, the decision maker is presented with information about the present state, beliefs about both the present and the future (and, thus, beliefs about what the acts of Nature and others will be), and a set of feasible acts to do in during the next phase. A rational decision maker chooses a plan that selects the best among the feasible acts at each future state (with respect to achieving the intended end).

Here, it may be helpful to make some comparisons with game theory. In game theory, what we are calling a *plan* is referred to as the decision maker's *strategy*. Typically, game theory presents an interactive decision situation (i.e., in which players' decisions interact to affect the outcome of the game) as an extensive for game. Fig. 8 expands Fig. 1 to depict a situation in which the toys ring and the toddler is unsure of which toy is best. Here, the information sets (nodes connected by dashed lines) indicate that, at the time of his decision, $i$ knows which toy is ringing but not which is best. However, $i$'s beliefs are that the ringing toy is best. Then, in this game, Nature determines

the state, $i$ forms beliefs based upon what is happening, and chooses rationally.

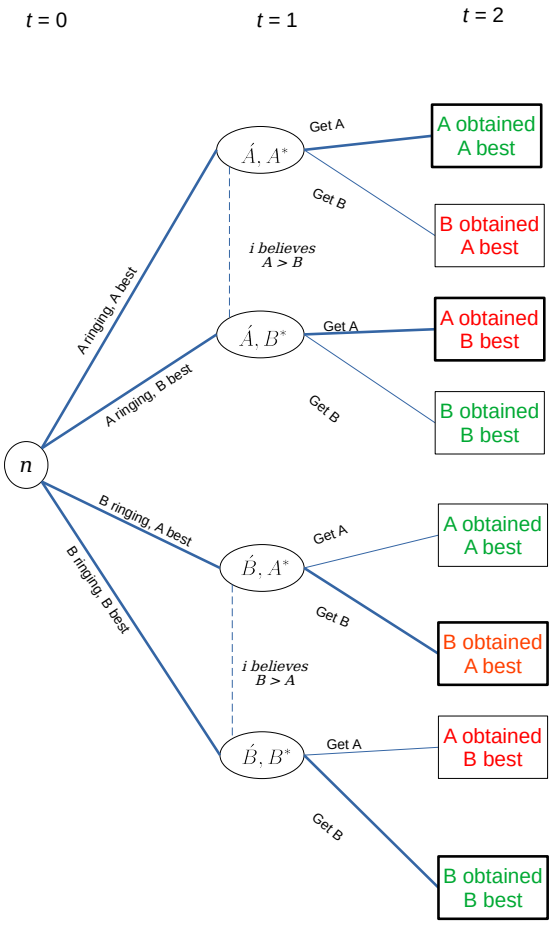


Figure 8: The complete but imperfect information game in which Brian's Toddler must act

Here, $i$'s rational strategy is: If $A$ is ringing, get A, otherwise get $B$. Note well that "rationality" does not imply perfect foresight. Rather, it means that $i$ chooses the subjectively optimal action given his beliefs. Beliefs may or may not be consistent with objective reality, depending upon the primitives of the game. In our example, a well-calibrated Bayesian player would believe that Nature selects among the four possible states according to some probability distribution. His beliefs about which node is really best would equal the true conditional probability of which is best given which toy is ringing. In most games, a players' desires are represented by a dollar-denominated utility number associated with each path through the game (the terminal nodes indicate the outcome of

play). Then, the rational Bayesian player selects the action that maximizes his expected utility. More generally, players' subjective beliefs need not be consistent with objective reality, in which case the "rational" player selects according to his subjective expected utility.

This brings us to the notions of *complete* versus *perfect* information games. A game of complete information is one in which the players know everything about the structure of the game, including everyone's payoffs: that is, the players know what game they are really playing. Games of perfect information are those in which, at every decision node, the associated player knows everything that happened in the past (i.e., there is no uncertainty regarding which node each player is on at any point). In Fig. 8, the assumption is that Brian's Toddler knows everything about the game, but at the time of his choice of action, is uncertain about which toy is best. Games of incomplete information are typically analyzed by transforming them into games of imperfect information: players know the set of possible games, Nature chooses the true game at the beginning of time and, faced with a decision, the player's form beliefs about both which game they are playing and where in that game they find themselves. The game in Fig. 8 is a game of complete but imperfect information.

The parsimonious game theoretic treatment focuses upon the actions of the decision makers. In Fig. 8, Nature acts by choosing the true state, $i$ arrives at a decision about which toy to choose given uncertainty about which is best. His choices are $A$ or $B$. Following his choice, the game resolves with an outcome. The question being asked in Fig. 8 is: *suppressing the details* of how he comes to formulate his strategy in the depicted situation (but making some assumptions about rationality), what strategy would that be? The treatment is "parsimonious" because the Phases 1-3 in our framework are left implicit.

In **Phase 4**, $i$ arrives at $t = 3$ and – if the state at that point is consistent with the plan developed in Phase 3 – acts in accordance with it. One way a state would be inconsistent with the plan is if it was not contemplated as a possibility in the plan's "menu" of possible states. (Such inconsistencies are not possible in a complete information game.) We will have more to say about what constitutes an inconsistency later but, for now, *we claim that running into an inconsistency causes the decision maker to revert to Phase 1.*

## 4.3 Where the unawareness model is richer

### 4.3.1 When everything is going smoothly

We begin by illustrating the case in which everything goes smoothly. **Phase 1:** $i$ hears $A$ ringing and decides to move forward to Phase 2. **Phase 2:** Toy $A$ continues to ring and, influenced by this information, $i$ believes that $A$ is best, thereby forming an intention to get $A$. **Phase 3** $A$ continues ringing and $i$ plans to effectuate his intention by crawling to the location of Toy A and picking it up. **Phase 4** $A$ continues ringing and $i$ crawls to Toy A and picks it up, thereby concluding the decision-action process. Notice that $i$'s intention to obtain the best toy may or may not have been satisfied at $t = 4$. If so, the process concludes. If not, $i$ may choose to think about obtaining $B$.

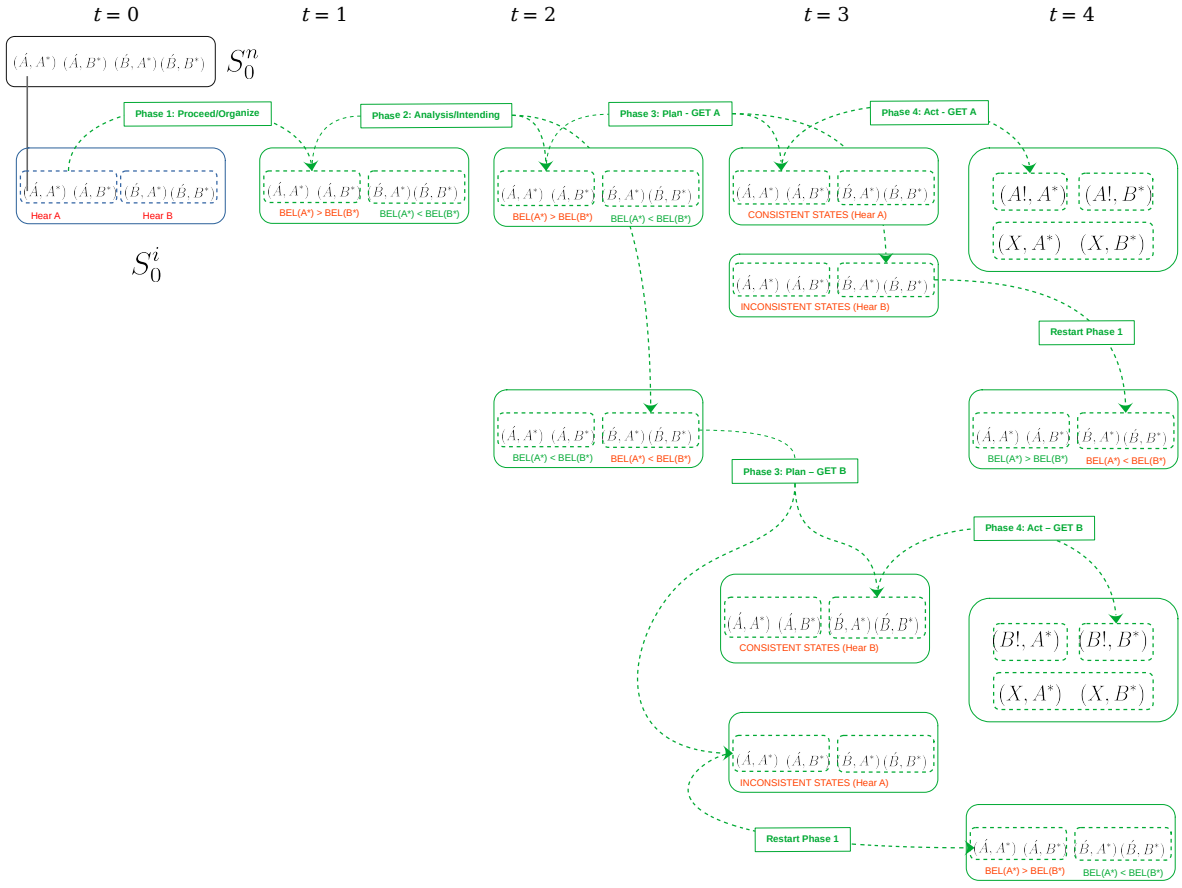


Figure 9: The four-phase process as considered by $i$ at $t = 0$

The situation in $t = 0$ is illustrated in Fig. 9. Phase 1 generally involves limiting the Phase 2

FOA to a subset of focal states. Thus, given all the elements of the world of which $i$ is passively aware at $t = 0$ (other people, the TV, and the barking dog), he will limit his attention to a strict subset of states thought to be decision-relevant. In our depiction, the decision-irrelevant features of reality mentioned above are not present at $t = 0$ (we erase them from $S_0^n$ to simplify the diagram – but the reader is welcome to imagine a much larger state space at $t = 0$ that includes all these elements).

Specifically, we define the states of the world at $t = 0$ as $S_0^n \equiv \{(\acute{A}, A^*), (\acute{A}, B^*), (\acute{B}, A^*), (\acute{B}, B^*)\}$ where the accent indicates that the toy is ringing a bell: e.g., $(\acute{A}, A^*)$ is the state in which A is ringing and A is best. Assume this is the true state. Further, assume $i$ is fully aware but uncertain which toy is best. This is indicated by the blue box, and labelled $S_0^i$. As shown, $i$ is aware of all the states and knows $A$ is ringing. However, as shown by the dashed lines (*information sets*), he is uncertain whether $A^*$ or $B^*$. The solid grey line shows that the FOA depicted is the one that arises in state $(\acute{A}, A^*)$. At $t = 0$, $i$'s has some belief endowment.

In addition to what the individual is aware of with respect to the present situation, he also has thoughts about the future. Tracking all four phases, our individual considers the future unfolding as shown by the green objects. Moving to Phase 2, the FOA may be reduced or expanded, as discussed, and Nature may act in a way that changes beliefs. Once the analysis is complete, and the intention is formed to obtain the preferred toy, at which point $i$ formulates a plan.

To emphasize, while the individual is thinking, analyzing, planning, and acting, the world evolves. Our individual $i$ considers how things might unfold. For example, during the analysis, which ends at $t = 2$, new information may come in to change $i$'s assessment of which toy is best. During this interval, a plan to get A or to get B is being shaped. The planning phase follows the intention. In this example, $i$ is rational – he plans to get whichever toy he believes best. In this case, real-world events may once again intrude upon the process, this time in ways that disrupt the plan. This is illustrated in the figure. For example, $i$ may plan to get A yet experience something that happens to indicate that B is really best (e.g., Toy B starts ringing). When the plan is disrupted, we assume that the decision maker must backtrack to the analysis phase.

Then, if the state of the world in $t = 3$ is consistent with the plan, $i$ proceeds to act. So, in the top row of individual 1's projected future, following the plan to Get A, a state of the world occurs in which 1 continues to believe that A is best and, hence, 1 acts to obtain A. However, $i$ considers the possibility that he will land in an inconsistent state in $t = 3$.

If the FOA evolves in a fashion consistent with the plan, then $i$ moves to the final phase in which he acts to obtain $A$. When a toy is obtained, the state is indicated so with an exclamation mark (e.g., $A!$). Here, $i$ allows for the possibility that he fails to obtain his planned objective, indicated by $X$ instead of $A!$ or $B!$. Note also, that $i$ believes that, upon successful completion of the action to get a toy, he will know whether it is the best one or not (indicated by the green dashed lines) and will be able to reason about the counterfactual (obtaining the planned toy but realizing it is not the best). If something happens to disrupt the plan, $i$ will be uncertain about which is best.

*Suppose everything goes according to expectations.* Fig 10 shows the situation at $t = 1$. The world has evolved to $S_1^n$ in which state $(\acute{A}, A^*)$ continues to hold. Reality is illustrated on the top row. Individual $i$ begins the analysis phase based upon the information organized in Phase 1.

Keep in mind that the state labelled $(\acute{A}, A^*)$ in $S_1^n$ is not identical to the one so labelled in $S_0^n$, even though the state spaces appear to be the same. First, the states in $S_1^n$ include the information about the sequence of preceding states (i.e., $s_1^n = (\acute{A}, A^*)$) as well as about the acts that caused them (i.e., $A$ *continues* ringing and $i$ chooses to consider forming an intention in Phase 1). Second, individual 1's state of mind has changed (and, remember, this is also summarized by the state). He recalls what he knew before, $S_0^i$ as well as what actions were taken by himself and Nature. This is indicated by the blue dashed line. Another difference is that he projects the decision process from the present into the future. This is indicated by the green objects – although they are consistent with his state of mind in $t = 0$, they are what is in his mind in $t = 1$.

In Fig. 11, Phase 2 concludes in $t = 2$ with $i$ forming an intention to get $A$. Toy $A$ continues to ring. This is relevant because we have assumed that $i$ believes whichever toy is ringing is best. Since $A$ rings throughout, there is nothing happening on the part of Nature to disrupt the process.

In Fig. 12, Phase 3 concludes in $t = 3$ with a plan formulated to act towards the attainment of the intention. The sound of $A$ ringing continues. This is consistent with the plan to get $A$. Therefore, $i$ proceeds to act according to plan – that is, to actually get $A$.

Finally, as shown in Fig. 13, the process draws to a conclusion: toy $A$ is obtained and, as it turns out, this is indeed the best. Notice that all uncertainty has been resolved – by obtaining $A$, it is discovered to be the best.

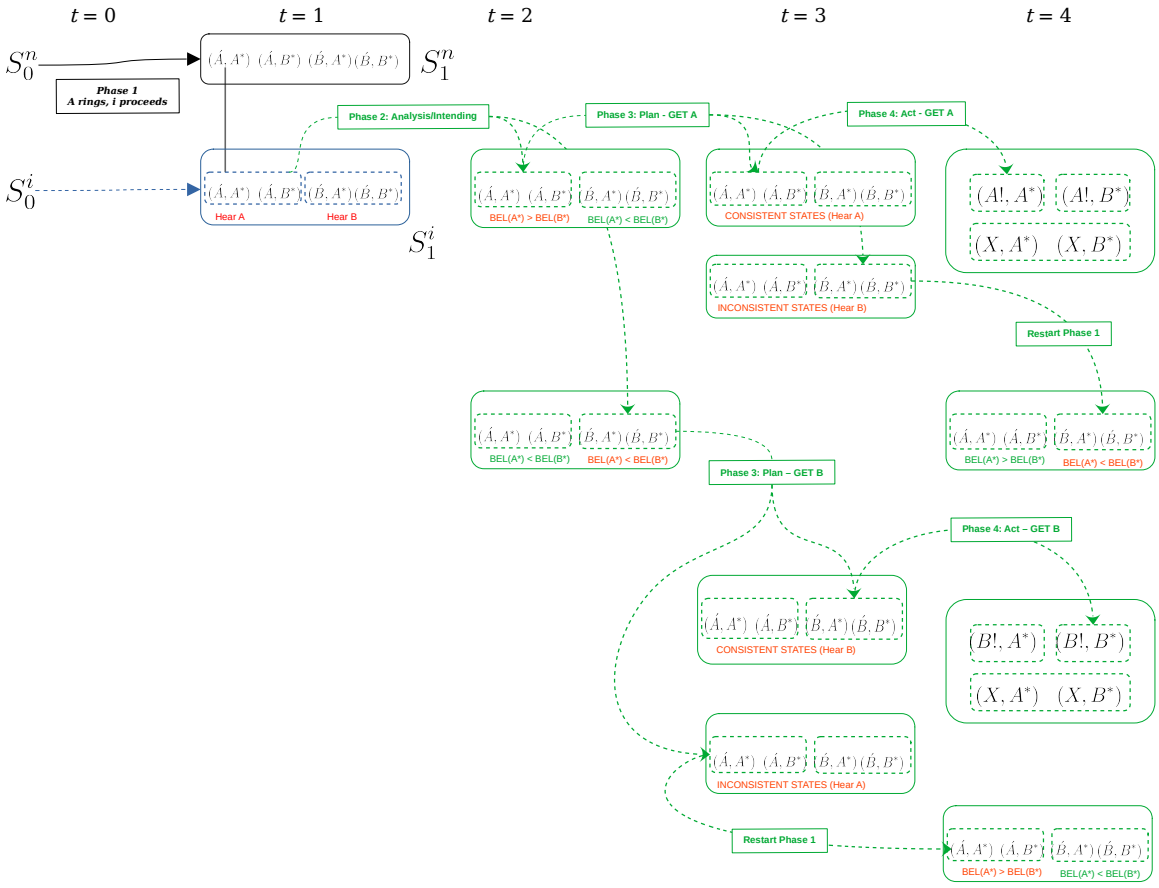


Figure 10: Environment stable, Brian's Toddler begins analysis,

### 4.3.2 The smooth case can be represented as an extensive-form game

So far, we have not written anything down that can't be shown as an extensive form game. This is not surprising given that, thus far in our example, individual $i$ continues to have full awareness of the situation and, as such, has expectations about the future that are consistent with reality. To represent what we are doing as an extensive form game, Fig. 8 would need to be expanded to include all the acts of $i$ (organize information, form intention, set up plan, act to get a toy, etc.) as individual decisions. As, well, it would need to account for the acts of Nature in each intervening period (choosing a ringing toy).

Unlike an extensive form game, our diagrams provide an elaboration of what is going on in $i$'s mind (explicitly: recalled history, FOA, and projected futures). However, this level of detail

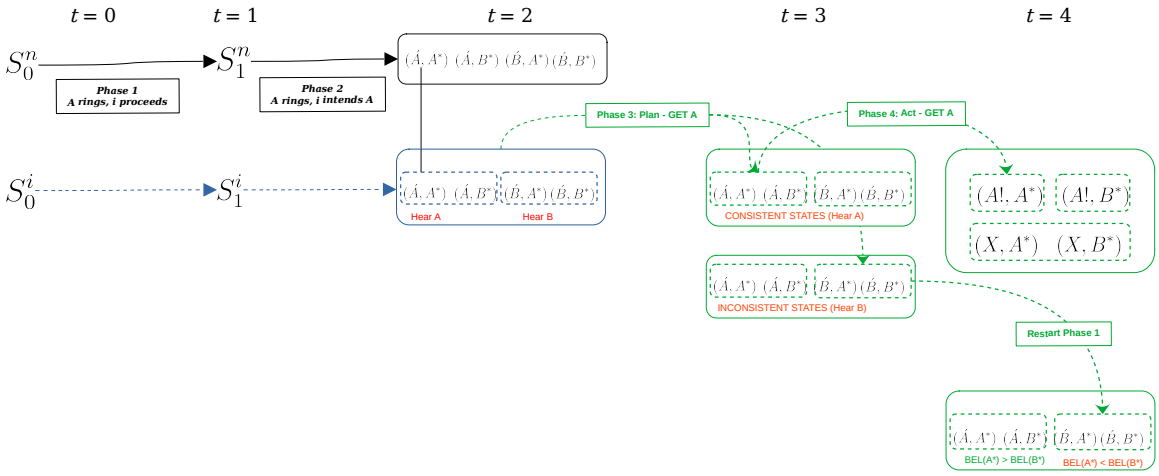


Figure 11: Environment stable, intention formed, planning begun

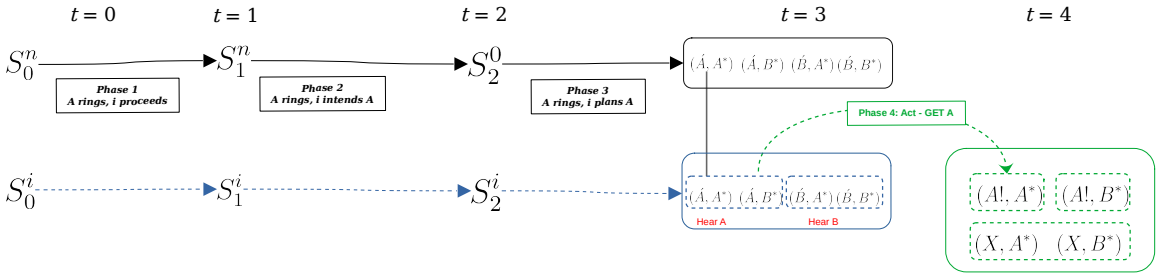


Figure 12: Environment stable, plan complete, action begun

comes at the expense of illustrating everything happening over time in one diagram (.e.g, Fig. 8). Game theorists would surely point out that, while these diagrams are not incorrect, they are also not necessary. Elaborating all these details in our example is superfluous since they are all implied by the joint stipulation that the setting is one of complete information and perfect recall (i.e., $i$'s assessments about present and future possibilities is correct and he remembers everything he knew at in an earlier period). Nevertheless, since this paper is attempting to bridge idea between two disciplines, we judge the additional elaboration to be worthwhile.

The next step is to illustrate what can go wrong – how an even in a complete information

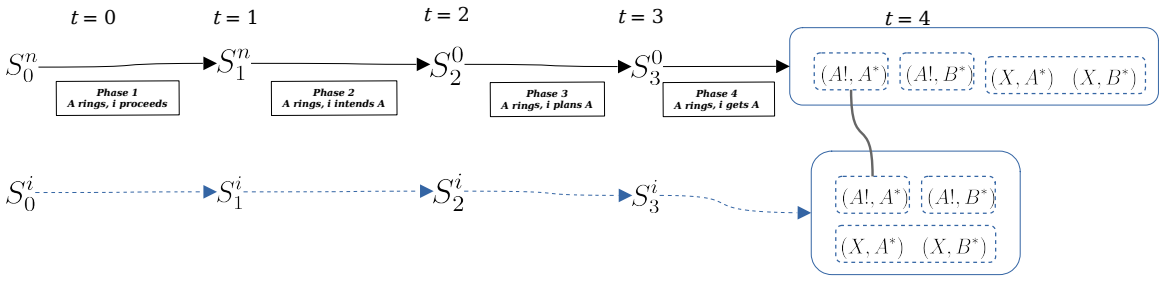


Figure 13: Process complete, intention satisfied!

setting, a decision maker can get stuck in a "paralysis by analysis" situation. This, then will bring us to introduce our idea of intentional unawareness and its effects on the decision process.

### 4.3.3 Paralysis by analysis

The previous discussion suggested that things can go wrong when nature evolves state spaces more quickly than the decision maker's response process. This situation is illustrated in Fig. 12: the interpretation now is that the state actualized in period $t = 2$ reflects an act by Nature (not illustrated but, e.g., Toy B sounding a bell) that interrupts the intended plan by causing the toddler to switch beliefs from $A > B$ to $B > A$. Given this new state of mind, a new plan is formulated – to get $B$.

This alternating flow of information can go on indefinitely, as illustrated in Fig. 15. Under the standard, Bayesian belief-desire model, a rational agent is one who immediately updates beliefs based upon new information. This is true in the sense that, provided the information is indeed informative (and not just noise), more information will lead to better decisions. The problem highlighted here is that taking new information into consideration requires an allocation of time and cognitive resources, both of which are in finite supply.

By explicitly accounting for this fact, we see that a truly rational decision maker must weigh the benefits of recalibrating while postponing acting versus ignoring new information and moving forward. The purpose of making a decision is to act and of acting is to achieve an end. If one never decides, then one attains the desired end – the means to which is the deciding. (Note that"never" is too high a bar – if one discounts future utility streams, then there is always a tension between

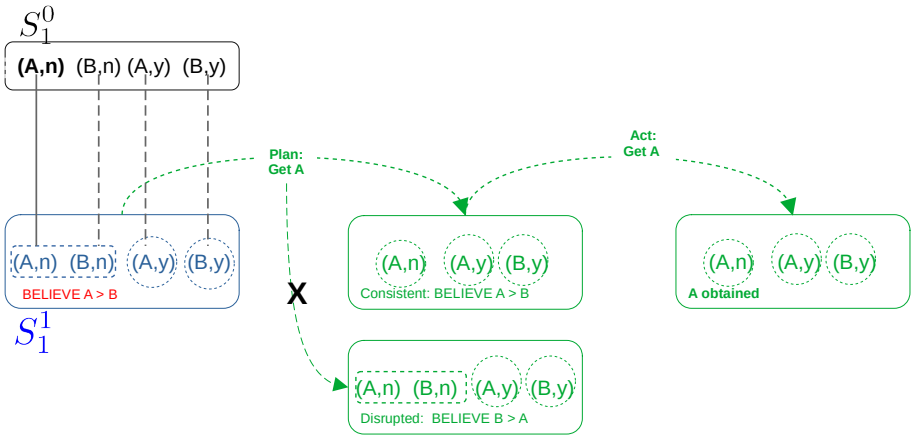


Figure 14: Nature disrupts the plan

taking time to analyze in lieu of acting.)

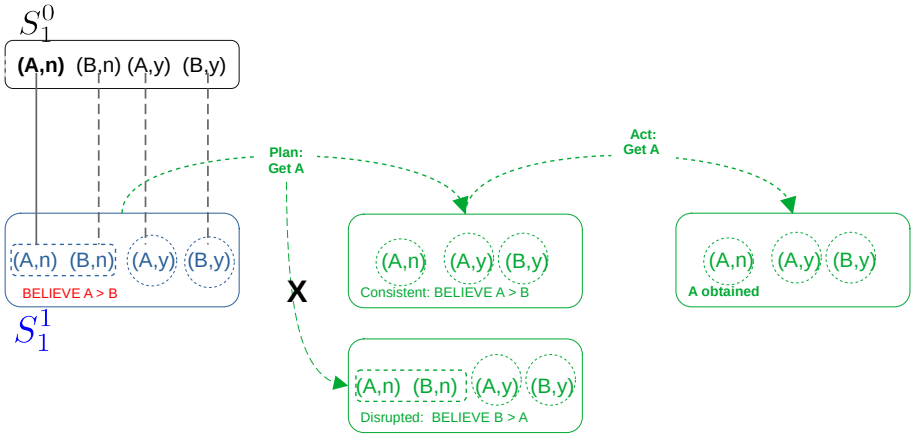


Figure 15: Nature disrupts the plan again

### 4.3.4 Intentional unawareness

We can now see how unawareness solves the problem. Here, there are three interrelated aspects: intention, planning, and awareness. How these are distinguished from one another in a fashion amenable to philosophers is a question that requires further discussion. Speaking roughly, at some point, an individual "commits" to making an act.

It seems that an intention (a commitment to attain some end) always implies a corresponding commitment to a plan (a program of action designed to achieve the intended end). Even in the case of split-second decision making, e.g., a policeman's intention to stay alive during a drug raid upon observing an unidentified person moving in the room behind the door, a "plan" is required – e.g., aiming the rifle and pulling the trigger. Thus, a random twitch in one's leg does not count as an intentional act. The converse is not true: one can make many plans without intending to put them into action.

The new idea that we are proposing is that along with the plan comes another piece of the puzzle – intentional unawareness. In the most disaggregated elaboration, there are at least two places where unawareness arises. The first is at the observation/analysis stage. Here, the individual decides what aspects of the present state of the world to pay attention to (we will call this *weak* unawareness). Having focused upon a particular set of aspects about the world, the individual proceeds to conduct an analysis. The conclusion of the analysis is a decision either to continue thinking about and analyzing the situation or to commit to the attainment of some end.

In the preceding paragraph, I mention weak unawareness because it seems there is an important distinction worth making. One can be unaware of some things which one could call to mind (weak) as well as of some things about which one cannot call to mind (strong). As I was writing a moment ago, I had music playing in the background. I was unaware of the name of the band playing the music. It was not in my mind at all. I was not thinking about it. Then, as I started thinking about examples to write down, this one occurred to me. When it did, I naturally recalled the name of the band. A moment ago I was also unaware of what engineering details are required for a working teleportation system (which could be the null set if such systems are impossible). Now, I am aware of the question, but not of the answer – nor will I ever be.

Here, there is another distinction. A moment ago, I was not thinking about what the temperature is in Hong Kong. It was not in my mind at all. Now, the question is in my mind – I am aware of it. In this case, I have introduced a new information set into my FOA. It includes a *range*

of temperatures within which I think the true temperature lies. Presumably, I also form beliefs over that range and can report an expected temperature. I can also take an action (look on the internet) to discover the actual state of the world.

All of these awareness distinctions seem important to intending and planning.

With this commitment in place, the individual then develops a committed plan of action. The program may be simple, i.e., just selecting among one's presently available acts. Or, it may be complex, requiring much analysis – including deciding the best among several plans required to attain the end. In any event, the conclusion of this process is the committed plan. Unawareness arises here because a plan can be enacted without further analysis. An important caveat is that there must be some specification of what states the plan are included in the plan's FOA with the proviso that should a state arise that is not part of the plan's FOA – that is, should the individual become aware of something that falls outside the plan FOA, then the plan is disrupted. In other words, the plan allows the individual to put activity on "auto-pilot" for some FOA. Awareness of new states or events outside the scope of the plan have the potential to disrupt it – at a minimum, a reevaluation is required. [This bit needs further thought and refinement.]

With all of this in mind, return to the problematic case discussed above. Fig. 16 illustrates the situation in $t = 1$ in which the intention, plan, and planned unawareness all occur in the period. Notice the change from Fig.XXX: now, instead of a more refined set of future FOAs, the $(B, n)$ state has been eliminated from the future plan. The shift is intentional (or, at least, is implied by the plan). The potential for state $(B, n)$ to disrupt the plan is eliminated.

The implementation of the plan is shown in Fig. 17. The world evolves and, once again, Nature does her best to disrupt the plan. Now, however, the toddler is resolutely focused upon getting Toy A – he is unaware of the signals being sent by Nature to reevaluate the situation. The process is happily concluded as illustrated in Fig. 18.

# 5 Some applications

Here we need some payoff, in terms of what can be accomplished with a more realistic example. At least some discussion, if not some gestures toward models. How does incorporating unawareness change how we might act in

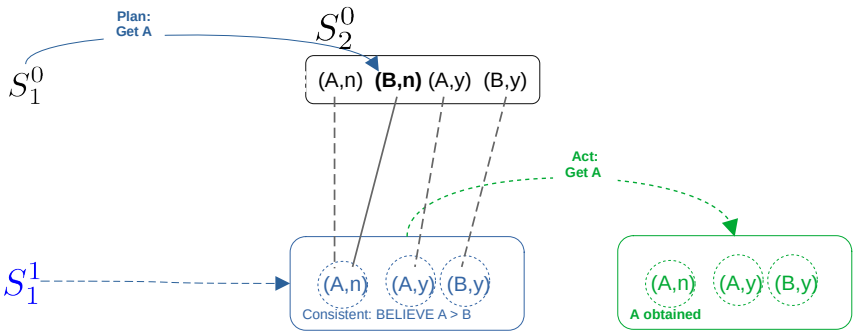


Figure 16: Individual 1's intended unawareness

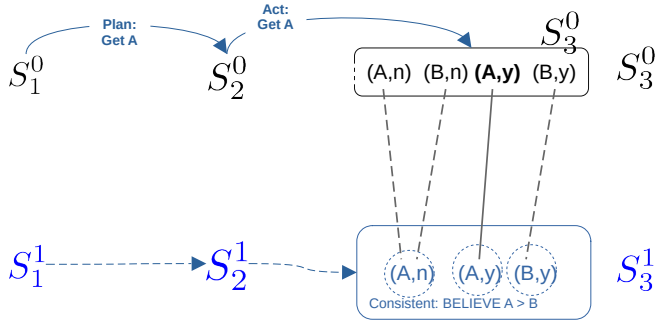


Figure 17: Actualized state $(B, n)$ does not disrupt the plan

## 6 Generalizing diachronic awareness

We now extend this framework to the dynamic setting. Assume time is discrete and limit attention to some finite number of periods, $T$. Time subscripts are added to indicate the period. For example, *Nature's state space at time t*, is denoted $S_t^0 \subset S^0$, where with an arbitrary element denoted $s_t^0 \in S_t^0$ Then, $S^0$ contains all the states that could possibly be actualized at $t$ expressed in their richest level of detail. We continue to use index subscripts to identify specific states when necessary; e.g., $s_{3,t}^0$ is the state indexed as number 3 in Nature's state space in priod $t$.

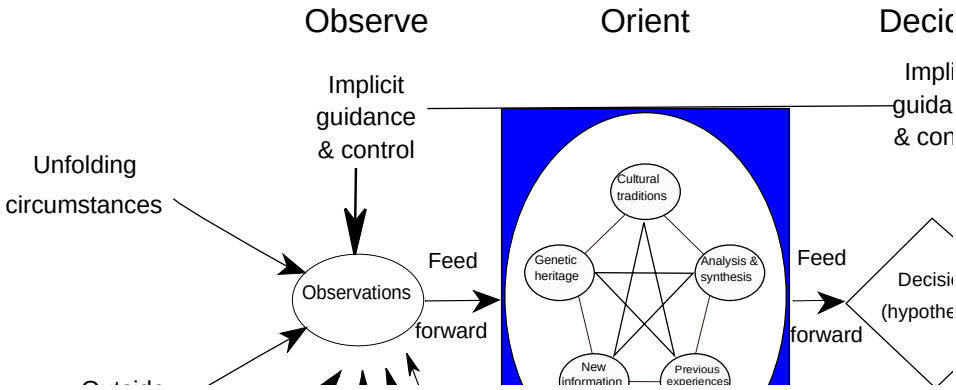


Figure 18: Mission accomplished

### 6.1 Acts and actions

The sequence of states actualized over the period of analysis is effected by the acts of the individuals in the population in conjunction with acts of Nature (i.e., all the causes that, in conjunction with the acts of the individuals, determine the actualization of a particular state from an immediately preceding, previously actualized state). For each individual $i \in N$ and each state $s \in S^0$, $A^i(s)$ indicates the set of *feasible acts available to individual* $i$ *in state* $s$ with arbitrary element $a^i \in A^i(s)$.[4]

We adopt the convention that $A^i(s) = \emptyset$ indicates that individual $i$ has no available acts in state $s$. An *act profile* is a list of acts, one for each individual, denoted $\mathbf{a} \equiv (a^0, \ldots, a^n)$. Recall, Nature is "Individual 0" so that $a^0$ summarizes all the developments that, in conjunction with the individuals' acts, determine which state is actualized following $s$. The set of *all act profiles at state* $s$ is $\mathbf{A}(s) \equiv \times_{i=0}^{n} A^i(s)$; the set of *all possible act profiles at time* $t$ is $\mathbf{A}_t \equiv \cup_{s \in S_t^0} \mathbf{A}(s)$; and the set of *all possible act profiles* is $\mathbf{A} \equiv \cup_{s \in S^0} \mathbf{A}(s)$.

### 6.2 Dynamics

As indicated above, the act profiles summarize all the conditions required to actualize one state from the previously actualized state. To formalize this, let $\omega : \mathbf{A} \times S^0 \to S^0$ be the *state-contingent*

[4]Notice that we use a capital letter to indicate that $A^i$ is a set-valued function: $A^i : S^0 \to 2^A$. Also note that feasible acts for individual $i \neq 0$ are determined by reality (states in $S^0$), not by $i$'s awareness of reality (states in $S^i$). Because we consider the intentional formation of some mental attitudes as choices available to individuals, we use the term "act" to describe the choices available to someone in a broad way. We think of "action" as describing the narrower category of act associated with physical movement.

*actualization function*, where $\omega(\mathbf{a}_t, s_t^0) = s_{t+1}^0$indicates that if the act profile at state $s_t^0 \in S_t^0$ is $\mathbf{a}_t \in \mathbf{A}(s_t^0)$, then the next state actualized is $s_{t+1}^0$. Notice that the engine of change operates at the level of Nature's reality: individuals find themselves in some true state $s_t^0$; they implement their human acts alongside Nature's act (this act elaborates the things that occur beyond the acts of the individuals represented in the analysis), as summarizedby $\mathbf{a}_t$; after which, the next state $s_{t+1}^0 = \omega(\mathbf{a}_t, s_t^0)$ is actualized. Assume that, for all $t$, $\omega$ is bijective from $\mathbf{A}_t \times S_t$ to $S_t^0$. In other words, each feasible act profile in a given state at time $t$ leads to a unique state in period $t+1$ and each state in period $t+1$ can be traced back to a singlepredecessor state in period $t$ by a unique act profile that links the two. Thus, the inverse $\omega^{-1}$ exists, where $\omega^{-1}(s_{t+1}^0) = (\mathbf{a}_t, s_t^0)$ indicates if $s_{t+1}^0$ is actualized, then the immediately preceding state was $s^0t$ and $\mathbf{a}_t$ was the enacted act profile. Suppose, for example, that two distinct sequences of acts could lead to an identical footprint in the snow. In that case, we consider there to be two states in which that identical footprint exists, each associated with one of the sequences of acts that lead to it. The world begins at state $s_0^0$. To allow for uncertainty or partial knowledge with respect to various aspects of the world at the beginning of time, we assume Nature's acts entirely determine $s_1^0$. That is, $\mathbf{a}_0 = (a_0^0, \emptyset, \ldots, \emptyset)$, where $a_0^0$ represents all the actualized historic factors that lead individuals to their first decision state, $s0_1 = \omega(\mathbf{a}_0, s_0^0)$. Uncertainty with respect to the state of the world in $t = 1$ (e.g., about the intentions or other individuals) is, thus, formalized as uncertainty about "Nature's act" $a_0^0 \in A^0(s_0)$ prior to the first decision period.

We define the *history at state* $s_t^0$ as a profile of states that starts at $s_0^0$ and ends at $s_t^0$, denoted $\mathbf{h}(s_t^0) = (s_0^0 \ldots, s_t^0)$. A history $\mathbf{h}(s_t^0)$ is *feasible* if there exists a sequence of action profiles $\mathbf{a}_0, \ldots, \mathbf{a}_{t-1}$ such that $s_1^0 = \omega(\mathbf{a}_0, s_)^0, \ldots, s_t^0 = \omega(\mathbf{a}_{t-1}, s_{t-1}^0)$. Feasible histories are the only ones that can be actualized according to objective reality.[5]

The set of all *feasible histories at time t* is $\mathbf{H}_T$ and the set of all subsets of histories is $\mathcal{H}_T$. An arbitrary *history at time t* is denoted $\mathbf{h}_t \in \mathbf{H}_t$, where we start with the *null history* $\mathbf{h}_0^0 = (s_0^0)$ at the beginning oftime (so, $\mathbf{H}_0^0 = \{\mathbf{h}_0^0\}$ and $S_0^0 = \{s_0^0\}$). Because there is a single root node and $\omega$ is a bijection, the set of paths in $\mathbf{H}_T$ form a tree. Thus, $S^0$ can be partitioned according to subsets of states corresponding to time periods: $S^0 = S_0^0 \cup \cdots \cup S_T^0$ and $S_0^0 \cap \cdots \cap S_T^0 = \emptyset$. Note also that

[5]This distinction allows for situations in which individuals subjectively consider infeasible histories to be possible. For example, individual $i$ may believe that act $a_t^i \in A^i(s_t^0)$ is consistent with the actualization of $s_{t+1}$ even though $a_t^i$ is not in the profile $\mathbf{a}_t$ that leads from $s_t$ to $s_{t+1}$. We do not examine these cases in this paper.

each $S_t$ implies a partition of $\mathbf{H}_T$ according to the sets of paths intersecting the states in $S_t$.

++++++++++++++++++++++++++++++++++++++STOPPED HERE++++++++++++++++++

### 6.2.1 Diachronic events

A *diachronic event* is a subset $D \in 2^{\mathbf{H}_T}$; i.e., a subset of paths in the tree associated with $\mathbf{H}_T$. Note that diachronic events are subsets of whole paths from $s_0^0$ to subsets of states in $S_T^0$. Therefore, they do not have time subscripts. Let $\mathcal{D} \equiv 2^{\mathbf{H}_T}$ be the set of all diachronic events. Given the preceding discussion, every synchronic event $\sigma_t \in \Sigma_t$ is associated with a unique event $E \in \mathcal{E}$.

To see how we use states and understand how these objects work, consider the canonical example of rolling a six-sided die. We use functions on $S^0$ to "extract" information from the states. Here, for example, we can let $d(s_t^0)$ indicate the outcome of a die roll in state $s_t^0$: for all $s_t^0 \in S_t^0$, $d(s_t^0) \in \{1, 2, 3, 4, 5, 6\}$; i.e., $d$ maps from each state in $S_t^0$ to a number between 1 and 6, indicating the side of the die that landed up in that state (where $s_t^0$ includes *all* features of the world besides how the die landed). Now, the synchronic event "the die roll is even" is described by $\sigma_t \in \Sigma_t$ such that $\Sigma_t \equiv \{s_t^0 \in S_t^0 | d(s_t^0) = 2, 4 \text{ or } 6\}$. Alternatively, suppose $T = 2$. Then, the diachronic event, "snake-eyes were rolled" is described by $E \in \mathcal{E}$ such that $E \equiv \{(s_0^0, s_1^0, s_2^0) \in \mathbf{H}_T | d(s_1^0) = d(s_2^0) = 1\}$.

## 7 Locating intentions (and other attitudes) in the model

**Beliefs** Beginning with beliefs, let $\Delta(H)$ denote the set of all probability distributions on the set of histories. Then, $\mu_i : S \to \Delta(H)$ is a function that maps from states to individual $i$'s beliefs on histories $H$. We write $\mu_i^s$ to indicate $i$'s subjective probability distribution on $H$ at state $s$. This distribution induces a distribution on history events, $\mathcal{H} \equiv 2^H$. Note that each $\mu_i^s$ induces a probability distribution on $S$. For example, the probabilities of the elements of $Z$ (terminal nodes) are equal to the probabilities of the complete histories they terminate. The probability of some arbitrary state $s_t$ is equal to the sum of the probabilities of the complete histories running through it, and so on. Since all of this is implied by $\mu_i$, we will slightly abuse notation and write, e.g., $\mu_i^s(Z) = \mu_i^s(H)$, even though $Z \in \mathcal{S}$ while $H \in \mathcal{H}$.

It is important to note that the existence of more than one element in $S_0$ means that individuals may be uncertain about which tree is the objective one and, hence, the true history they have experienced. If so, they will be uncertain about which state they are in. In addition, there will be

uncertainty about how the future unfolds. At the moment, we have the objective world starting at $s_0^*$ and unfolding in accordance with $\omega$ and the sequence of everyone's act choices. Since acts are free choices by individuals, it is possible they are selected randomly ("now, I will decide what to do by flipping a coin"). This includes acts of Nature. All of individual $i$'s speculation with respect to the history, state and unfolding of events is summarized by $\mu_i$.

Like in the case of incomplete information, we proceed by introducing probability distributions on state-spaces. For any state space $S \in \mathcal{S}$, let $\Delta(S)$ be the set of probability distributions on $S$. Even though we consider probability distributions on each space $S \in \mathcal{S}$, we can talk about probability of events that, as we just have seen, are defined across spaces. To extend probabilities to events of our lattice structure, let $S_\mu$ denote the space on which $\mu$ is a probability measure. Whenever for some event $E \in \Sigma$ we have $S_\mu \succeq S(E)$ (i.e., the event $E$ can be expressed in space $S_\mu$) then we abuse notation slightly and write

$$\mu(E) = \mu(E \cap S_\mu).$$

If $S(E) \not\preceq S_\mu$ (i.e, the event $E$ is not expressible in the space $S_\mu$ because either $S_\mu$ is strictly poorer than $S(E)$ or $S_\mu$ and $S(E)$ are incomparable), then we leave $\mu(E)$ undefined.

To model an agent's awareness of events and beliefs over events and awareness and beliefs of other groups, we introduce type mappings. Given the preceding paragraph, we see how the belief of an agent at state $\omega \in S$ may be described by a probability distribution over states in a less expressive space $S'$ (i.e., $S \succeq S'$). This would represent an agent who is unaware of the events that can be expressed in $S$ but not in $S'$. These events are "out of mind" for him in the sense that he does not even form beliefs about them at $\omega$: his beliefs are restricted to a space that cannot express these events.

More formally, for every agent $i \in N$ there is a *type mapping* $t_i : \Omega \longrightarrow \bigcup_{S \in \mathcal{S}} \Delta(S)$. That is, the type mapping of agent $i \in N$ assigns to each state $\omega \in \Omega$ of the lattice a probability distribution over some space. Now a state does not only specify which events affecting value creation may obtain, and which beliefs agents hold over those events, but also which events agents are aware of. Recall that $S_\mu$ is the space on which $\mu$ is a probability distribution. Since $t_i(\omega)$ now refers to agent $i$'s probabilistic belief in state $\omega$, we can write $S_{t_i(\omega)}$ as the space on which $t_i(\omega)$ is a probability distribution. $S_{t_i(\omega)}$ represents the *awareness level* of agent $i$ at state $\omega$. This terminology is intuitive because at $\omega$ agent $i$ forms beliefs about *all* events in $S_{t_i(\omega)}$.

For a type mapping to make sense, certain properties must be satisfied. The most immediate one is *Confinement:* if $\omega \in S'$ then $t_i(\omega) \in \triangle(S)$ for some $S \preceq S'$. That is, the space over which agent $i$ has beliefs in $\omega$ is weakly less expressive than the space contains that $\omega$. Obviously, a state in a less expressive space cannot describe beliefs over events that can only be expressed in a richer space. We also impose Introspection, which played a role in our prior discussion of incomplete information: every agent at every state is certain of her beliefs at that state. In AppendixXX, we discuss additional properties that guarantee the consistent fit of beliefs and awareness across different state-spaces and rule out mistakes in information processing.

It might be helpful to illustrate type mappings with an example. FigureXX depicts the same lattice of spaces as in FiguresXX and XX. In addition, we depict the type mappings for three different groups. At any state in the upmost space $S_{pq}$, the blue agent is aware of $p$ but unaware of $q$. Moreover, she is certain whether or not $p$ depending on whether or not $p$ obtains. This is modeled by her type mapping that assigns probability 1 to state $p$ in every state where $p$ obtains and probability 1 to state $\neg p$ in every state where $\neg p$ obtains. (The blue circles represent the support of her probability distribution that must assign probability 1 to the unique state in the support.) An analogous interpretation applies to the red agent except that she is an expert in $q$. In contrast, the green agent is aware of both $p$ and $q$ but knows nothing with certainty, modeled by her probabilistic beliefs in the upmost space that assigns equal probability to each state in it.[6]

Unawareness structures allow us to model an agent's awareness and beliefs about another agent's awareness and beliefs, beliefs about that, and so on. This is because, as in the incomplete information case, beliefs are over states and states also describe the awareness and beliefs of groups. Return to FigureXX. At state $pq$ the green agent assigns probability 1 that the blue group is aware of $p$ but unaware of $q$. Moreover, he assigns probability 1 to the blue agent believing with probability 1 that the red group is unaware of $p$.[7]

**Desires** For all $i \in N$, define the state-dependent *desire relation* such that, for all $s \in S$, $D_i^s \subset P \times P$ where, $(p', p'') \in D_i^s$ means that individual $i$ in state $s$ desires the path $p''$ at least

[6]The example is taken from **?** who shows how a generalist (i.e., the green agent) emerges as an entrepreneur and forms a firm made of specialists (i.e., the blue or red agents) in a knowledge-belief and awareness-based theory of the firm using strategic network formations games under incomplete information and unawareness.

[7]We note, it has been shown that under appropriate assumptions on spaces $S \in \mathcal{S}$ and the type mapping, unawareness structures are rich enough to model any higher order beliefs of agents (see the working paper version of **?**).

as much as the path $p'$. Having described the mathematical structure of desires, we use the more intuitive notation $p' \preceq_i^s p''$, which is defined to mean $(p', p'') \in D_i^s$. We use $\prec_i^s$ and $\approx_i^s$ to indicate strict preference and indifference, respectively.

Why make preferences over paths? Because we assume individuals care about how they get to an end as well as the end itself. To take a canonical example, a homeowner may have a renovated kitchen in mind as the desired end. However, even if the kitchen specs are provided in extensive detail (so the owner knows exactly what the end will be), there may be many contractors who can deliver it. In this case, assuming there are several contractors from which to choose, each of which identify with a different path with states encoding costs at each step of the way and the final quality of the work, the owner's choice will be based upon the path (costs) as well as the final state (quality). Similarly, an individual sensitive to the time value of money will prefer shorter paths to longer ones, other things equal. Or, individuals may value portions of the paths themselves. For example, even though a student drops out of school (thereby, not completing the degree), he or she may nevertheless value the portion of the education that was completed. Our approach allows for special cases in which all these details are elaborated as primitives of the situation. For our discussion, we simply assume preferences are over paths.

**Intentions** Finally, define the state-contingent *intention* for individual $i$ as a function $\gamma_i : S \to \mathcal{S}$, where $\gamma_i(s) = E$ means that in state $s$ individual $i$ intends event $E$. We assume that individuals have desires and beliefs in all states, but not necessarily intentions. The idea here is that, e.g., in some states Mike intends the end "Mike has a cup of coffee" and in others, Mike has yet to form intentions. We adopt the convention that $\gamma_i(s) = \emptyset$ means that $s$ is a state in which individual $i$ has not formed an intention. We highlight that states may be differentiated only by changes in mental attitudes. For example, it may be that the only change from $s_t$ to $s_{t+1}$ is $\gamma_i^{s_t} = \emptyset$ to $\gamma_i^{s_{t+1}} = E$. This suggests that the interval between time periods may be very short (measured in milliseconds).

This raises the question of how an individual moves from being in a state without an intention to one in which the intention is formed. Here, we can require an act of commitment to cement the intention. That is, if $s_t$ is a state in which $i$ does not have an intention, then the set of feasible acts, $A_i^{s_t}$, can include an *act to form the intention* to "get a cup of coffee," which would then take him to a state $s_{t+1}$ in which $\gamma_i^{s_{t+1}} = X$ where $X$ contains all the states consistent with $i$ having a cup of coffee.

For all $i \in N$, individual $i$'s *mental attitudes* are summarized by a triple denoted $\theta_i \equiv (\mu_i, D_i, \gamma_i)$.[8] A *profile of mental features* for all the individuals is given by the profile $\theta \equiv (\theta_1, \ldots, \theta_n)$. Given our conventions, we can write $\theta_i(s)$ and $\theta^s$ without ambiguity.

——- STOP HERE ————

# 8 An interesting connection

While thinking about this example, I came across the idea of OODA Loops (Observe, Orient, Decide, Act), which gained popular use in the military. The wiki discussion is https://en.wikipedia.org/wiki/OODA$_{l}$oo

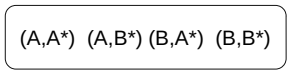


Figure 19: John Boyd's OODA Loop

The articles are interesting for at least two reasons. First, they are dealing with split-second decision situations, for example a soldier entering a building in hostile territory and having to decide whether to shoot at someone moving past an open doorway in the room ahead. The fact that this

[8] In setting up mental features in this way, we are following a version of the familiar "type-space" approach used in game theory (See **??**).

model is used for training people to make split-second decisions in life-or-death situations suggests there is something to it. Most importantly, it suggests that some process like this is going on at the basic cognitive level and not only, e.g., at the level of higher-level, complex decisions that may involve explicit, more extended data gathering and analysis phases.

The other interesting angle is that this model assumes that interrupting the OODA loop resets the competitor's loop all the way back to the first stage. Hence, the explicit reason soldiers want to get inside the enemy's OODA loop is precisely that disrupting the process disrupts the enemy's ability to act. One way of thinking about our previous example is that Nature is "getting inside" the decision maker's "OODA loop".e

## 9 A deeper dive into the state spaces and FOAs in this example

The preceding discussion was intended as a first, rough cut exposition designed to outline and illustrate some key ideas. Having done that, there are some aspects worth refining.

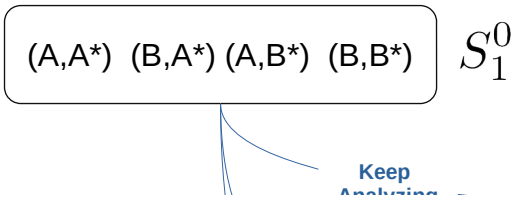


Figure 20: A more accurate set of Nature's states in $t = 1$

The previous elaboration of nature's state spaces was compressed in order to get at the overall framework without too much clutter. A more accurate representation is shown in Fig. 20. In $t = 1$, Toddler hears one of the Toys ringing a bell and believes one of the toys is best. The states are $(x, y)$ where $x$ is the toy that is ringing and $y$ is the toy that is best. The toddler always believes that the ringing toy is going to be the best toy. Since acting is not allowed without a plan, there are not states in the first period in which a toy is obtained. Presumably, the truly best toy is determined by Nature at the start, in period $t = 1$.

In this example, the transition from $S_1^0$ to $S_2^0$ depends upon the act of the toddler in period 1. There are three possibilities: i) continue to gather information; ii) commit to a plan to get $A$; or iii) commit to a plan to get $B$. Each act leads to a corresponding Nature's state space ($S_{2,i}^0$

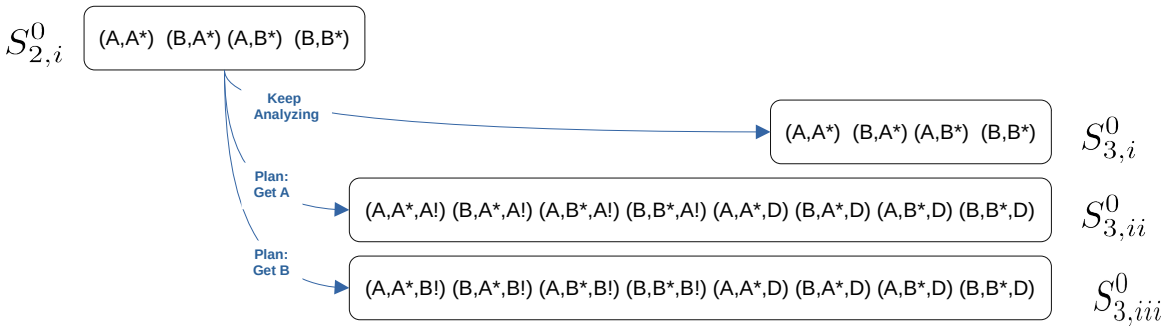


Figure 21: Possible state spaces in $t = 2$ depend upon toddler's act

through $S^0_{2,iii}$ as illustrated in Fig. 21). In the second period, the actualized state $(x, y, z)$depends upon which toy is ringing in $t = 2$ $(x)$, which toy is actually best $(y)$ and, if operating under a committed plan, whether the state is consistent with the plan (e.g., $z = A!$ if the plan is to get $A$) or the toddler becomes aware of something that disrupts the plan $(z = D)$.

The tree expands substantially in period $t = 3$. The possibilities are illustrated in Fig. 22. What happens in $t = 3$ depends upon the state space actualized in $t = 2$ in conjunction with the act of the toddler in $t = 2$. If the toddler continued analysis, then the possibilities are $S^0_{2,i}$ through $S^0_{2,iii}$, mirroring the possibilities in period $t = 2$. If the committed plan was to get $A$, then three possibilities are illustrated: Space $S^0_{3,iv}$, corresponding to the plan being disrupted in $t = 2$ and the toddler choosing to reanalyze the situation; Space $S^0_{3,v}$, corresponding to the plan being disrupted in $t = 2$ and the toddler committing to a plan to get $B$; and Space $S^0_{3,vi}$, corresponding to an actualized state consistent with the plan and the toddler acting to get $A$. The possibilities following $S^0_{2,iii}$ mirror these (shown as $S^0_{3,vii}$ through $S^0_{3,ix}$).

It is important to note that, although $S^0_{3,i}$, $S^0_{3,iv}$, and $S^0_{3,vii}$ appear to be identical (the analysis state space), in fact they are not. The reason for this is that each state also contains its own history. Since the histories leading to each of these state spaces is different, technically, the states they contain are not identical.

Finally, what do the toddler's FOAs look like in this example? Let us consider the sequence described in the previous, problematic case: Nature switches between ringing toys, starting with $A$ , the toddler commits to the plan to get $A$, following which there is no disruption. As shown in

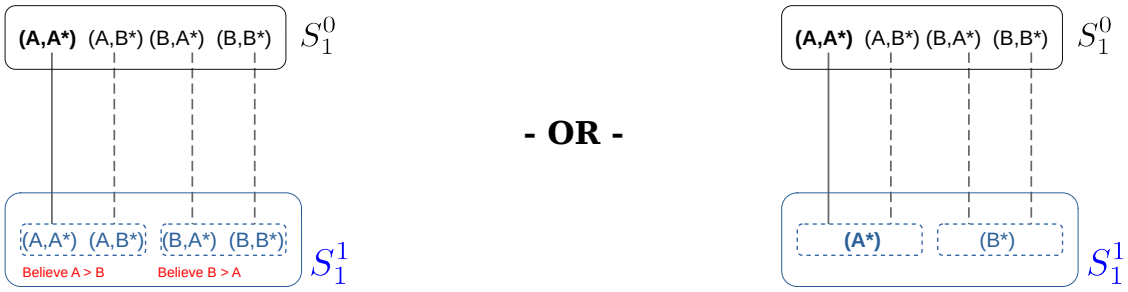


Figure 22: Possible state spaces in $t = 2$ depend upon toddler's act

Fig. 23, there are a couple of possibilities for the toddler's FOA that come immediately to mind. On the left-hand side, the toddler is aware of all the states but uncertain about which toy is best. Her beliefs are such that he believes the ringing toy is best. On the right-hand side, the toddler is unaware of all the possibilities. Rather, she is certain that whichever toy is ringing is best. The latter is probably a good model of a toddler, the former would be better for a sophisticated decision maker.

In $t = 2$, illustrated in Fig. 24, the plan and its associated unawareness kick in. Even though toy $B$ is ringing, the toddler's FOA projects all the plan-consistent states into a single, plan-to-get-$A$-consistent state, $(A!)$. Here, the toddler is shown as being aware of the possibility that something

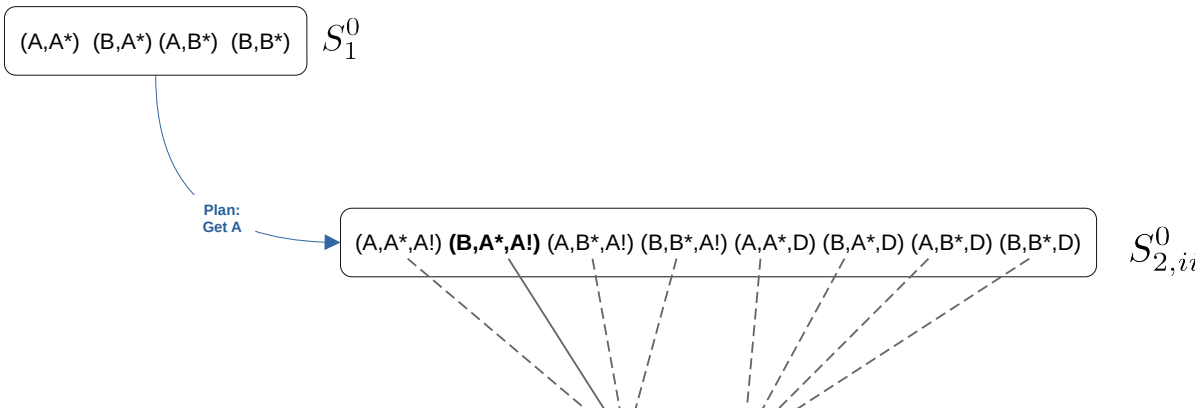


Figure 23: Toddler believes the ringing toy is best

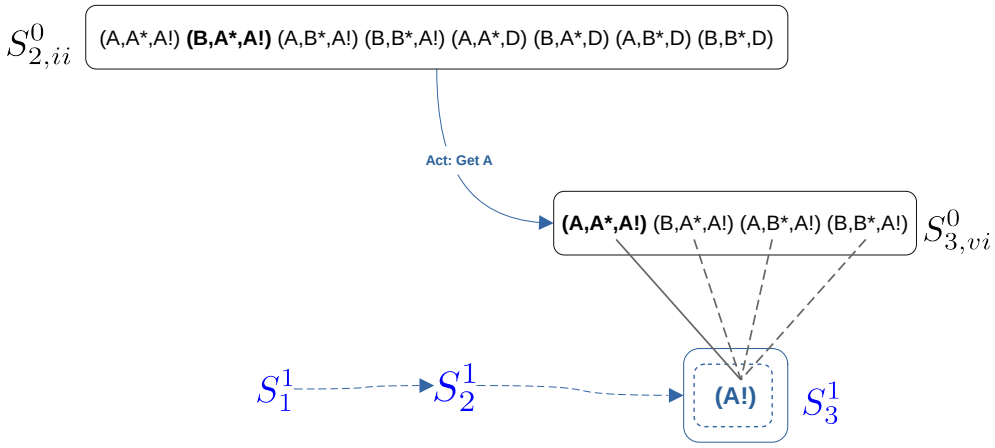


Figure 24: Plan coupled with intentional unawareness

could happen to disrupt the state.

I am not sure this depiction is accurate. Since everything is going according to plan, including $(D)$ may be incorrect. Rather, if something actually happened to disrupt the plan (Mom enters the room and says, "It is time for bed!"), then that would count as an act of Nature, resulting in the FOA shown (with the $(D)$ state included in the FOA).

Finally, we come to the last stage, in which the toddler obtains $A$. This is shown in Fig. 25. This is fairly straightforward, though there are a couple of options here as well. In the top version,

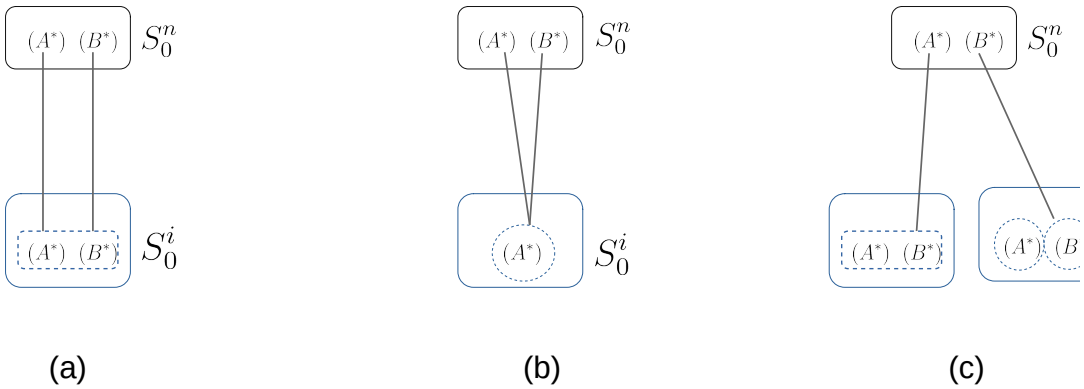


Figure 25: Final outcome possibilities: only aware of getting $A$ or all relevant state details

toddler gets $A$ and is only aware of that – having achieved her intended end, she simply moves on to other things. Alternatively, the toddler may be aware of the true state (it really is the best toy!) and, as well, may be able to reason about the other possibilities. Note that all states in the toddler's FOAs have the $A!$ (get $A$) indicator, since this is accomplished by his act in $t = 2$. Therefore, it is redundant (and could be removed). But, keeping the indicator there is fine since it is consistent with what happens in all states.

## 10 Consistency conditions

Having structured the objects of interest, we now explore various conditions required to impose the regularities between the various mental attitudes and between those attitudes and the external world that are appropriate to a rational human being.

**Reality Alignment** Beginning with the latter, our setup allows individuals to believe (place positive probability on) things that are not objectively true. However, it is difficult to square rationality with someone whose beliefs are completely divorced from reality. Therefore, we assume beliefs align with reality at least to some extent.

**Condition 1** (Grain of Truth)**.** *For all* $i \in N$, $s_t \in S$, $\mu_i^s(h_t^*) > 0$.

That is, rational individuals do not rule out the true state of affairs. This implies that, although an indivual's beliefs about an event may be wildly inaccurate, that belief is not completely irrational: i.e., for all $W \in \mathcal{H}$ such that $\mu_i^s(W) > 0$, $h_t^* \in W$. Going in the other direction, for all $h_t^* \in H^*$, there exists some $W \in \mathcal{H}$ such that $\mu_i^s(W) > 0$. This condition is not without controversy as it does rule out situations in which an individual is surprised by being confronted with a state of affairs he or she had previously thought impossible. There are formal approaches to dealing with such situations. For now, however, we sidestep such issues.

**Learning** We can also think of consistencies implied by learning. Even with the Grain of Truth Condition in place, our setup presently allows a person's beliefs through time to be completely inconsistent in all ways except $\mu_i^s(h_t^*) > 0$. For example, suppose $X, Y \in \mathcal{H}$ and $\mu_i^{s_t}(X) = 1$ and $\mu_i^{s_{t+1}}(Y) = 1$ ($X$ and $Y$ contain all the states $i$ believes are possible in periods $t$ and $t+1$, respectively). Then, even if $X$ and $Y$ are quite large, there is nothing in the setup preventing $X \cap Y = h_{t+1}^*$; i.e., the *only* consistency from period to period is belief in the possiblity of the objectively true history. Such situations seem inconsistent with any reasonable concept of learning. The following condition is a notion of learning that admits a wide range of learning models. For example, Baysian updating is consistent with this (though, by no means requred).

**Condition 2** (Weak Learning)**.** *Let* $X, Y \in \mathcal{H}$. *For all* $i \in N$, $s_t, s_x \in S, x > t$, *if* $\mu_i^{s_t}(X) = 1$ *and* $\mu_i^{s_x}(Y) = 1$, *then* $Y \subseteq X$.

Notice that learning is, indeed, weak in the sense that one may never learn anything ($Y = X$ through time). However, we imagine that as individuals experience the world, their grasp of it becomes more refined. Again, this condition is also not without controversy since it seems to rule out "conversion" experiences in which an individual shifts from one worldview to another, apparently inconsistent worldview. Whether or not such experiences are, in fact, inconsistent with Condition 2 we leave for another discussion.

**Introspection** It seems reasonable to assume that an individual knows his or her own mental features (but may be uncertain of those of others). For example, being certain of one's own beliefs rules out some peculiar mistakes in information processing (e.g., **?**, **?**). As described above, the probability distribution representing an individual's beliefs in may vary by state. Introspection entails that, at any given state, the agent's belief assigns probability 1 to the set of states in which he has the same belief as in that state. Formally,

**Condition 3** (Introspection)**.** *For each agent $i \in N$ and state $s \in S$, the agent's belief at $s$, $\mu_i^s$, assigns probability 1 to the set of states in which $i$ has precisely these beliefs:* $\mu_i^s(\{s' \in S \mid \mu_i^{s'} = \mu_i^s\}) = 1$.

**Ordering of desires** It is also typical to add some structure to desires, namely that they be a partially ordered. Formally, for all $i \in N$, $\preceq_i$ is a partial order relation on the set of paths, $P$; i.e., the following conditions hold for all paths in $\Gamma$:

1. $\forall p' \in S, (p', p') \in D(p)$: the relation ip reflexive,
2. $\forall p', p'' \in p, (p', p'') \in D(p) \wedge (p'', p') \in D(p) \Rightarrow p' = p''$: the relation ip antipymmetric,
3. $\forall p', p'', p''' \in p, (p', p'') \in D(p) \wedge (p'', p''') \in D(p) \Rightarrow (p', p''') \in D(p)$: the relation is transitive.

These conditions simply assume that there is a certain degree of consistency in an individual's desires over states.

**Intentions** An intention differs from both beliefs and desires in that this mental attitude implies the individual possessing it has made a commitment to take action toward a desired end. The desired end is an event, such as "Mike buys a cup of coffee," which may be actualized by a large number of states of the world; e.g., buying at McDonalds, or at Starbucks, or alone, or with friends,

or while believing the dark roast is probably sold out. Thus, in state $s$, the object of individual $i$'s intention is an event in $\mathcal{S}$. It is not enough for an individual to simply intend some outcome. Rather, we assume that at the time an intention is formed, it is coupled with a concrete plan of action designed to achieve the desired end.

To formalize this, for each individual $i$, define an *action plan* as a function $\sigma_i : S \to A$ where $\sigma_i(s) = a_i \in A_i(s)$ indicates that when individual $i$ arrives at state $s$ she selects an act $a_i$ from the set of acts $A_i(s)$ available at that state. Since every state has a single history leading to it, action plans may be history-contingent. Notice that, as defined, the action plan indicates what act the individual will implement at every state. Of course, we do not expect the individual to have thought through a contingency plan for every state in the state space. Rather, we impose a means-ends consistency condition on $\sigma_i$ that joins the action plan to the intention.

**Condition 4** (Weak Means-Ends Consistency)**.** *Suppose individual $i$'s intention is given by $\gamma_i(s) = X \in \mathcal{S}$. Let $P_X^s \subset P$ denote all the paths in $\Gamma$ that begin at $s$ and terminate in $X$. Then $\sigma_i$ is said to be weak means-ends consistent with $\gamma_i(s)$ if at no state $s'$ along any path in $P_X^s$ does $\sigma_i^{s'}$ force actualization of a state $s''$ that is not on any path in $P_X^s$. By "force" we mean that $\sigma_i^{s'}$ indicates an act that actualizes some state outside of $P_X^s$ regardless of the acts of all the other individuals and Nature.*

**Condition 5** (Strong Means-Ends Consistency)**.** *Suppose individual $i$'s intention is given by $\gamma_i(s) = X \in \mathcal{S}$. Let $P_X^s \subset P$ denote all the paths in $\Gamma$ that begin at $s$ and terminate in $X$. Then $\sigma_i$ is said to be strong means-ends consistent with $\gamma_i(s)$ if at every state $s'$ along any path in $P_X^s$, $\sigma_i^{s'}$ forces actualization of a state $s''$ that continues along a path in $P_X^s$. By "force" we mean that $\sigma_i^{s'}$ indicates an act that actualizes some state on a path in $P_X^s$ regardless of the acts of all the other individuals and Nature.*

In other words, Condition 4 says that the individual's plan never has him unilaterally driving the world to a state from which the intended event cannot be reached. When this condition is met, it may nevertheless be the case that the world is driven to such a state. However, this will need to be the result of the acts of others and/or Nature and nothing to do with the acts of individual $i$. The strong form, Condition 5, says that individual $i$ has a plan of action by which he can gaurantee his intended even regardless of what anyone else does. There is another case which is this: no matter what $i$ does, the intended $X$ will happen. In this case, I do not think we would properly call $X$ intention.

We also need some rationality conditions that tie the preferences over paths to the action plan. This is subtle because paths are determined by the entire act profile (i.e., and not just the acts of $i$. So, how do you tie in preferences. One possiblity is to use $i$'s may have beliefs about what the other agents are going to do (remember all of this would be encoded in the states) and, based upon this, choose an action plan that implements the most preferred path possible given the plans of the others. This would then tie beliefs, desires, intentions and plans of action together.

[STOP HERE]

# 11 Groups

### 11.0.1 Group composition and existence

Often, we are interested in the individuals that comprise a group. With that in mind, define the *group composition* function $c : M \times S \to \mathcal{N}$ where $c(k, s) = G$ indicates that in state $s \in S$ the group indexed by $k \in M$ is comprised of those individuals whose indices are contained in $G \in \mathcal{N}$. Notice that, using this approach, group composition can differ across states and a given individual can belong to multiple groups in the same state. Indeed, the same collection of individuals can comprise the memberships of different groups; i.e., we can have $c(k, s) = c(k', s)$ for $k \neq k'$.

If $k$ is a potential group in state $s \in S$, then $c(k, s) = \emptyset$. Thus, $c$ maps every element of $M$ (potential or existing) in every state to some element of $\mathcal{N}$ (possibly, $\emptyset$). Yet, because $c$ need neither be injective (one-to-one) nor surjective (onto), the inverse of $c$ need not be implied by $c$ itself. However, we can still define an *inverse group composition* function as $c^{-1} : N \times S \to \mathcal{M}$ where $c^{-1}(i, s) = H$ indicates that in state $s \in S$ the individual corresponding to index $i \in N$ belongs to the groups whose indices are contained in $H \in \mathcal{M}$. We adopt the convention that if $s$ is a state in which $i$ does not belong to any group, $c^{-1}(i, s) = \emptyset$. Then, $c^{-1}$ is a well-defined function that, like $c$, is neither injective or surjective.

From the preceding setup, we see that a state elaborates all the groups which exist in it. To keep track of this, let $e : S \to \mathcal{M}$ be the group *existence* function $e(s) \equiv \{k \in M | c(k, s) \neq \emptyset\}$. Essentially, $e$ "pulls out of $s$" the groups that exist in that state. Thus, we can define the "*no-group-exists*" event as $E_\emptyset \equiv \{s \mid e(s) = \emptyset\}$. Assume that $S$ is sufficiently expressive to permit the existence of any combination of groups: for all $H \in \mathcal{M}$, $\exists s \in S$ such that $e(s) = H$. Since states also summarize mental features of individuals, there may be many states corresponding to a

particular set of existing groups.

# 12 Initial conditions

### 12.0.1 Modest social groups

It appears promising to begin with an analysis of modest social groups and then build to to more complex, formal organizations like firms. Our interest is in *modest social groups*. The conditions required for the existence of a modest social group are stated later. However, we assume that $k$, contingent upon it existing as a modest social group, has the following informally stated features:

1. It is informally constituted,
2. It consists of two or more individuals,
3. It aims to accomplish a one-dimensional end, and
4. It is one-shot.

This eliminates from initial consideration groups: 1) whose grounding conditions include a concrete explication of group principles (e.g., a contract); 2) which are not singletons; 3) whose purpose is to achieve a single goal (e.g., *take a walk* or *play a duet*, but not *engage in money laundering and kidnapping*); 4) persist beyond the completion or failure of the intended purpose. According to Modest Social Group Condition 2, existing groups have two or more members: $\forall s \in S, c(k,s) \neq \emptyset \Rightarrow |c(k,s)| > 1$.

### 12.0.2 Analytical sequence

The idea is to begin with the simplest case of an intentional group, one in which the group is constituted simply by its individuals and their relations to each other and the group. Our present interest is in seeing how far we can get in articulating some mutually suitable description of what we mean by group intentions and their associated group acts.

Therefore, assume that the initial state of the world is $s_0^* \in E_\emptyset$, a state in which no groups exist. The profile of mental features is a primitive of the model. Therefore, everyone begins with mental states $\theta(s_0^*)$. These imply a profile of intended actions $a(s_0^*)$. According to these primitives, in a fashion not yet described, some new state of the world, $s$, obtains in which the groups $e(s)$ come

into existence along with the updated mental features $\theta(s)$. Our task is to identify how these all hang together in a coherent metaphysics.

### 12.0.3 Human acts

To rule out cases of group formation via coercion, like being kidnapped by the mafia and taken to New York in the trunk of a car, we assume that group membership relies upon the classical notion of a *human act*: at the most basic level, $\sigma(s) = a_i$ implies that, in state $s$, $i$ intends act $a_i$ voluntarily in a fashion "consistent" with his or her desires – i.e., having given his choice some thought and without coercion (we will need to say more about how these features are connected later). One obvious situation that violates this assumption is $i$ finding himself limited to one act at a state $s$ such that $|A_i(s)| = 1$. To avoid this and simplify, assume that, in state $s_0^*$, all real individuals are free to join any *one* group: for all $k \in M$ and all $i \in N$, $A_i(s_0^*) \equiv \{a_i^{1+}, \ldots, a_i^{m+}\}$.

Note that we have not said anything about the conditions required for group existence. For example individual $i$ intending the act of joining group $k$, intention $\sigma_i(s_0^*) = a_i = k^+$ is, presumably, necessary but not sufficient to cause a state to arise, $s'$, such that $k \in e(s')$.

### 12.0.4 Discussion

Although we have still said nothing about how modest social groups come to exist, have group-level intentions or take group actions, we do have the machinery to say a number of things in a precise way. Here are some examples:

1. At $s$, $i \in N$ knows that the collection of groups $\Gamma$ exist: $\mu_i(s)(\{s' | H \subseteq e(s')\}) = 1$.

2. At $s_0^*$, the collection of individuals $G \in \mathcal{N}$ each intend to join group $k$: for all $i \in G$, $\sigma_i(s_0^*) = k^+$.

3. The event that the collection of individuals $G \in \mathcal{N}$ each intend to join group $k$: $E_{G \to k} \equiv \{s \mid \forall i \in G, \sigma_i(s) = k^+\}$.

4. In state $s_0^*$, $i \in N$ knows all the members of $G$ intend to join $k$: $\mu_i(s_0^*)(E_{G \to k}) = 1$.

5. The *event* that $i \in N$ knows that the individuals $G$ intend to join $k$: let $\bar{E}_i(s)$ denote the support of $\mu_i(s)$. Then, $K_i(E_{G \to k}) \equiv \{s \mid \bar{E}_i(s) \subseteq E_{G \to k}\}$, where $K_i$ denotes events

determined by what $i$ knows in their states. Thus, $K_i(E_{G\to k})$ is the collection of states in which, given $\mu_i$, $i$ knows $E_{G\to k}$.

6. It is *evident* to the individuals $G$ that they each intend to join $k$: For all $i \in G$, $E_{G\to k} \subseteq K_i(E_{G\to k})$. It can be shown that this implies $E_{G\to k} = K_i(E_{G\to k})$.

7. $E_{G\to k}$ is *common knowledge* at $s \in S$ if and only if there exists an event $E$ such that: $s \in E$ and, for all $i \in N, E \subseteq K_i(E)$ and $E \subseteq K_i(E_{G\to k})$. This is the **?** formulation, which is a restatement of **?** in terms of evident events. For example, $E$ can be the event "The individuals $G$ publicly and credibly announce their intention to join $k$." This announcement is evident to everyone (for all $i \in N, E \subseteq K_i(E)$) and, once it occurs, it implies that everyone knows the individuals $G$ will act to join $k$, knows that they know, that they know that they know that they know, etc. (for all $i \in N, E \subseteq K_i(E_{G\to k})$). Note that $E_{G\to k}$ is not necessarily evident knowledge: it is possible to have some state $s \in E_{G\to k}$ in which not everyone knows $E_{G\to k}$.

8. In state $s_0^*$, the individuals $G$ agree that being in $k$ is most desirable: For all $i \in G$ and all $s, s' \in S$ such that $k \in e(s)$ and $k \notin e(s')$, $s' \prec_i s$.

## 13 Group formation

Since we only have in mind such simple group activities as"we take a walk to NYC" we can think of a fairly simple sequence of acts and consequences that appear to be implied by them. Let us roughly follow (**?**, Ch. 2) to see how this setup relates.

Beginning with Section 1, "I intend that we $J$, and circularity." Let $B \subset N$ be a collection of individuals. For each individual $i \in B$, assume $a_i^* \in A_i(s_0^*)$ is the act that $i$ transports herself to NYC. Let $E_i^* \subset S$ be the event "$i$ is in NYC" and $E^* \equiv \cap_{i\in B} E_i^*$ be the event that all the individuals in $B$ are in NYC. Assume $E^*$ is nonempty and that the members do not start out in NYC: $s_0^* \notin E^*$. Then, the following are some things that Bratman says are *not* a group intention to go to NYC:

1. Each individual in $B$ intends to go to NYC: $\forall i \in B, \sigma(s) = a_i^*$.

2. Each individual thinks being in NYC is the best thing: $\forall i \in B, s' \in E_i^*, s \notin E_i^*, s \prec_i s'$.

Then, Bratman suggests that the key is framing the group intention as "we each intend that we go to NYC." This is where we run into problems because what is being "intended" is vague and, in

any event seems to be doing too much lifting. In our framework, an individual can intend his or her own acts – full stop. They cannot intend the intentions or actions of others. In our construction, Bratman's sentence of intention is nonsensical.

While Bratman does indicate that "each of us has the ability to pick out the other participants," [p. 41], I think he leaves out a crucial step: the act of group formation. My sense is that if we make this explicit, we can actually make better headway. The following set of conditions for group formation is incomplete:

1. In $s_0^*$, the individuals in $B$ jointly intend to bring a group $k$ into existence to go to NYC. This requires several sub-conditions:

   (a) A profile of intentions such that, for all $i \in B$, $i$ intends to join $k$ ( $\sigma_i(s_0^*) = a_i^{k+}$) and, for all $j \notin B$, $j$ does not intend to join $k$: $\sigma_j(s_0^*) \neq k^+$.

   (b) Group existence conditions are now required, such as that the individuals each prefer states in which $k$ contains exactly the individuals $B$ to any other state: for all $s, s' \in S$ such that $c(k, s) = B$ and $c(k, s') \neq B$, $s' \preceq_i s$. The idea is that, since the existence of this kind of group simply requires everyone's assent, $i$ won't remain in the group if the composition is not to her liking. But, to be complete, this needs another condition because we don't know what happens when individuals outside of $B$ also decide to join $k$. For example, although $s$ is preferred to $s'$, $s'$ may be preferred to any other state. In that case, $c(k, s')$ could, presumably, come to exist.

2. $E_{B \to k}$ (the joint intentions of $B$ to form $k$) is common knowledge in state $s_0^*$.

3. Following the intended acts, a new state of the world $s$ occurs in which $B$ forms $K$: $c(k, s) = B$.

4. In state $s$, the existence and composition of $k$ is common knowledge.

5. Once the group forms, there must be a plan to get the group to NYC. This is where the idea of group awareness may prove helpful. We may also need to add in structure for planning within groups. This end must be joined to the intentions, beliefs and preferences at play in $s_0^*$ to make everything hang together.

Once the preceding is sorted out, we can start talking about individuals intending and acting from a state of group existence. Thinking about this second part is the next challenge.

### 13.1 Group Awareness

**Example: awareness of corporate culture"** Let us illustrate the difference between zero-probability events and unawareness in a group context. Suppose we pick up the action at $t = 1$. Amee is the customer service manager who reports directly to Bob, who is the owner of Intentional Products, Inc. a business-to-consumer firm. Amee conceives of a state of world in $t = 2$ in which her group develops a "culture of customer-service excellence." Label this state $s_1^*$. The other possibility is a state in which IPI delivers a level of customer service consistent with industry standard practices, state $s_1$. Assume $S_2^0 = \{s_2^*, s_2\}$.

If Amee is successful, the firm is more profitable: $\pi(s_2^*) = 80$, whereas $\pi(s_2) = 60$. Suppose Amee incurs a personal cost of effort if she attempts to establish the culture of excellence of $c = 15$. Furthermore, if she does attempt it, state $s_2^*$ actualizes with certainty. If Amee does not attempt it, then she incurs no personal cost and $s_1$ (the status quo) actualizes with certainty. Clearly, Amee will not attempt it unless she receives some bonus payment $B \geq c$ in the event of success.

Suppose that Bob, being a member of IPI, is awaren of the possibility of a "culture of customer-service excellence." Thus, both Amee and Bob are aware of reality fully elaborated: $S_2^A = S_2^B = S_2^0$. Although Bob is aware of $s_2^*$, he believes the actualization of this state is impossible (zero probability). Amee believes the two states happen with equal probability.

Suppose Amee proposes a bonus of $B = 35$ if successful and nothing otherwise. Under this deal, Amee's expected payoff is

$$\begin{aligned} Pr_A(s = s_2^*) \times (B - c) + Pr_A(s = s_2) \times (0) &= 0.5 \times (35 - 15) + 0.5 \times (0) \\ &= 10 \end{aligned}$$

since she gets nothing under the status quo state $s_2$. Bob's expected payoff is

$$\begin{aligned} Pr_B(s = s_2^*) \times \big(\pi(s_2^*) - B\big) + Pr_A(s = s_2) \times \pi(s_2) &= 0 \times (80 - 35) + 1 \times 60 \\ &= 60. \end{aligned}$$

It is worth pointing out that there is no feasible deal (i.e., one constrained by the funds available under each state) that makes either individual strictly better off. Bob is happy to agree to give Amee all the surplus in $s_2^*$ because he believes it is impossible.

Now, switch things up and assume that Bob is outside the organization – say, Bob is a venture capitalist dealing with Amee the entrepreneur. Furthermore, assume that, exactly because Bob is

outside the organization, Bob is unable to conceive of what a "culture of customer-service excellence" might look like in IPI. Alternatively, Bob may be able to conceive of it, but realize that there is no way he could verify it as an outsider – either way, a problem arises.

Then, from Bob's perspective, $s_2$ is the *only* possibility. Thus, Bob expects to earn $\pi(s_2) = 60$ with certainty. He will not agree to a deal in which Amee is paid a bonus. First, he cannot conceive of (or, alternatively, verify) what Amee is talking about – which precludes any contractual agreement around this idea since, presumably, he will not agree to put clauses in a contract the meaning of which he does not understand. Second, even if one could get past the first problem, from his perspective, any payment to Amee (should conditions arise that he would have to make one, even though he does not grasp Amee's idea) would come out of his 60 which, sans any deal with Amee, he is fully guaranteed.

The point of the example, then, is that awareness within an organization permits a wide range of state-contingent arrangements – even when individuals disagree on the probabilities. Indeed, divergent beliefs may be the *source* of internal organizational opportunities such as exemplified here. Moreover, awareness may span a wide range of state-cintingent organizational phenomena, such as organizational intentions, plans, conditions, etc. Those outside the organization do not share in this awareness. This limits the sorts of state-contingent arrangements outsiders can make with an organization. It also limits the extent to which the inner workings can be understood by outsiders (e.g., competitors). In this example, a "culture of customer-service excellence" may be viewed by outsiders as the "secret sauce" that allows IPI to outperform the industry