Intentional Awareness[1]

Brian Epstein
Tufts University, Medford

Michael D. Ryall
University of Toronto

January 9, 2022

[1]

1 Introduction

This note presents the full mathematical description of the intentional awareness model. It is not meant to be a paper. Rather, it is a full elaboration of a model that can be summarized or referred to in a paper. Some discussion of how certain mathematical objects are intended to be interpreted is provided (though, these descriptions are not at the same level of detail required for a paper).

1.1 Overview

In what follows, we develop a four-phase model of intentional acts. The essential aim of this formalism is to take seriously the cognitive constraints we face as finite, material beings. 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. Under constrained awareness, intentions take on an important role that is distinct from beliefs and desires.

Our approach refines some existing discussions on this topic by distinguishing between states and acts. A state is a snapshot of the world at a given moment in time that describes the status of all the features that are relevant to the situation at hand. An act is a procedure that unfolds over time. Starting in a state of the world at time t, the willful acts of individuals and the brute acts of Nature jointly determine the state of the world at time $t+1$. This interaction is elaborated in the following sections.

Acts include both efforts that are inherently invisible to others (i.e., mental activities such as deliberating, judging, and choosing) and those that are observable (e.g., enrolling in a graduate course). We refer to the latter as *actions* to distinguish them from the sorts of acts that can only be observed by the acting individual. Thus, actions are a subcategory of act. Because states of the world include cognitive attitudes, all forms of act have the power to influence future states of the world.

An individual in our model proceeds from an initial state of the world at time t to a future action according to the following sequence of phases. Each phase is assumed to take *at least* one unit of time. During a phase, the individual and Nature may act, thereby bringing the world to a new state. Individuals recall their experiences from earlier phases in later phases.

1. **Problem Selection:** Contemplating their awareness, beliefs, knowledge, and preferences as

featured in the state at t, individuals identify the set of act-problems. An *act-problem* is an opportunity for the decision maker to achieve a desired goal by influencing the evolution of future states through her acts. The problem is to settle upon a plan by which to cause or contribute to the evolution of the world to a state in which the target goal is attained. *The output of this phase is the selection of an act-problem to solve.* Alternatively, the decision maker may choose to wait and evolve to a new state in which they may select another problem for consideration.

The world evolves to a new state.

2. **Deliberation:** Contingent upon the act of orientation to engage in an act of deliberation and given the awareness, beliefs, knowledge, and preferences featured in this new state, individuals conduct an analysis to determine which goal should be pursued. Individuals screen out infeasible and dominated goals and then rank-order the remaining ones according to their preferences. *The completion of this analysis is a conclusion about which goal to pursue.* If no goal is best, revert to a new Problem Selection phase.

   The world evolves to a new state.

3. **Judgment:** Contingent upon the goal selected as best and given the awareness, beliefs, knowledge, and preferences featured in this new state, individuals decide whether to pursue the goal. *The output of this phase is a commitment to formulate a plan to achieve the goal.* Failing that, the individual reverts to a new Problem Selection phase.

   The world evolves to a new state.

4. **Planning:** Contingent upon the commitment to plan and given the awareness, beliefs, knowledge, and preferences featured in this new state, individuals formulate a state-contingent plan of action. This plan includes the goal in the support of their beliefs (i.e., individuals believe that if the plan is implemented the goal will occur with positive probability). *The output of this phase is a plan and a commitment to activate the plan.* Alternatively, the individual may revert to a new Problem Selection phase.

   The world evolves to a new state.

5. **Acting:** Upon entering the new state, the individual checks her awareness, beliefs, knowledge, and preferences, then: i) if the state is a contingency included in the plan, then take the action

as proscribed; or ii) if not, revert to a new orientation phase. *The output of this phase is an action.* Alternatively, the individual may revert to a new Problem Selection phase.

The world evolves to a new state.

For comparison, Holdon's (p. 57) four phase characterization of a typical exercise of freedom of the will unfolds as follows:

1. **Deliberating**: Considering the options that are available, and their likely consequences; getting clear on one's own desires, and one's own prior plans and intentions; seeing how the options fit in with these desires and plans; establishing pros and cons.

2. **Judging** (deciding that): Making a judgment that a certain action is best, given the considerations raised in the process of deliberation. The upshot of the judgment is a belief.

3. **Choosing** (deciding to): Deciding to do the action that one judged was best. The upshot of this decision is an intention.

4. **Acting**: Acting on the intention that has been made, which involves both doing that thing, and coordinating other actions and intentions around it.

Comments The key differences between the two approaches are the following. First, there is a distinction between states of the world and acts which flow over time. Thus, we make explicit the idea that the world is changing as the decision maker tics through the phases. Our first phase recognizes that an individual lands in a state and, at that point, must make some sense of the situation, exercising a certain degree of discretion in organizing themselves for a deliberation. Holdon does not include such a phase. Our second phase, Deliberation, follows Holdon fairly closely. The main difference is that, in our case, the options are rank-ordered at the end of the phase but with no decision yet to advance to the planning stage. In our Judgment phase, the decision is whether or not to move on to the planning phase (which may be influenced by the evolution of states). Our Planning phase is like Holdon's Choosing phase in the sense that it constitutes the commitment to act. This is because, barring winding up in a state that falls outside the scope of the plan, the individual acts according to plan (there is no reason not to). The difference is that, in addition to the trivial case of simply choosing to do one action no matter what (as in Holdon's case), our plan can be dynamic and state-contingent. Our Acting phases are pretty much identical.

The one difference we have in mind is that all the phases are mutually exclusive in our setup except acting. That is, a person can be doing Phase 5 from a previous decision sequence while engaging in a new decision sequence.

I suggest we combine Phases 2 and 3. Separating out the Judgment phase seems important to Holden. But, I don't see what this adds and rolling in to Phase 2 puts us back to a 4-phase process. Note also that awareness, beliefs, and knowledge are evolving in every one of our phases. This seems to be in contradiction to Holden, where "beliefs" arise as a result of a judgment. (I am not even sure how to interpret this, by the way.) On the other hand, it is not clear where, exactly, in our approach, the "intention" appears. Each move to a new phase involves a conscious decision to do so. Hence, one could say, there are intentions at every step.

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. 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. 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 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.

1.2 Awareness

There are two conditions that must be met for an individual to be aware of some feature of the world.

1. The feature must be accessible to the individual for active consideration. The sources of accessible features include contemporaneous sense data, imagination, and knowledge—essentially, anything toward which an individual can mindfully attend.

2. An individual must choose to incorporate that feature into a conscious thought process.

These conditions reflect our focus on decision making, as opposed to what mental phenomena might be going on when an agent sits idly thinking with no purpose in mind. Because conscious capacity is finite, at any given moment an individual will be aware of a small subset of all the features constituting the state of the world.

In some cases, a feature of the world may force itself into an individual's awareness through sense data, such as the pilot becoming aware of an alarm screeching in the cockpit. Even though the pilot does not control the breaking-through of the noise into his consciousness when the alarm goes off, at the point it does he then has a choice as to whether to incorporate the fact of the alarm sounding into his decision process or not. If for some reason the pilot should decide that the alarm is not relevant to any of his active decision deliberations, we count him as being unaware of it—even though it remains audible (and even irritating to listen to), it is not a factor in any deliberation he is presently undertaking. Alternatively, a feature of the world may be intentionally called to mind, such as when a pilot in mid-flight calls upon his knowledge of how to navigate the jetliner. One cannot bring to mind aspects of the world that one does not know or cannot imagine, such as an airline pilot who has never been to medical school pondering the technical pros and cons of cutting-edge heart transplant procedures.

Central to our approach is the assumption that humans face extreme constraints in the number of features of the world to which they can mindfully attend in any given moment. Given these constraints and the fact that humans are constantly flooded with more information than they can effectively incorporate into any deliberation, we see that mental effort is required to keep relevant information in mind, as opposed to being required to banish unnecessary information from it. For example, unless the pilot chooses to maintain awareness of the cockpit alarm (presumably, because it is relevant to something he is trying to do), we are claiming that the fact of the alarm will

automatically fade to unawareness as part of a natural process of the pilot's cognitive architecture.

Although this strikes us as an uncontroversial position, it has non-trivial implications. In particular, an open issue in the philosophy of action is whether it is rational for individuals to make commitments to ignore new information which, properly considered, might cause them to change their future plans. From our perspective, "ignoring" information—in the sense of being unaware of it—is the baseline state of most information accessible for human cognition. Given that an individual has the mental capacity to focus upon only a tiny fraction of the world's features at any one time, the question for rationality is not whether one should reflect upon the set of all relevant information and then decide whether to brush some of it aside. Rather, it is to determine which one of the multitude of issues that are rendered mutually exclusive for the purpose of reflection (due to cognitional constraints) should be brought to into active consideration (at the expense of the benefits available from reflecting upon some other things instead).

Awareness and unawareness have long been a tricky problem for decision theorists. For example, in a standard Bayesian decision problem, unawareness of certain consequences could be modeled as zero-probability states according to the decision maker's subjective beliefs. However, Bayesian decision makers will be confounded should a subjectively impossible state occur. What then? Added to this is the problem of representing interactive decision makers with different states of awareness (i.e., when the acts of one affects the consequences of the acts of the others).

Dekel et al. (1998) demonstrate that standard state-space approaches cannot model unawareness. Schipper (2015) surveys various alternatives to modeling unawareness, including approaches from AI, logic, and game theory. We adopt a version of the framework developed in Heifetz et al. (2006) (also see Heifetz et al., 2008, 2013, for related extensions) that is both simpler, in the sense that we focus upon a single-agent decision problem, and extended, in the sense that an agent's space of awareness may vary.

2 Notational conventions

Capital letters (G, N, etc.) refer to sets and to set-valued correspondences. 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 $X \backslash Y$ 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]

3 Objective Reality

In this section, we describe the status and dynamics of reality—the world as it actually is, could have been, or might yet be along with the causes that drive it to unfold in a particular way. Once we describe the way the world works here, we move on to the next section to describe the way a single decision maker understands and thinks about it. In this setup, there are two *actors*, a human decision maker and Nature, labeled i and n, respectively.[2] Nature is included to account for a God's-eye view of the status of the world in all its richness as well as for the phenomena that occur outside the decision maker's acts that, jointly with them, determine the instantiations of the world through time. We focus on the action over a fixed period, from $t = 0$ to $t = T$. Time is indicated with subscripts.

3.1 States of the world

As outlined in the Introduction, at time t, individual i finds himself in a particular situation, which we term an *objective state of the world*. This state corresponds to objective reality, including the status *of all features of the world* in that moment. Importantly, these include both the mind-independent features of the world as well as the *mental attitude* of the individual acting in that world. Let S denote the *objective state space*. S is a (finite) set of all the objective potential states that can could be actualized from $t = 0$ to $t = T$, $T > 0$.

The states themselves are indexed by the time to which they correspond and, for a particular time, by an identifying index number. We indicate these with subscripts: $s_{t.k} \in S$ refers to the k_{th} state associated with time t. Intuitively, if there are z states associated with time t, then

[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 model.

[2] In future work, we will investigate interactions between agents and group decisions. This model lays the foundation for those extensions.

$S_t \equiv \{s_{t.1}, \ldots, s_{t.z}\} \subset S$ is the set of states that could actualize in period t. With an eye toward reducing notational clutter, arbitrary states are written without their full identifying subscripts, e.g., $s \in S_t$ or $s_t \in S_t$, when doing so introduces no ambiguities. If t is the present or some historic time, then *one* s_t is factual and the others are counterfactual. If t is some future time, then S_t elaborates all the possibilities that could occur at t.

In terms of interpretation, it does no harm to imagine that each state elaborates an uncountably infinite number of features of the world. However, we have assumed that the the number of states in S is finite. The rationale for this is two-fold. First, the number of features of the world that are relevant to an individual decision maker in a specific state is often finite (though, possibly quite large). Moreover, the number of possible instantiations of each feature may be finite or effectively approximated by a finite number of categories (e.g., profits in dollars or temperature ranges). If so, then the number of states required to elaborate all the possibilities is also finite. Second, even though we lose a measure of generality by making this assumption, doing so eliminates a substantial amount of mathematical complexity that, were it included to account for an uncountably infinite number of states, would add little in the way of philosophical insight.

3.2 Acts

In our approach the acts of Nature and the decision maker jointly cause the system to evolve from a state s_t to a new state s_{t+1}. Acts of Nature represent all the causes that, in conjunction with the act of the individual, co-determine the actualization of a particular state from an immediately preceding, previously actualized state.

For each $s \in S$, let A_s^j indicate the set of *feasible acts available to actor j in state s* with arbitrary element $a_s^j \in A_s^j$, where $j \in \{i, n\}$.[3] We adopt the convention that $A_s^j = \varnothing$ indicates that actor j has no feasible acts in state s. An *objective act profile at s* is a pair of acts, $a_s \equiv (a_s^i, a_s^n)$, one by Nature and one by the individual, respectively. The set of *all objective act profiles at state s* is $A_s \equiv A_s^i \times A_s^n$. Note the implication that the acts of i and n at s are simultaneous—that is, while i is in the process of acting, there are other things going on in the world that may also have an impact on the features of the world of interest to i. Because they are sets, the acts in each A_s^j are unique. This implies the act profiles in each A_s are also unique. The set of *all possible objective act*

[3]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.

profiles at time t is $A_t \equiv \bigcup_{s \in S_t} A_s$; and the set of *all possible objective act profiles* is $A \equiv \bigcup_{s \in S} A_s$. The uniqueness of act profiles in each A_s does not imply the same for A_t or A: an actor may have the same feasible acts in different states.

We wish to represent the uncertainty associated with Nature's acts as well as to allow for situations in which the individual randomizes over acts (e.g., i decides to "flip a coin" to determine what to do). Therefore, for each objective state $s \in S$, a *mixed state-contingent act for* $j \in \{n, i\}$ is a probability distribution over feasible acts at s, denoted $\sigma_s^j \in \Delta(A_s^j)$.[4] We write $\sigma_s^j(a_s^j)$ to indicate the probability that act a_s^j is selected at s. The use of probability distributions over feasible acts provides a nice level of generality. Typically, individuals make action choices with certainty: e.g., $\sigma_s^i(a_s^i) = 1$ for some $a_s^i \in A_s^i$. Thus, mixed acts are general in the sense that they can represent both deterministic and random behavior. Mirroring the objective act objects, define:

1. $\Sigma_s \equiv \Delta(A_s^n) \times \Delta(A_s^i)$, the *set of mixed act profiles at s* (where the membership of s in an objective or subjective state space will be clear from the context), with typical element $\sigma_s = (\sigma_s^i, \sigma_s^n)$;

2. $\Sigma_t \equiv \bigcup_{s \in S_t} \Sigma_s$, the set of *all possible mixed act profiles at t*, with typical element σ_t; and

3. $\Sigma \equiv \bigcup_{s \in S} \Sigma_s$, the set of *all possible mixed act profiles*, with typical element σ.

3.3 Dynamics

As indicated above, the act profiles summarize all the conditions required to actualize one state from the previously actualized state. It may be helpful to think of act profiles as the "flow" variables between states—the activities that occur over a unit of time that cause the world to move from one state to the next.

To formalize this, define the *dynamic objective state graph*, $\Gamma \equiv (S, \lessdot)$ which elaborates of all objectively possible sequences of states. Specifically, assume Γ is a directed, rooted tree with nodes S ordered by the precedence relation $\lessdot$; i.e., the predecessors of each $s \in S$ are totally ordered by $\lessdot$. Thus, $S_0 = \{s_0\}$, where s_0 is the root node at the beginning of time. Let $\omega : A \to S$ be a function mapping act profiles available at a given state to its children. Thus, if $s_t \lessdot s_{t+1}$, then $\omega(a_{s_t}) = s_{t+1}$ means a_{s_t} is the act profile that causes the world to evolve from s_t to $s_{s_{t+1}}$. Assume ω is bijective from A_s to the immediate successors of s (the elements in A_s uniquely label the edges

[4] Where $\Delta(A_s^j)$ is the set of probability distributions over actor j's feasible acts at s.

from s to its successors, one-to-one with no extras or shortfalls). This means there is no state in which a specific act profile can lead from one state to more than one successor state.[5] Because the relationship is bijective, we can also identify the acts leading to states; i.e., $\omega^{-1}(s_{t+1}) = a_{s_t}$.

At the beginning of time, in state s_0, i must choose one from a number of mutually exclusive act-problems. In future periods, i is generally free to continue working on the current problem or abandon it and take up another. In s_0, however, i's only feasible act is to choose a problem to get started upon. Therefore, $A^i_{s_0} = \{p_1, \ldots, p_k\}$ is the indexed set of k act-problems objectively available to i in state s_0. To refer to *the act-problems available in an arbitrary state, s*, we define $P_s \equiv \{p_1, \ldots, p_k\}$.[6]

For each state s, let $h_s = (s_0, \ldots, s)$ denote the *history at s*; i.e., the unique path from s_0 to s in Γ. Because ω is bijective on A_s, which contains no duplicates, each history h_s is also associated with the unique sequence of act profiles that actualize it. Abusing notation a bit, let this be represented by $\omega^{-1}(h_s) \equiv (\omega^{-1}(s_0), \ldots, \omega^{-1}(s))$.

Let H_T denote the set of *objective terminal histories*; i.e., of the form $h_s = (s_0, \ldots, s)$ where $s \in S_T$. Note that the relationship between H_T and S_T is bijective: every terminal state corresponds to a unique history in Γ. It is not difficult to show that each $\sigma \in \Sigma$ implies a probability distribution on the set of terminal histories. We will abuse notation some more and write $\sigma(h)$ to be the probability of terminal history h implied by σ.

Example: Brian's Infant, objective reality We consider the problem of Brian's Infant, an extended example that we will use to illustrate the formalism as we develop it. The situation is as follows. Brian's child, i, finds herself in $t = 0$ presented with two, mutually potential act-problems: to choose a toy with which to play or to choose a TV show to watch. Let the toys be labeled A and B. One of these toys is better than the other. Let A^* indicate that A is best and B^* indicate that B is best. Alternatively, the child can choose to watch TV show W or C. One of these shows is better than the other. Let W^* indicate that W is best and C^* indicate that C is best.

Let $P_{s_0} = \{p_1, p_2\}$, where p_1 is the choose-a-toy problem and p_2 is the choose-a-show problem. The idea is that each act-problem presents an actor with a distinct choice deliberation that occupies

[5] Two or more states may yet have the same the set of feasible act profiles. That is, for $s \neq s'$ it it is possible that $A_s = A_{s'}$. For example, at 11am i may be finished with work and get a cup of coffee or, alternatively, may not be finished with work and, yet, still have the option to get a cup of coffee. However, it is never the case—holding Nature's acts constant—that getting a cup of coffee in one state has an ambiguous effect on the future.

[6] In later periods, $t > 0$, A^i_s may contain numerous acts in addition to choosing a problem from P_s. Hence, the distinction is helpful.

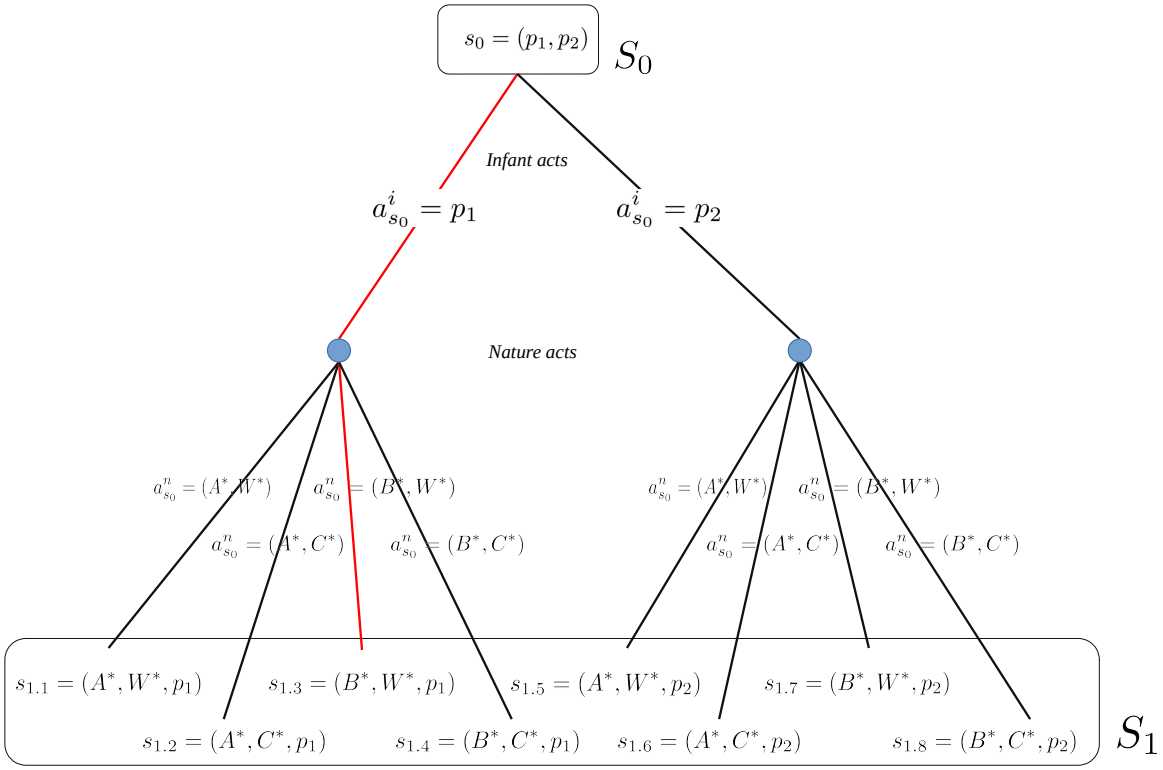

Figure 1: Evolution of Nature's state spaces from $t = 0$ to $t = 1$

their full awareness capacity. That is, cognitive constraints render the elements of P_{s_0} mutually exclusive. Here, i's feasible acts in state s_0 are $A^i_{s_0} = \{p_1, p_2\}$. At the same time, Nature determines which toy and which show are best. Therefore, Nature selects one of among four possible acts: $A^n_{s_0} = \{(A^*, W^*), (A^*, C^*), (B^*, W^*), (B^*, C^*)\}$.[7]

Figure 1 illustrates the evolution of the system from $t = 0$ to $t = 1$. From s_0, the objective action profiles are of the form $a_{s_0} = ((Y, X), Z)$ where $a^n_{s_0} = (Y, X)$ is Nature's choice of Y, the best toy, and X, the best show, and $a^i_{s_0} = Z$ is i's choice of which act-problem to solve. Then, S_1 contains eight possible states (one for each combination of these variables, as shown).

Here we illustrate the notational convention of identifying states by time and index number by using "$t.\#$" subscripts. For example, $s_{1.3} = (B^*, W^*, p_1)$ is state number 3 in period 1. In this state: B is the best toy; W is the best show; and i has decided to focus on choosing the toy with which to play. This state was actualized by the act profile $a_0 = ((B^*, W^*), p_1)$. The mapping

[7] When a^j_s is a compound act, as is the case with Nature in this example, we use parentheses to list the individual elements of the act.

function with respect to this state is $\omega((B^*, W^*), p_1) = s_{1.3}$. The history associated with this state is $h_{s_{1.3}} = (s_0, s_{1.3})$. This history corresponds to an act profile sequence that includes a single element: $\omega^{-1}(h_{s_{1.3}}) = (a_0)$ where $a_0 = ((B^*, W^*), p_1)$.

3.4 Events

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 probability theory usage) as *synchronic events*, and those associated with properties extended through time (the philosophy usage) as *diachronic events*.

We define synchronic events as subsets of a state space at a moment in time. For example, the event "Mike intends to get a cup of coffee at t," includes *all* the states in S_t 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.

In the Brian's child example in Figure 1, the synchronic event "A is the best toy," is the subset $\{s_{1.1}, s_{1.2}, s_{1.5}, s_{1.6}\} \subset S_1$. Alternatively, the event, "i is solving p_1," is $\{s_{1.1}, s_{1.2}, s_{1.3}, s_{1.4}\}$. Thus, if i is aware of all the states in S_1 and knows that "A is the best toy," and that she is solving p_1, then she knows that one of $\{s_{1.1}, s_{1.2}\}$ is the actual state of the world.

Diachronic events are defined as subsets of the set of terminal histories, H_T. These refer to sequences of events over time and, implicitly, the sequences of action profiles that cause them. For example, the set of terminal histories in the Brian's child example is

$$H = \{(s_0, s_{1.1}), \ldots, (s_0, s_{1.8})\}.$$

The diachronic event, "i's period 0 act is $a_0^i = p_1$," is given by the subset

$$\{(s_0, s_{1.1}), (s_0, s_{1.2}), (s_0, s_{1.3}), (s_0, s_{1.4})\} \subset H_T.$$

Diachronic events that will be of interest are individual acts, sequences of act profiles, and features

of the world that persist through time—all of which are associated with sequences of states which, themselves, are associated with subsets of H. For example, the diachronic event *state s occurred* is the set of terminal histories that include state s as a component. These are

It is worth noting that synchronic events imply diachronic events. For example, the synchronic event, "A is the best toy in $t = 1$," corresponds to the diachronic event that includes all histories in which this is true.[8]

3.5 Summing up reality

Pulling together the key parts of the formalism, mind-independent reality is described by a triple $\Xi \equiv (\Gamma, A, \omega)$. These elaborate the objective state space, S, and their dynamic interrelationships, $\prec$, as specified by the dynamic objective state graph Γ, all the available act profiles, A, and the mapping from these to edges in the dynamic tree according to ω to indicate which act profiles cause transitions from one state to another. Along with A and ω, Γ describes dynamic reality in all its detail—including each potential state of the world in each period and the sequences of act profiles that are required to actualize them.

4 Awareness

We begin to structure the individual's mental attributes by assuming that—in each objective state of the world—the individual has in mind a subjective version of Ξ plus some additional cognitional components that allow him to navigate through act problems. Specifically,for each $s \in S$, let $\Xi^s \equiv (\Gamma^s, A^s, \omega^s)$ be i's mental representation of Ξ.[9] Each element in Ξ^srepresents i's subjective awareness of the corresponding element in Ξ. In particular, the elements of *awareness state space* S^s (as specified by Γ^s) are subjective states that describe the features of the world as i is aware them when i is in s. These include i's awareness of the world's features in past actualized states, the present state, counterfactuals to these, and future states. To avoid additional complexities for now, we assume that i does not reflect on his awareness[10] Thus, like S, S^s is partitioned by the subsets of states corresponding to each time period, S^s_t.

[8] A synchronic event is a subset of states at some time t. Its diachronic counterpart is the subset of all terminal histories that pass through those states.

[9] Our notational convention going forward is that state superscripts indicate the subjective objects that correspond to the indicated objective state.

[10] That is, in this analysis we do not raise the issue of awareness of awareness. This will be an important issue in the multi-agent case, but we sidestep it for now.

The awareness graph Γ^s describes the individual's subjective organization of reality when he finds himself in objective state s. This construct is part of the reality elaborated by s. This raises the central question of how (ontological) reality, as described by Γ, corresponds to i's (epistemological) perception of it in a given state, as described by Γ^s. What we assume is that i is aware of a limited but accurate version of reality: although i may not be aware of all its features, those of which he *is* aware are correct.

To formalize this, for each objective state $s \in S$, define the surjective *state-s awareness mapping* $r^s(\cdot) : S \to S^s$ where $s'' = r^s(s')$ means that—when i is in objective state $s \in S$—the subjective state $s'' \in S^s$ describes the features of the world of which he is aware regarding some other objective state $s' \in S$. For any time t, we assume that r^s is surjective (onto) from S_t to S_t^s. This means that the elements of S_t^s typically represent a coarsening of reality.

Example *To see the idea, suppose $s \in S_t$ is a state in which i's coffee is too hot to drink and this is the only feature of reality of which i is aware. Presumably, S_t is huge, with each state capturing features of the world beyond the temperature of the coffee—such as the weather, where i is geographically located, what is on TV, the level of gas in the car, and so on, many of which feature coffee that is too hot to drink. Then, r^s projects all the states in which the coffee is too hot to drink into a subjective awareness state, say $s' \in S_t^s$. Now, suppose i is also be able to reason about (is aware of) a state in which the coffee is not too hot to drink. Then, r^s would project all the other states into a counterfactual subjective awareness state, say $s'' \in S_t^s$, in which this is true. Because r^s is surjective when restricted to S_t, the inverse $r^{s^{-1}}$ exists on this part of its domain: e.g., $r^{s^{-1}}(s') \subset S_t$ is the objective synchronic event, "the coffee is too hot to drink." Suppose $s''' \in S_t$ is an objective state in which the coffee is not too hot to drink. Then, $s'' = r^s(s''')$.*

Here, $s' = r^{s_t}(s_t)$ means that s' elaborates only the features of the world of which i is aware in t given the actualization of objective state s_t in that period; it is what i is aware of regarding his current state when that state is s_t. Here, it is important to point out that each Γ^s is considered to be a distinct mental (and mathematical) object. Therefore, we often add decorations of some kind to distinguish states in one awareness graph from another. For example, we might write $s'_0 = r^{s_0}(s_0)$ to emphasize that the features of which i is aware when he is in s_0 is captured by the awareness state s'_{s_0}. Even if i is perfectly aware in objective state s_0, we would not write $s'_0 = s_0$ because this would amount to equating apples and oranges: the former represents a mental construct while the

latter represents reality. This rule also holds for comparisons across subjective awareness graphs. Suppose i is in a second-period state, say $s_{2.k}$, reflecting back on what he knew about the world in period 0: $s_0'' = r^{s_{2.k}}(s_0)$. Even if his recollection of his awareness of s_0 in s_0 is exactly the same in $s_{2.k}$, it is still technically incorrect to write $s_0'' = s_0'$. The mental constructs of i in s_0 and $s_{2.k}$ are distinct, even though the *contents* of some of their elements may be identical.

Example *For example, suppose $s, s'' \in S$ are objective states in which it is raining and sunny outside, respectively. Further, assume that in both states, i is aware of a cat meme on his computer, but not the weather. Then, $s' = r^s(s) \in S_t^s$ is the subjective state in which i is aware of a cat meme on his computer. Here, it is also the case that $s' = r^s(s'')$: being in state s, i cannot distinguish his present actualized state in which it is raining from the counterfactual state in which it is sunny—the weather is simply not on his mind. In the fully-elaborated objective world, $r^{s^{-1}}(s')$ is the synchronic event "i is aware of a cat meme"—i.e., all the states in which this is true. Now, in the present state s, i may also be aware that, tomorrow, he will be engrossed by a new meme—which must correspond to a different subjective state, say $s''' \in S_{t+1}^s$.*

At this point, we have said: i) the collection of states in the individual's awareness graph correspond to a partition of the objective state space; and ii) this correspondence maintains time consistency in the sense that real world states at t never map to awareness states in some other period. We have said nothing about consistency with respect to feasible acts or dynamic consistency.

The structure of feasible act profiles at awareness state $s' \in S^s$ is assumed to be $A_{s'}^s \equiv A_{s'}^{s.n} \times A_{s'}^{s.i}$, where $A_{s'}^{s.j}$ is the set of feasible acts available to actor j at s' of which i is aware when i is in objective state s. The product structure implies *Act Independence*: the set of act profiles includes *all combinations* of the feasible acts for n and i at s' of which i is aware. As in the objective case, for each $s' \in S^s$, we require ω^s to be bijective (one-to-one and onto) from $A_{s'}^s$ to the immediate successors of s'.

We impose two consistency conditions on the relationship between A^s and A. First, if $s' = r^s(s)$, then $A_{s'}^s \subseteq A_s$. In other words, when i is in objective state s, the feasible acts of which he is aware is a subset of the feasible acts that really are available to him. We permit (though, do not require) the possibility that i is unaware of some his feasible acts. However, the feasible acts of which he *is* aware really *are* feasible. Second, if $s' = r^s(s)$, then $\omega^{s^{-1}}(h^s(s')) = \omega^{-1}(h(s))$. In other words, along the actualized history, i labels the edges with their correct act profile labels. This will come

up later when we assume that i does not forget the actualized act profiles that got him to his present state (see the discussion in the Uncertainty section).

Outside of the preceding restrictions on act profile names and mapping, i is free to label edges in Γ^s as he wishes. In many cases, edges in his awareness space will be an amalgam of edges in reality. Outside of the actualized act profiles that i has observed and those that are presently feasible to him, the names he assigns to other (i.e., counterfactual or future) edges in Γ^s is not crucial. Rather, the mapping function r is the central determinant of the relationship between reality and i's perception of it (i.e., between Γ and Γ^s, respectively). With this in mind, we impose the following consistency conditions on r^s:

1. Dynamic consistency: for $s, s', s'' \in S$ such that $s' \prec s''$, if $s''' = r^s(s')$ and $s'''' = r^s(s'')$, then $s''' \prec^s s''''$ in Γ^s.

2. No contemplation of future unawareness: for $s \in S$, if $s' = r^s(s)$, then the subgraph in Γ^s starting at s' is a tree rooted rooted at s'.

3. Awareness of counterfactuals: for $s \in S_t$, $t > 0$, if $s' = r^s(s) \in S_t^s$, then there exists at least one $s'' \in S_t^s$ such that $s' \neq s''$.

Item 1 is important as it implies that terminal histories in the real world correspond to histories in the awareness graph (i.e., as i is thinking about them). Thus, histories in H_T^s correspond to diachronic events in H_T. Item 2 is intuitive in the sense that, as i look looks to the future while contemplating how he will solve his act-problem, he does not envision becoming unaware of what he is doing as he travels along one of his potential paths. Item 3 says that, whatever i is aware of with respect to his actualized state (at any period beyond the the start of time), he is also aware that things could have turned out differently (perhaps in the very rough sense of, "there could have been a state of the world that is NOT the one I am presently in").

It is worth noting that awareness graphs need not be trees. Typically, as i moves forward in time, his awareness of certain parts of reality becomes more refined while other parts fade into unawareness (with the recall caveats described above and in the section on mental attitudes).

Finally, one interesting possibility is to impose some hard constraints on awareness graphs and actions. For example, although this might be arbitrary, we could analyze individuals capable of: i) 10 awareness histories in any Γ^4; and ii) one act per period consistent with phases I through IV in Figure 4 plus conducting acts according to an established plan (discussed below). While

the constraints so imposed would, barring findings from cognitional science, be arbitrary, they would allow us to analyze intentional unawareness in the sense of an individual trying to use the cognitional resources in an optimal fashion.

Example: Brian's Infant, Part 2 Picking up the example of Brian's Infant where we left off, let consider the infant's awareness of reality. Suppose the act profile is $a_{s_0} = ((B^*, W^*), p_1)$. According to Figure 1, this brings the world to objective state $s_{1.3} = ((B^*, W^*), p_1)$ in $t = 1$. Assume that, having decided to solve the toy-choice problem, i becomes aware of which toy is best. In this state, i can reason about B being the best toy and the counterfactual that A could have been the best toy. Here, we see that i is also aware that she could have chosen to solve p_2, which would have put her in a different state of the world. Then, $S_0^s = \{s_0'\}$, such that $P_{s_0'} = P_{s_0}$, and $S_1^s = \{(A^*, p_1), (B^*, p_1), (p_2)\}$.[11]

In this state, her awareness tree is $\Gamma^s_{s_{1.3}}$, as depicted in Figure 2. In this diagram, the top half shows Γ tipped on its side. The bottom half illustrates $\Gamma^s_{s_{1.3}}$. Several of the awareness mappings are labelled. For example, i finds herself in $r(s_{1.3}|s_{1.3}) = (B^*)$. Notice that $r^{-1}((B^*)|s_{1.3}) = \{s_{1.3}, s_{1.4}\}$, which corresponds to the objective event, "B is the best toy and i chooses to solve the indoor-play problem. The diagram labels several of the other awareness mappings associated with state $S_{1.3}$.

The central take-away from Figure 2 is that i is completely unaware of the TV show choices. In particular, it is *not* the case that i is *uncertain* about the best TV show. Rather, according to $\Gamma^s_{s_{1.3}}$, i is simply not thinking about TV shows at all—they are not in her field of awareness and, therefore, are not a factor in her future deliberations.

We can also check the conditions on the awareness mapping. Condition 1 is met because each objective state $s \in S_t$ maps to a state in S_t^s. Condition 2 is met because all combinations of the acts of which i is aware are included as act profiles in $\Gamma^s_{s_{1.3}}$.[12] Condition 3 is clearly met. Condition 4 is met, which can be seen from the fact the terminal states in S^s imply a partition of the terminal states in S_1 (and, hence, of H). Finally, Condition 5 is met as well: i is able to recall her own act p_1, which co-actualized her present state along with Nature's act (of which she is partially aware).

[11] Keep in mind, the states in i's subjective, state-contingent awareness trees are mathematically distinct from each other and from the objective states. So, we write $S_0^s = \{s_0'\}$ rather than $S_0^s = \{s_0\}$ even though $P_{s_0'} = P_{s_0}$ because s_0' is not the same object as s_0

[12] The acts of which i is aware are $\{p_1, p_2\}$ for herself and $\{A^*, B^*\}$ for Nature. The set of act profiles is $\{A^*, B^*\} \times \{p_1, p_2\}$. These are all included as branches in i's awareness tree.

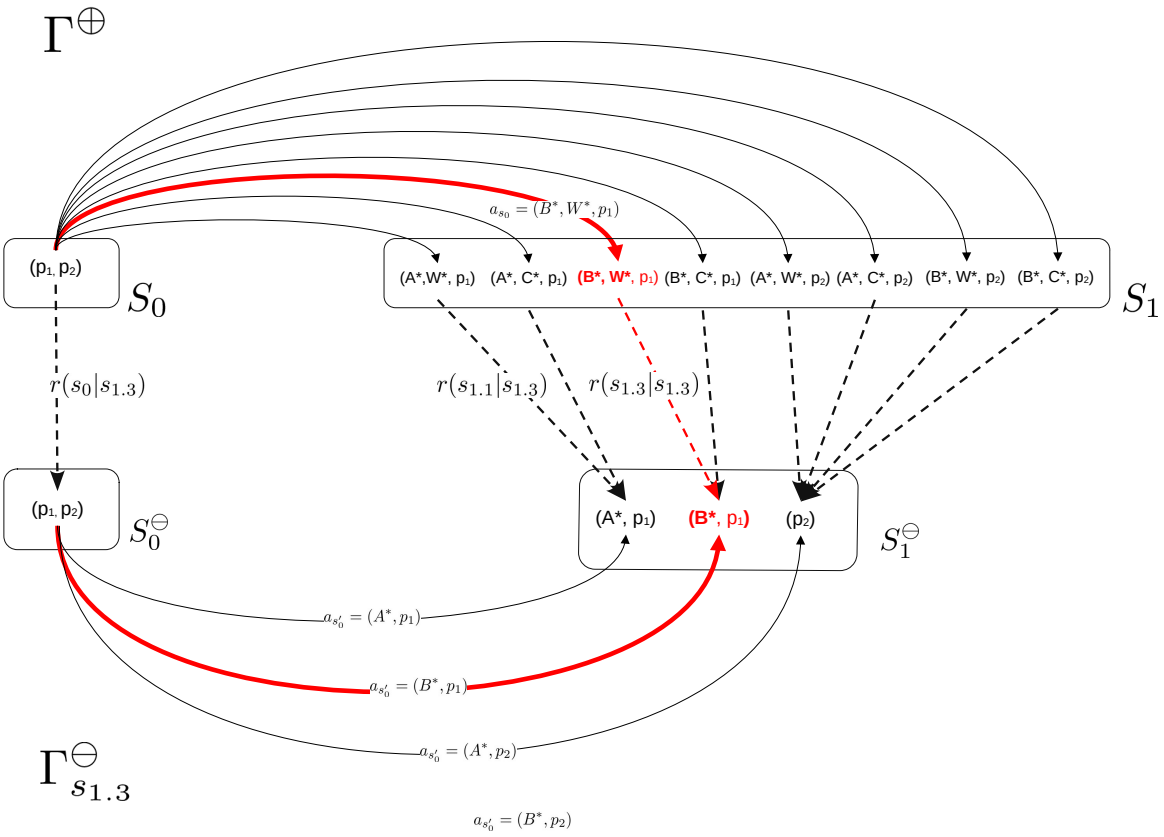

Figure 2: Awareness tree for Brian's Infant in objective state $s_{1.3} = ((B^*, W, p_1))$.

5 Mental attitudes

5.1 Uncertainty

In addition to the states about which i is aware, we wish to account for uncertainty. Uncertainty is not the same as unawareness. For example, in the Brian's Infant example, i could be aware that A^* and B^* are possible yet harbor uncertainty about which is consistent with the true state of the world—and all the while be utterly unaware of the weather.

5.1.1 Information sets

For each objective state $s \in S$, define the *information set at subjective state* $s' \in S^s$, denoted $I^s_{s'}$, to be a subset of S^s_s. Information sets serve to distinguish states about which i is aware but uncertain.

Information sets are assumed to meet the following conditions:

4. No-delusion: $s' \in I^s_{s'}$.

5. Introspection: if $s'' \in I^s_{s'}$, then $I^s_{s''} = I^s_{s'}$.

6. Feasible act consistency: For all $s', s'' \in S^s$, if $A^{s,i}_{s'} \cap A^{s,i}_{s''} \neq \varnothing$ then $A^{s,i}_{s'} = A^{s,i}_{s''}$.

7. Distinct acts at disjoint information sets: if $s', s'' \in S^s_t$ and $A^{s,i}_{s'} = A^{s,i}_{s''}$, then $I^s_{s'} = I^s_{s''}$.

8. Time consistency: if $s', s'' \in I^s_{s'}$ and $s' \in S^s_t$, then $s'' \in S^s_t$.

Condition 6 is self-explanatory. Condition 7 prevents states from being in more than one information set (which would make no sense: if i is uncertain about whether s' or s'' is true and about whether s'' or s''' is true, then he should also be uncertain about whether s' or s''' is true). Condition 8 partitions the states in S^s_t into equivalence classes according to the acts available to i. Condition 9 implies that i can distinguish between states based upon differences in their feasible acts. Condition 10, which is similar to Condition 1 for awareness, ensures that all the states in a given information set are possibilities at a specific time t. The idea is that i is always aware of the time and, therefore, never places any weight on the proposition that he is presently in a state occurring at a time different than the one he is in.

Notice the implication that, for each $s \in S$, a state space in a given period, S^s_t is partitioned by its information sets. Moreover, since the time-stamped state spaces $S_{t,s}$ themselves partition S_s, the information sets also partition S_s. Therefore, let $\mathcal{I}^s$ denote the collection of all information sets associated with Γ^s. This allows us to refer to an arbitrary information set simply as $I \in \mathcal{I}^s$. Then, because Condition 9 says that all states in the same information set have the same feasible actions (of which i is aware), we can write A^i_I without ambiguity.

Example: Brian's Infant, Part 3 Information sets are illustrated for the Brian's Infant example in Figure 3. In this scenario, i is in objective state $s_{1.3}$ and has the same awareness as in Figure 2. Now, information sets I and I' have been added to illustrate the case in which i is uncertain about whether A or B is the best toy. These are indicated by the dashed lines connecting states. I indicates that i knows she is working on p_1 but uncertain about which toy is best. She is aware that should could have known that she was working on p_2 had she chosen that problems (according to I').

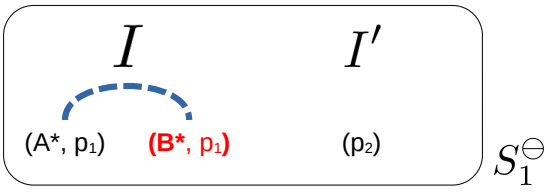

Figure 3: Partially aware infant is also uncertain about which toy is best

When an information set is not a singleton (i.e., i is uncertain about something), we assume she has beliefs about which state is true that are represented by a probability distribution on the states in the information set. We use ρ^i to indicate these assessments. In the preceding example, assuming $s_t = (C, A^*)$ is understood from the context, $\rho^i_{0.1}(R, A^*)$ is the probability i assigns to the possibility that the true state is (R, A^*).

5.1.2 Beliefs

Given an objective state $s \in S$, let $\Delta(H^s_T)$ denote the set of all probability distributions on the set of subjective terminal histories. Then, $\mu^s \in \Delta(H^s_T)$ is i's belief about how dynamics play out, with $\mu^s(h)$ indicating the probability i places on terminal history $h \in H^s_T$. Because act profiles and states imply events in H^s_T by the terminal histories that pass through them, μ^s can be used to compute conditional probabilities. In particular, when i is in subjective state $s' \in S^s$, the probability i assigns to s' is $\mu^s(s'|I^s_{s'})$. We adopt the convention that if $\mu^s(I_{s'}) = 0$, then $\mu^s(s'|I^s_{s'}) = 0$.

For example, consider S^s_1 in Figure 2 in which each state is in its own, singleton, information set. Then, $\mu^{s_{1.3}}((B^*, p_1)|(B^*, p_1)) = 1$: i knows she is at (B^*, p_1). Instead, suppose the situation is identical except that the information sets are as shown in Figure 3, with $\mu^{s_{1.3}}$ assigning equal probability to both states in I. Then, $\mu^{s_{1.3}}((A^*, p_1)|I) = \mu^{s_{1.3}}((B^*, p_1)|I) = 0.5$.

For now, it is sufficient to understand that they are probability distributions on terminal histories and that they may vary state by objective state.

5.2 Desires

For all $s \in S$, define the state-dependent *desire relation* $D^s \subset H_T^s \times H_T^s$ where, $(h, h') \in D^s$ means that i in objective state s desires the subjective terminal history h' at least as much as the history h. We use the intuitive notation $h \preceq^s h'$, which is defined to mean $(h, h') \in D^s$. We use $\prec^s$ and $\approx^s$ to indicate strict preference and indifference, respectively. Assume that each D^s is complete (in the sense that any pairs of histories $h, h' \in H_T^s$ are comparable) and transitive. Then, the desire relation D^s can be represented by a numerical function $d^s : H_T^s \to \mathbb{R}$, where $d^s(h) < d^s(h')$ and $d^s(h) = d^s(h')$ if and only if $h \prec^s h'$ and $h \approx^s h'$, respectively.

Note the implication that desires may change from objective state to objective state. The objective state provides a fully-featured snapshot of the world, including the status of i's cognition. The desire function says that i is capable of comparing the desirability of complete histories—up to his awareness of the world according to his present, objective state.

We wish to impose some consistency between i's desires over the terminal histories as he is aware of them in a given state and his desires over the objective histories underlying them.[13] To this end, let $d : H_T \to \mathbb{R}$ be the desire function for i under conditions of perfect awareness (i.e., a situation in which i is aware of all the terminal histories and, hence, can assess each one). Then, we assume that, for all $s \in S, h \in H_T^s$, d^s meets the following consistency condition:

9. Let

$$x = \min_{h' \in r^{-1}(h)} d(h'),$$

and

$$y = \max_{h' \in r^{-1}(h)} d(h').$$

Then $d^s(h) \equiv \alpha x + (1 - \alpha)y$ for some $\alpha \in [0, 1]$.

We imagine that i is equipped with primitive desires (d) at the beginning of time which would permit him to compare his desires with respect to any two objective histories, *were he aware of them all* (H_T). When he is not perfectly aware, each history of which he is aware corresponds to an objective diachronic event. THus, the desire function d implies an interval of values $[x, y]$ for each such event (which may be trivial in some cases—i.e., $x = y$). Then, the structure of d^s means

[13] Remember, the histories of which i is aware at any point correspond to collections of objective histories, the latter being the most refined. Although i's awareness limitations coarsen his perception of reality, the consequences of his actions will correspond to the actual unfolding of a specific objective history. This raises the question of how consistent with reality his desires should be.

that i's desires at lower levels of awareness provide an implicit ranking of the objective intervals determined by d. Notice that the α parameter does not vary by state. It is a constant primitive of i's desires from state to state.[14] That said, as i's awareness changes from objective state s to objective state s', the partitions on H_T implied by H_T^s and $H_T^{s'}$ also vary, thereby implying the possibility of variation in the structure of desires from d^s to $d^{s'}$.

Again, i's desires over objective histories under perfect awareness (d) can be thought of as a cognitive primitive that assigns numerical intervals to the subjective histories under imperfect awareness, Then, in objective state s, the ranking of subjective histories provided by d^s is consistent with a ranking function on these intervals. This is the key consistency condition. One interpretation is that this structure ensures that i's desires under unawareness do not depart too far from reality That is, the individual's valuations of awareness histories are required to be roughly aligned with the evaluations of the objective histories with which they are associated. This alignment is systematic (α is constant on S) and within the bounds of the values that could actually happen ($0 \leqslant \alpha \leqslant 1$).

Why consider desires over histories? 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.

5.3 Intentions

For all $s \in S$, define the state-dependent *intention* for the individual as $\gamma^s \subseteq H_T^s$, where $\gamma^s = E$ means that in objective state s individual i intends subjective event E. We assume that individuals

[14] That is α does not vary by s.

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 $\gamma^s = \varnothing$ 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 $\gamma^{s_t} = \varnothing$ to $\gamma^{s_{t+1}} = E \neq \varnothing$. 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 an objective state in which i does not have an intention, then the set of objective feasible acts, $A^i_{s_t}$, can include an *act to form the intention* (e.g., to "get a cup of coffee"), which would then take him to a state s_{t+1} in which $\gamma^{s_{t+1}} = E$ where E contains all the states consistent with i's intention (e.g., having a cup of coffee).

Summing up, in each objective state, individual i's *mental attitudes* are summarized by a triple denoted $\theta^s \equiv (\mu^s, D^s, \gamma^s)$. In setting up mental features in this way, we are following a version of the familiar "type-space" approach used in game theory (See Harsanyi, 1967; Mertens and Zamir, 1985) in which θ_s is i's mental type in objective state s.

Thus, i's mental life in state $s \in S$ is fully described by Ξ^s and θ^s.

Object	Description	Comments
S	All objective states of the world	S contains all possible states
A_s^i	i's feasible acts in state $s \in S$	$a_s^i \in A_s^i$, arbitrary act by i in s
a_s	A profile of acts in s, $a_s \equiv (a_s^n, a_s^i)$	
A_s	All act profiles in s	
h_s	History at $s \in S$	
$\mathcal{H}_t$	Set of all subsets of histories at t	

Table 1: Notation Reference (NEEDS TO BE UPDATED AND EXPANDED)

6 Notation Reference

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

7 Four-Phase Model of Action

With the previous setup in place, we can now elaborate the model discussed in the Introduction. The phases are: 1) Problem Selection; 2) Deliberation; 3) Planning; and 4) Acting. Each phase requires completion of the preceding phases before it can begin. As we elaborate the phases, we also add certain consistency conditions which are pertinent to them.

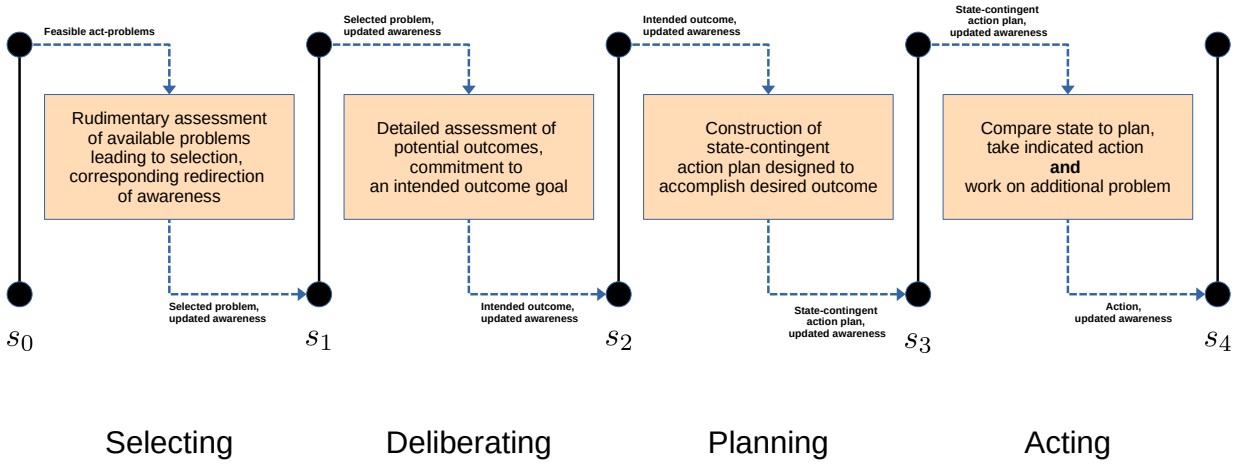

Figure 4: Overview of states and decision phases.

The objective tree Γ elaborates all the ways in which the world can evolve. Consistent with Γ, H contains all the terminal histories, each of which elaborates a particular evolution from the beginning state, s_0, to some terminal state T periods later. The elements of A describe the feasible act profiles at every state. The act mapping ω connects the act profiles available in one state lead to the new states in the following period which they actualize. See Figure 9 for a summary.

Brian's Infant, Part IV Figure 5 illustrates Γ for the Brian's Infant example. Here, we introduce some further details. First, to illustrate the four phases of the decision-act process, we extend the tree through $T = 4$. Next, we add the details of what could happen at each phase. At $t = 0$, i decides whether to solve p_1 or p_2. At the same time, Nature fixes the best toy and the best TV show. For now, we assume Nature does not act until the penultimate period.

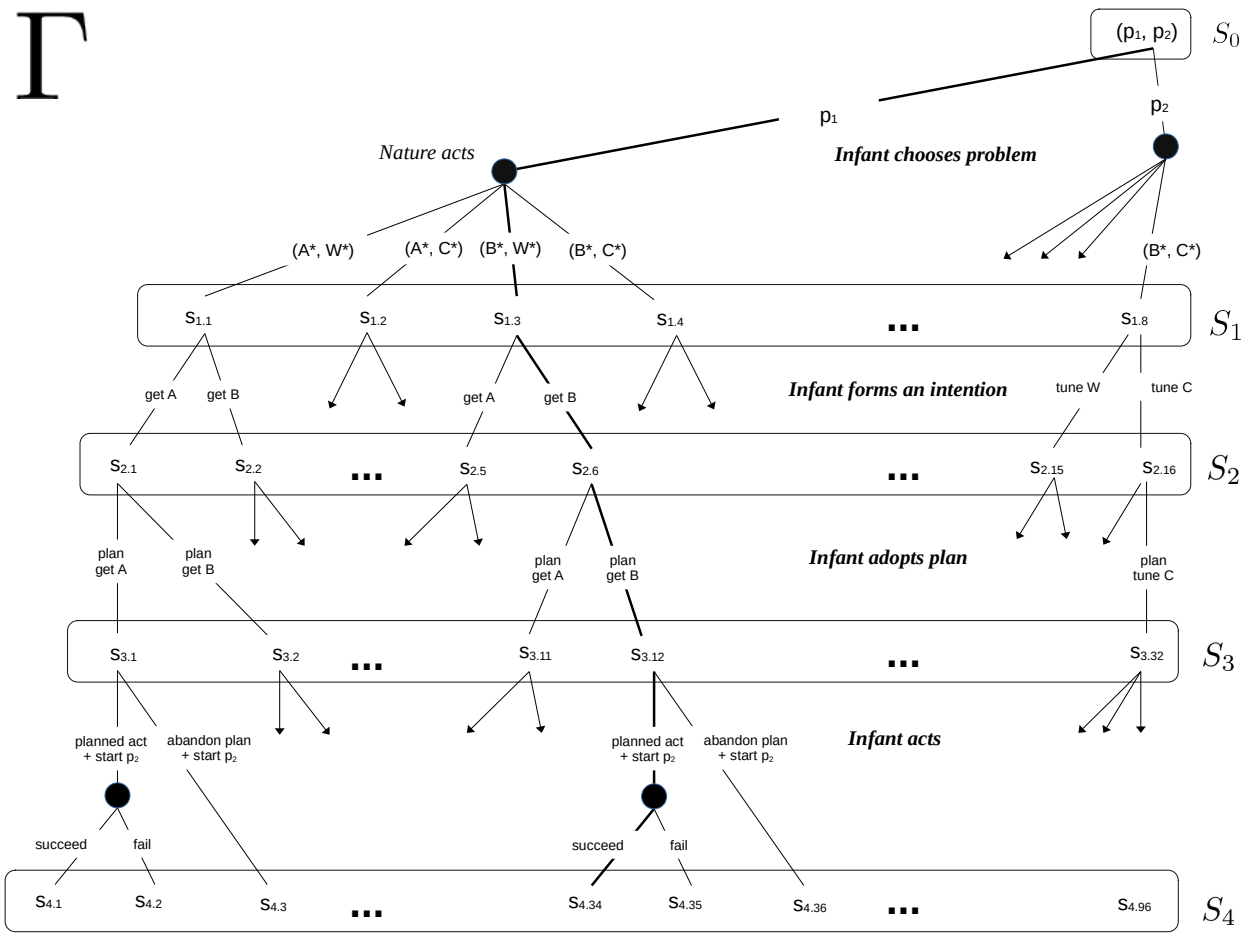

Figure 5: Objective tree spanning four periods; bolded path will be focal.

In $t = 1$, i deliberates and commits to an intention. The intention is a commitment to influence events such that a particular diachronic event is actualized. It is a goal. Here, ff i chooses p_1, then her intentions can be either to get A or to get B. Alternatively, if she chooses p_2, then her intentions can be to either to watch W or C.

With an intention in place at the start of $t = 2$, i then formulates a plan by which to accomplish the intention. Here, we assume the plan under p_1 is to approach and grab the intended toy. Under p_2, the plan is to pick up the TV remote and click to the intended show. Typically, state-contingent plans will be more elaborate, but in this simple example, the plan is a single-step.

With the plan in place at the start of $t = 3$, i acts according to the plan. At this point, we

assume Nature acts once again to determine the success of i's act with respect to the intention commitment at $t = 1$. For example, failure could involve i dropping the toy or Brian entering the room, sweeping up the toys, and placing them into a toy box out of i's reach.

To avoid clutter, Figure 5 does not illustrate every history in Γ. Still, there should be enough detail elaborated in the diagram to infer the entire structure of the objective tree. The number of states are expanding through time, from a single state at $t = 0$ to thirty-two at $t = 4$. Also omitted from each state's feasible actions for i is the option to halt solving a particular act-problem and begin another. As discussed, a set of available act-problems is present in every state. Here, i can always choose to drop what she is doing on one problem and start working on another. With this option included, the number of histories would multiply significantly, so we omit them—but readers should keep in mind this is always an available option.

Finally, let us quantify $d\oplus$ in this example as follows:

1. $d(h) = 100$ if h includes successfully obtaining the best toy;
2. $d(h) = 20$ if h includes successfully obtaining the second best toy;
3. $d(h) = 50$ if h includes successfully choosing the best TV show;
4. $d(h) = 10$ if h includes successfully choosing the best TV show;
5. $d(h) = -10$ if h includes failing to achieve any of the above.

Since the states in S_4 each correspond to a terminal history, we can write, e.g., $d(s_{4.1})$ without ambiguity. Thus, $d(s_{4.1}) = 100$, $d(s_{4.2}) = -10$, $d(s_{4.7}) = 25$, $d(s_{4.31}) = 50$, and so on.

7.1 Problem Selection

As discussed above, at the beginning of time, in objective state s_0, the individual may select from a set of mutually exclusive act-problems. The mutual exclusivity may be due to something inherent in the problems themselves (for example, Brian's infant may be faced with a decision about how to play indoors or how to play outdoors—these are inherently mutually exclusive), or they may be mutually exclusive due to the cognitive capacity constraints (for example, Brian's infant can play with a toy while watching TV, but only has the capacity to deliberate which toy is best or which show is best, but not both simultaneously).

For each objective state, s, the act-problems that are available for deliberation are given by P_s. At s_0, choosing an element in P_{s_0} is the *only* act available to i. Individual i assesses which problem to solve at s_0 based upon desires D_{s_0} and beliefs μ_{s_0}. In any future state, the individual may stop what he is doing and select a new act-problem to solve.

For now, we assume that i is aware of all the possible act-problems at s_0. However, we also assume that i has only a vague sense of the dynamics involved with each of these problems. Specifically, let Γ_{s_0} be a tree with $|P_{s_0}|$, T-length branches (one for each problem in P_{s_0}).[15] In this state, i consults his desires and chooses a problem that maximizes d_{s_0}. *This choice of an act-problem is the output of of the Problem Selection phase.*

Brian's Infant, Part V Figure 6 illustrates a partial illustration of the situation in s_0 (shown are periods $t = 0, 1$). The purpose of this diagram is to show how the state projections work. The objective tree Γ is shown on the top half of the diagram which accounts for all the feasible acts by Nature and i. The bottom half illustrates the corresponding parts of i's awareness tree Γ_{s_0} Here, i is aware of both act-problems that are available to her at the beginning of the decision process. However, at this early stage, she is only vaguely aware of what each of these problems entail. She has not analyzed either one in deep detail. Therefore, as she anticipates the unfolding of events, she is only aware of the broad histories consistent with "solving p_1" and "solving p_2."

The complete awareness tree is shown in Figure 7. Here, the awareness states are labeled with primes to distinguish them from their objective counterparts. What i envisions as she contemplates which problem to tackle is a sequence in which she chooses either to play with the best toy or to watch the best show. She knows the four phases required to get to an action. However, she is not thinking about the details of Nature's act at $t = 0$, nor is she aware of her feasible acts at each state. She does realize that, at the conclusion of her efforts, she may succeed or fail depending upon events outside of her control.

Notice that the four histories of which she is aware each correspond to eight distinct histories in reality ($4 \times 8 = 32$). For example, the terminal synchronic event associated with $s'_{4.1}$ is

$$r_{s_0}^{-1}(s'_{4.1}) = \{s_{4.1}, s_{4.3}, s_{4.5}, s_{4.7}, s_{4.9}, s_{4.11}, s_{4.13}, s_{4.15}\}.$$

These correspond to the diachronic event "the infant chose to solve p_1 and succeeds in her intended

[15] So, i is comparing the consequences of solving each problem over an equal time horizon.

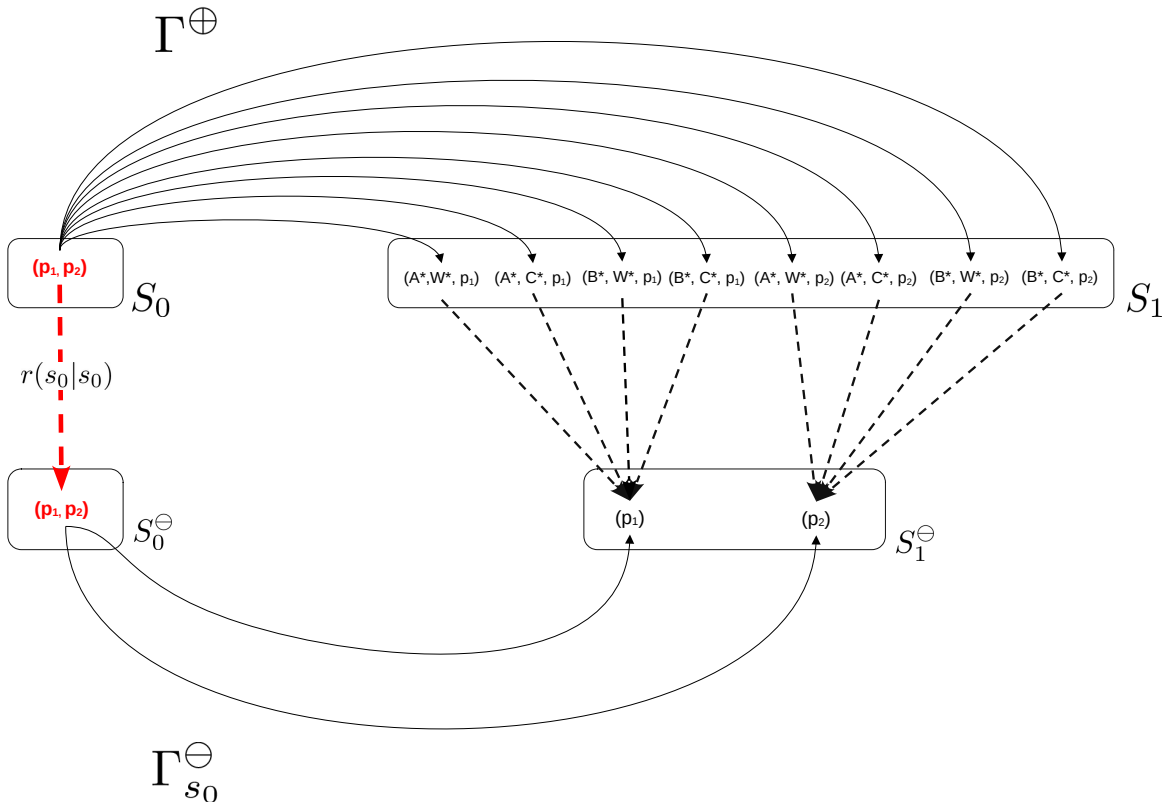

Figure 6: Partial awareness tree for Brian's Infant at the beginning of time s_0.

outcome." In terms of payoff values, this event includes two kinds of histories: i) those that include success in obtaining the best toy ($d(h) = 100$); and ii) those that include success in obtaining the second best toy ($d(h) = 20$). With respect to Item (ii), note that at $t = 1$ the infant is free to choose either toy as her intended plaything, even though she knows which toy is, in fact, best. Notice that intending second best *is* a feasible act at that point (even if her desires motivate her to intend otherwise).

In the p_1/success event, the min value is 20 and max value is 100, whereas in the corresponding p_2 event, the min value is 10 and the max value is 50. Suppose $\alpha = 0.5$. Then, the terminal history through $s'_{4.1}$ in Figure 7 has value $d^s_{s_0}(s'_{4.1}) = 60$ and the one through $s'_{4.3}$ has value $d^s_{s_0}(s'_{4.3}) = 30$. The fail histories, terminating in $s'_{4.2}$ and $s'_{4.4}$, both have value of -10.

To choose between p_1 and p_2, i's beliefs must come into play. Specifically, i must assess the probability of success or failure associated with each problem. Suppose that in state s_0, i believes the probability of success is 0.5 for both problems. Then, the expected value of choosing p_1 is

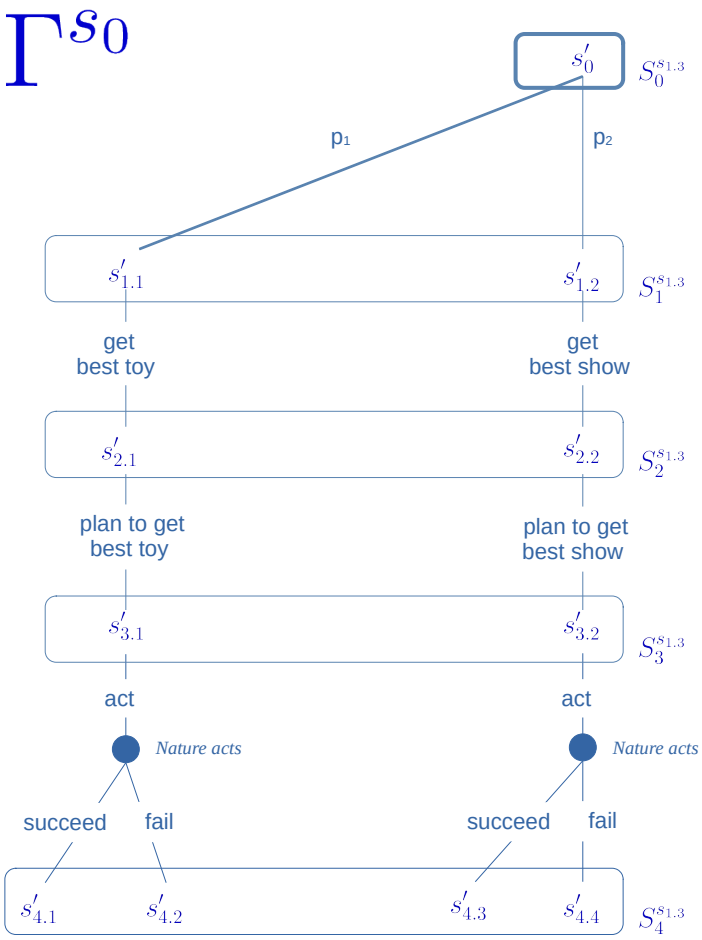

Figure 7: Awareness tree for i in objective state s_0.

$0.5 * 60 - 0.5 * 10 = 25$ and the expected value of choosing p_2 is 10. Therefore, i chooses p_1.

7.2 Deliberation

The Deliberation phase begins in $t = 1$ in objective state $s_1 = \omega(a_{s_0}) \in S_1$. During the time interval during which the state of the world evolves from s_0 to s_1, the individual shifts his awareness to support a deliberation on how to resolve his chosen act-problem. Typically, the awareness tree in $t = 1$ will refine certain aspects of the original awareness tree in $t = 0$ and coarsen others. In the new awareness tree, i is aware of the history of his own moves. He considers the terminal histories that are reachable from his present state and, based upon his preferences and beliefs, forms a

concrete intention.

Intentions take the form $\gamma_{s_1} \subset H^s_{s_1}$. *An intention is the output of the Deliberation phase.* Here, it is worth reminding ourselves that the awareness state in which i finds himself is given by $s'_1 = r(s_1|s_1)$. We assume the intention must be subjectively feasible. That is, the subjective histories included in γ_{s_1} must all pass through s'_1.[16]

Brian's Infant, Part VI Assume that the world evolves according to Figure 5; i.e., from objective states s_0 to $s_{1.3}$, where $s_{1.3}$ is actualized by the action profile $a_{s_0} = ((B^*, W^*), p_1)$ This shift in awareness causes i to have a more refined sense of the opportunities and issues associated with p_1. This is illustrated by the evolution from the awareness situation depicted in Figure 7 to the one depicted in Figure 8. In the latter, i is aware that B is the best toy, that A could have been the best toy, and that she could have embarked on a solution to p_2 instead of p_1. She is not thinking about anything having to do with which show to watch. Finally, although she is presently aware that Nature could have chosen A as the best toy (counterfactual state $S''_{1.1}$), she does not imagine that she will be thinking about that going forward.

There are a number of items of which to take notice at this point. First, referring to Figure 8, in $\Gamma^s_{s_{1.3}}$, i is aware of being in state $s''_{1.2}$. This state is the projection of the state $s_{1.3}$ depicted in Figure 5. As such, i is also aware of the subjective history that leads to $s''_{1.2}$. From this point, the feasible histories are the ones terminating in $s''_{4.1}$ through $s''_{4.4}$. The desire is to succeed at obtaining the best toy. Therefore, the intention going into period $t = 2$ is $\gamma_{s_{1.3}} = \{s''_{4.4}\}$.

[16] In most cases, the intention will be to effect a single terminal awareness history; i.e., $\gamma_{s_1} = \{h\}, h \in H^s_{s_1}$. However, this is not required. For example, the intention "Mike obtains a fresh cup of coffee," may be a diachronic event that includes many states (e.g., in which Mike's work being complete may or may not be true).

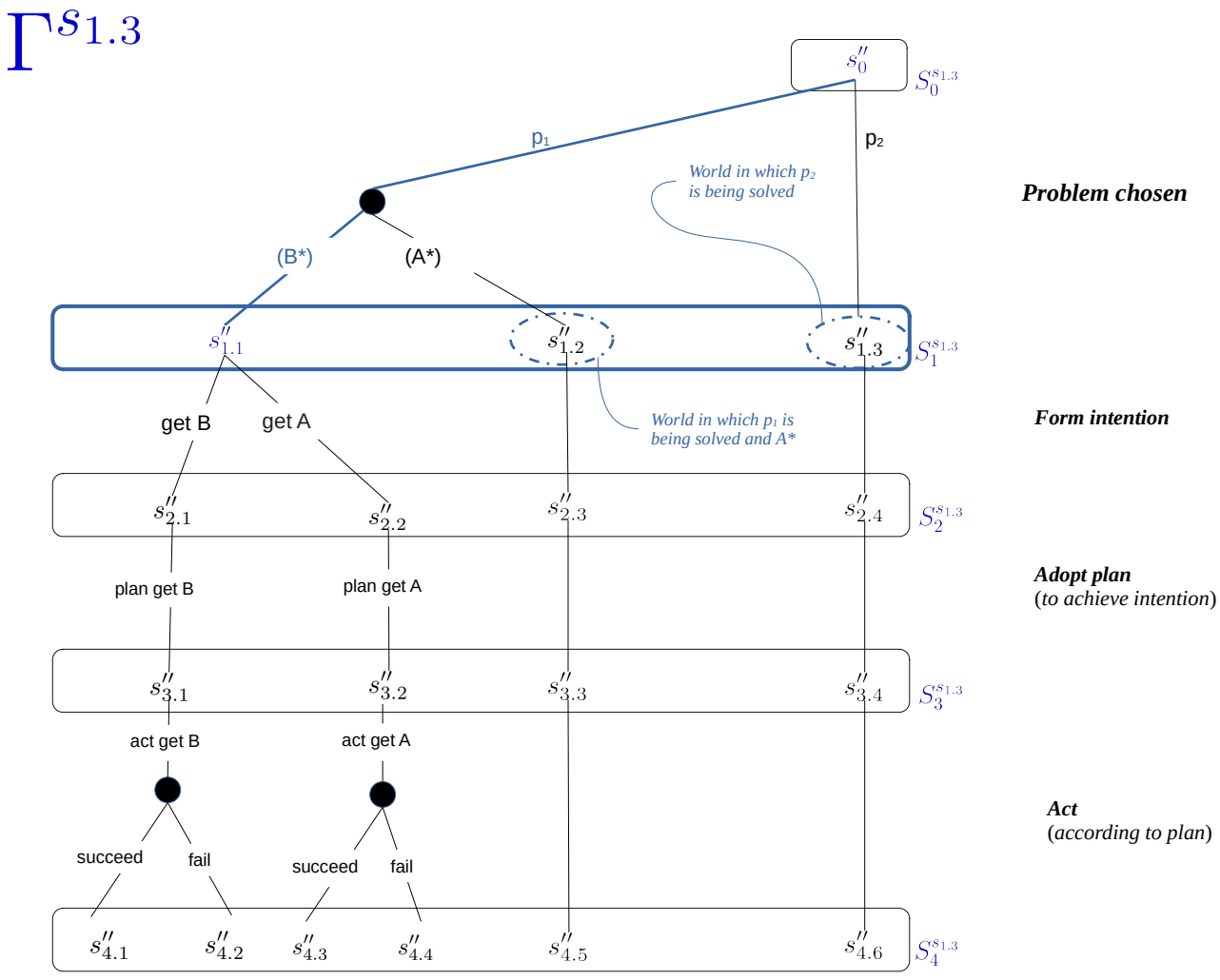

Figure 8: Awareness tree for i in objective state $s_{1.3}$.

7.3 Planing

Others have suggested that, in addition to helping us think through how to attain a desired end, plans serve the additional function of unencumbering the mind of some portion of its cognitive load. We agree and incorporate this aspect of planning explicitly into our analysis.

We write $\sigma_s^j(a_s^j)$ to indicate the probability that act a_s^j is selected at s. The use of probability distributions over feasible acts provides a nice level of generality. We note that, typically, individuals make action choices with certainty: e.g., $\sigma_s^i(a_s^i) = 1$ for some $a_s^i \in A_s^i$. However, this setup also allows for situations in, e.g., i comes to a decision point and decides to "flip a coin" to determine which act to choose.

The plan is intentionally developed by i to achieve an intention consistent with his desires and beliefs. Notice that, in particular, $\sigma_{s',s}^i$ specifies an act for i at $I_{s',s}$ where $s' = r(s|s)$ is the subjective state into which s projects when i is in s. Therefore, we write σ_s^i to indicate the act i will take at the objective state s as implied by his subjective plan. Finally, note the difference between an agent randomizing over acts versus being uncertain about where in the tree he is. Even when i is certain about his state (i.e., is at a singleton information set), he can still choose to randomize over his feasible acts.

Although we do not assume Nature is similarly intentional, we can nevertheless describe her behavior in the same way. That is, Nature's "plan" is also a list

$$\sigma^n \in \Sigma^n \equiv \prod_{s \in S} \Delta(A_s^n).$$

Whereas, we imagine individuals typically choose their acts with certainty, Nature's choices almost surely involve some measure of randomness. At the same time, Nature is never "uncertain" about where she is in Γ—all her information sets are singletons. Therefore, Nature's plan is a description of her behavior at every objective state. At state s, $\sigma_s^n \in (A_s^n)$ is the random act Nature takes at s.

Brian's Infant, Part VII

7.4 Action

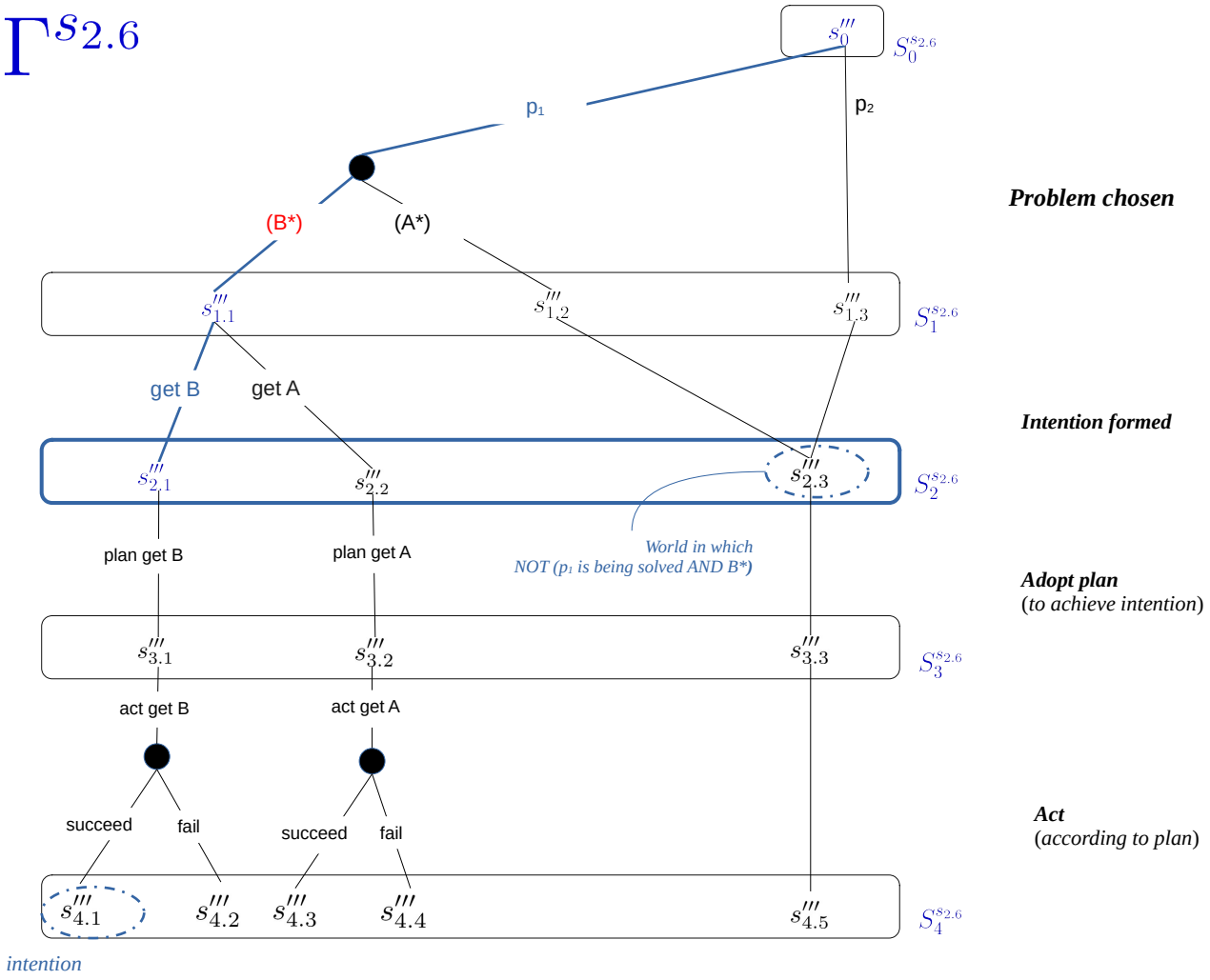

Figure 9: Awareness tree for i in objective state $s_{2.6}$.

References

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics 4*(6), 1236–1239.

Bratman, M. (2014). *Shared agency: A planning theory of acting together.*

Bryan, K., M. D. Ryall, and B. C. Schipper (2021). Value-capture in the face of known and unknown unknowns. *Strategy Science (forthcoming).*

Dekel, E., B. L. Lipman, and A. Rustichini (1998). Standard state-space models preclude unaware-

ness. *Econometrica 66*(1), 159–173.

Geanakoplos, J. (1989). Game theory without partitions, and applications to speculation and consensus. Technical report.

Harsanyi, J. C. (1967). Games with incomplete information played by "bayesian" players, i–iii: Part i. the basic model. *Management Science 14*(3), 159–182.

Heifetz, A., M. Meier, and B. C. Schipper (2006, sep). Interactive unawareness. *Journal of Economic Theory 130*(1), 78–94.

Heifetz, A., M. Meier, and B. C. Schipper (2008, jan). A canonical model for interactive unawareness. *Games and Economic Behavior 62*(1), 304–324.

Heifetz, A., M. Meier, and B. C. Schipper (2013, jan). Unawareness, beliefs, and speculative trade. *Games and Economic Behavior 77*(1), 100–121.

Mertens, J. F. and S. Zamir (1985). Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory 14*(1), 1–29.

Samet, D. (1990). Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory 52*(1), 190–207.

Schipper, B. C. (2015). *Awareness, in Handbook of Epistemic Logic*, Chapter 3. College Publications.

Schipper, B. C. (2016). Network formation in a society with fragmented knowledge and awareness. Technical report.