Intentional Awareness[1]

Brian Epstein
Tufts University, Medford

Michael D. Ryall
University of Toronto

December 1, 2021

[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

2.1 General

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 $X \setminus Y$ 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 model.

2.2 Notation Reference

Table **??** elaborates all the mathematical objects used in the paper.

3 Objective Reality ($\oplus$)

In this section, we describe the status and dynamics of "brute" reality; that is, 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 *agents*, 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 evolution of the world through time. We focus on the action over a fixed period of time, from $t = 0$ to $t = T$. Time is indicated with subscripts.

When you encounter an "$\oplus$" superscript attached to an object, think of it as indicating its objective reality, whereas an "$\ominus$" superscript indicates the object as subjectively perceived by the individual. In general, $\oplus$-objects are richer and more refined than $\ominus$-objects. Alternatively, n and i superscripts are used to indicate that an object belongs to or is chosen by Nature or the individual, respectively.

3.1 States of the world

As outlined in the Introduction, at time t, individual i finds himself in a particular situation. Specifically, at that time he exists in an *objective state of the world*, denoted $s_t^{\oplus}$. Think of $s_t^{\oplus}$ as a snapshot describing the world in all its detail. This state corresponds to objective reality, including the status *of all features of the world* in that moment. Importantly, these include the mind-independent features of the world as well as the *mental attitudes* of the individuals acting in that world.

In addition to elaborating how the world actually exists at t, we want to equip our individual with the ability to consider counterfactuals and future possibilities—that is, to be able to think about the way the world is, the way it might have been and the way it could be. With this in mind, let $S_t^{\oplus}$ denote the collection of states that constitute the *objective state space at time t*. If t is the present or some historic time, then one state in $S_t^{\oplus}$ is factual and the others are counterfactual. If t is some future time, then $S_t^{\oplus}$ elaborates all the possibilities that could occur at t. Let $S^{\oplus} \equiv \cup_{t=0}^{T} S_t^{\oplus}$

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

be the set of all objective states.

In terms of interpretation, it does no harm to imagine that a state, $s_t^{\oplus}$, elaborates an uncountably infinite number of features of the world. However, we assume the the number of states in $S_t^{\oplus}$ 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 Act-Problems

Here and in the sections that follow, we adopt the convention of using functions and correspondences to "pull" some information of interest out of a state of the world. Presently, we are interested in the set of mutually exclusive act-problems available to i for selection, deliberation, planning, and action. Therefore, let $P(s_t^{\oplus}) = \{p_{t.1}^{\oplus}, \ldots, p_{t.k}^{\oplus}\}$ denote the indexed set of act-problems objectively available to i in state $s_t^{\oplus}$ (where we adopt the notational convention of appending the time subscript with an object's index number, i.e., "$t.\#$"). In particular, $P(s_0^{\oplus})$ is the initial set of problems objectively available to i at the beginning of time. At any future state, $s_t^{\oplus}$, i may stop whatever he is doing and select some new problem to solve from $P(s_t^{\oplus})$.

3.3 Acts

In our approach the acts of Nature and the decision maker jointly cause the system to evolve from a state $s_t^{\oplus} \in S_t^{\oplus}$ to a new state $s_{t+1}^{\oplus} \in S_{t+1}^{\oplus}$. 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 $s_t^{\oplus} \in S^{\oplus}$, let $A^j(s_t^{\oplus})$ indicate the set of *feasible acts available to actor j in state $s_t^{\oplus}$* with arbitrary element $a_t^j \in A^j(s_t^{\oplus})$, where $j \in \{i, n\}$.[3] We adopt the convention that $A^j(s_t^{\oplus}) = \emptyset$

[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.

indicates that actor j has no feasible acts in state $s_t^\oplus$. An *objective act profile* is a pair of acts, $a_t^\oplus \equiv (a_t^n, a_t^i)$, one by Nature and one by i. The set of *all objective act profiles at state* $s_t^\oplus$ is $A(s_t^\oplus) \equiv A^n(s_t^\oplus) \times A^i(s_t^\oplus)$. Note the implication that the acts of i and n at $s_t^\oplus$ 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. Also, keep in mind that the act profiles in $A(s_t^\oplus)$ are unique. The set of *all possible objective act profiles at time* t is $A_t^\oplus \equiv \cup_{s_t^\oplus \in S_t^\oplus} A(s_t^\oplus)$; and the set of *all possible objective act profiles* is $A^\oplus \equiv \cup_{t=0}^T A_t^\oplus$.

Individual i's set of feasible acts at the beginning of time corresponds to choosing one of the available act-problems to solve. That is, $A^i(s_0^\oplus) \equiv P(s_0^\oplus)$ and $a_0^i = p_{0.j}^\oplus$ means that i chooses the j^{th} period 0 problem to solve.

3.4 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, assume that Nature kicks things off at the beginning of time. Specifically, $S_0^\oplus = \{s_0^\oplus\}$. That is, the world begins in state $s_0^\oplus$, in which i is presented with a number of mutually exclusive act-problems from which to choose. From this state, i chooses which problem to solve (the Problem Selection phase, which will be followed by the deliberation, planning and then acting phases).

Let $\Gamma^\oplus \equiv \langle S^\oplus, E^\oplus \rangle$ be a graph in which the nodes are the elements of $S^\oplus$ and $E^\oplus \subset S^\oplus \times S^\oplus$ is a set of edges. Assume: (i) $\Gamma^\oplus$ is a tree with root node $s_0^\oplus$; and (ii) if $(s_t^\oplus, s_k^\oplus) \in E^\oplus$, then $k = t+1$. Then, we wish to associate each edge $(s_t^\oplus, s_{t+1}^\oplus)$ a single action profile $a_t^\oplus$ that identifies the causes that move the world from state $s_t^\oplus$ to $s_{t+1}^\oplus$. To this end, let $\omega : E^\oplus \to A^\oplus$ be the *state-contingent actualization function*, where $\omega(s_t^\oplus, s_{t+1}^\oplus) = a_t^\oplus = (a_t^n, a_t^i)$ indicates that the the act profile that causes the world to evolve from $s_t^\oplus$ to $s_{t+1}^\oplus$ is $a_t^\oplus$.

Let $C(s_t^\oplus) \subset E$ be the set of edges of the form $(s_t^\oplus, \cdot)$. We assume that: (i) ω is surjective (no superfluous act profiles—every act profile in $A^\oplus$ is associated with at least one edge in E); and (ii) for all $s_t^\oplus$, ω is bijective from $C(s_t^\oplus)$ to $A(s_t^\oplus)$ (each feasible action at $s_t^\oplus$ is associated with a unique edge emanating from $s_t^\oplus$ and visa versa). Therefore, since the act profiles in $A(s_t^\oplus)$ are unique, there is no situation in which a specific act profile can lead from one state to more than one future state.

Now, it may well be that two or more states have the same the set of feasible act profiles. That is, for $s_t^{\oplus} \neq \hat{s}_t^{\oplus}$ it may be that $A(s_t^{\oplus}) = A(\hat{s}_t^{\oplus})$. For example, at 11am I may be finished with my work and get a cup of coffee or, alternatively, I may not be finished with my 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 my getting a cup of coffee in the finished-work state leads to two or more distinct states in the future.

We define the *objective history at state* $s_t^{\oplus}$ as the path in Γ^n that starts at $s_0^{\oplus}$ and ends at $s_t^{\oplus}$, denoted $h(s_t^{\oplus}) = (s_0^{\oplus} \ldots, s_t^{\oplus})$. Notice that a history $h(s_t^{\oplus})$ is actualized by the sequence of act profiles $\omega(s_0^{\oplus}, s_1^{\oplus}), \ldots, \omega(s_{t-1}^{\oplus}, s_t^{\oplus})$. The set of all *objective histories at time t* is $H_t^{\oplus}$, with an arbitrary element denoted $h_t^{\oplus}$ The world begins with the *null history* $h_0^{\oplus} = (s_0^{\oplus})$ at the beginning of time: so, $H_0^{\oplus} = \{h_0^{\oplus}\}$ and $S_0^{\oplus} = \{s_0^{\oplus}\}$. The set of all objective complete histories is $H_T^{\oplus}$.

Because $\Gamma^{\oplus}$ is a tree, each state is associated with a unique history leading to it. Therefore, we can think of a state of the world $s_t^{\oplus}$ as also carrying the information $h(s_t^{\oplus})$ as well as the associated actions $\omega(s_0^{\oplus}, s_1^{\oplus}), \ldots, \omega(s_{t-1}^{\oplus}, s_t^{\oplus})$. Suppose, for example, that two distinct sequences of acts could lead to a physically identical footprint in the snow. In our setup, there will be two states in which that identical footprint exists, each associated with its own distinct history.

3.5 Example: Brian's Infant, Part 1

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 exclusive act-problems: to play indoors or, alternatively, to go outside. If i decides to play indoors, then she must pick between one of two new toys with which to play. Alternatively, i can choose to play outdoors, in which case she must decide whether to put on flip-flops or boots. 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 play outdoors. If so, the weather outside can be warm, W or cold, C.

At the start of time, the world presents the infant with a list of mutually exclusive act-problems from which to choose. Here, $P(s_0^{\oplus}) = \{p_1, p_2\}$ where p_1 is how to play indoors and p_2 is how to play outside. Therefore, $A^i(s_0^{\oplus}) = \{p_1, p_2\}$. Simultaneously, Nature will determine which toy is best and

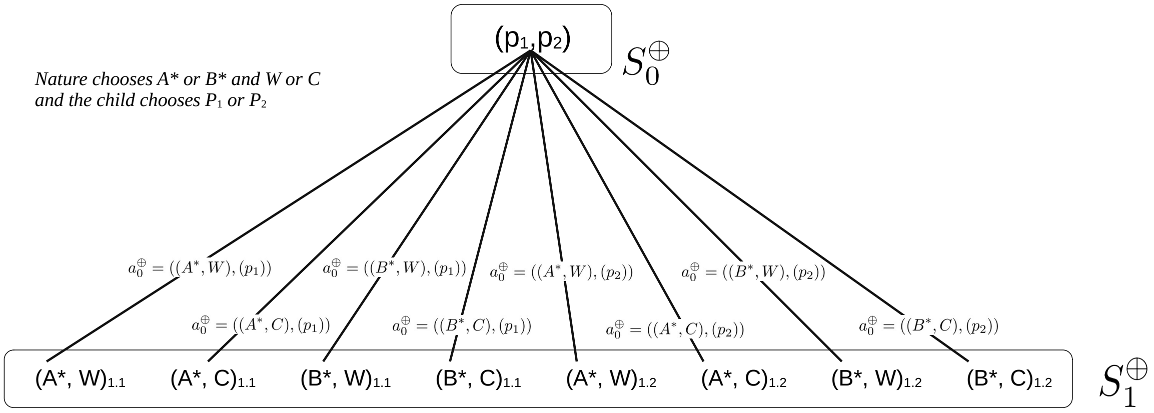

(p1,p2)
$S_0^{\oplus}$
Nature chooses A* or B* and W or C
and the child chooses P1 or P2
$a_0^{\oplus} = ((A^*, W), (p_1))$
$a_0^{\oplus} = ((B^*, W), (p_1))$
$a_0^{\oplus} = ((A^*, W), (p_2))$
$a_0^{\oplus} = ((B^*, W), (p_2))$
$a_0^{\oplus} = ((A^*, C), (p_1))$
$a_0^{\oplus} = ((B^*, C), (p_1))$
$a_0^{\oplus} = ((A^*, C), (p_2))$
$a_0^{\oplus} = ((B^*, C), (p_2))$
(A*, W)1.1
(A*, C)1.1
(B*, W)1.1
(B*, C)1.1
(A*, W)1.2
(A*, C)1.2
(B*, W)1.2
(B*, C)1.2
$S_1^{\oplus}$

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

what the weather will be at time $t = 1$. Therefore, Nature selects one of among four possible acts: $A^n(s_0^{\oplus}) = \{(A^*, W), (A^*, C), (B^*, W), (B^*, C)\}$.

Figure 1 illustrates the evolution of the system from $t = 0$ to $t = 1$. From $s_0^{\oplus}$, the objective action profiles are of the form $a_0^{\oplus} = ((Y, X), (Z))$ where $a_0^n = (Y, X)$ is Nature's choice of Y, the best toy, and X, the weather, and Z is i's choice of which act-problem to solve. Then, $S_1^{\oplus}$ contains eight possible states (one for each combination of these variables). Note that some states describe the same features of the world; e.g., $(A^*, W)_{1.1}$ and $(A^*, W)_{1.2}$. The differences between these is the history by which they come to be actualized.[4] Thus, if $(A^*, W)_{1.1}$ where actualized, i might be aware of A^*, W, and/or the fact that she had reached that state by deciding to solve p_1.

4 Subjective Reality ($\ominus$)

The individual's subjective world includes awareness of certain features of the objective world, desires, and beliefs. Each of these elements are elaborated below. Certain elements, such as beliefs, which belong uniquely to the individual are indicated with i superscripts. Others, such as states of awareness, are impoverished versions of some element of objective reality (an $\oplus$-indicated object). These are indicated with $\ominus$ superscripts.

[4] Keep in mind that the history by which each state comes to be is embedded in the state. Thus, for example, two states of the world could have identical footprints in the mud but be distinguished from each other by being made by different people.

4.1 Awareness

Our basic approach is to assume that, in each objective state of the world, the individual has in mind a subjective version of $\Gamma^{\oplus}$. Specifically, for all $s_t^{\oplus} \in S^{\oplus}$, let $\Gamma(s_t^{\oplus}) \equiv \langle S(s_t^{\oplus}), E(s_t^{\oplus})\rangle$ represent the individual's subjective awareness of $\Gamma^{\oplus}$ when the individual is in state $s_t^{\oplus}$. When $s_t^{\oplus}$ is arbitrary or understood from the context, we write $\Gamma^{\ominus} \equiv \langle S^{\ominus}, E^{\ominus}\rangle$ and denote individual states of awareness, $s_t^{\ominus}$, the awareness state space at time t, $S_t^{\ominus}$, and the set of all awareness states, $S^{\ominus}$. The idea is to map the objective, full-featured description of how the world may evolve (according to $\Gamma^{\oplus}$) to impoverished versions that include only the features of the world of which individual i is aware in any given objective state. We extend the state-contingent actualization function so that $\omega(s_k^{\ominus}, s_{k+1}^{\ominus}|s_t^{\oplus}) = a_k^{\ominus}$ indicates that, according to i's awareness in $s_t^{\oplus}$, the act profile that causes the world to evolve from $s_k^{\ominus}$ to $s_{k+1}^{\ominus}$ is $a_k^{\ominus}$. Assume that ω meets the same surjectivity and bijectivity conditions for each $\Gamma(s_t^{\oplus})$ as for $\Gamma^{\oplus}$.

Thus, if i is in the objective state $s_t^{\oplus}$, then $s_k^{\ominus} \in S(s_t^{\oplus})$ represents i's subjective awareness of a state in period k. If $k < t$, then $s_k^{\ominus}$ includes the features of which i is aware of a past state of the world that did occur or that could have occurred; if $k = t$, then it represents his awareness of the features either of the actual objective state in which he finds himself or of a counterfactual; and, if $k > t$, then it represents his awareness of the features a state in the future.

Then, for the period-k awareness state $s_k^{\ominus} \in S(s_t^{\oplus})$, $P(s_k^{\ominus}) = \{p_{k.1}^{\ominus}, \ldots, p_{k.k}^{\ominus}\}$ denotes the set of available act-problems that i perceives are available in $s_k^{\ominus}$. Similarly, $A^j(s_k^{\ominus})$ is the perceived set of acts available to agent j in $s_k^{\ominus}$ according to i's awareness in $s_t^{\oplus}$. The set of all subjective act profiles associated with $s_k^{\ominus}$ is $A(s_k^{\ominus}) \equiv A^n(s_k^{\ominus}) \times A^i(s_k^{\ominus})$. When $s_t^{\oplus}$ and $s_k^{\ominus}$ are arbitrary or understood from the context, the set of all period-k act profiles (again, as i is aware of them) is $A_k^{\ominus}$ and the set of all possible act profiles of which i is aware is $A^{\ominus} \equiv \cup_{t=0}^{T} A_t^{\ominus}$.[5]

Finally, given $s_k^{\ominus} \in S(s_t^{\oplus})$ the *subjective history at state* $s_k^{\ominus}$ is the path in $\Gamma(s_t^{\oplus})$ that starts at $s_0^{\ominus}$ and ends at $s_k^{\ominus}$ and is given by $h(s_k^{\ominus})$. With this context understood, the set of all *subjective histories at time t* is $H_t^{\ominus}$, with an arbitrary element denoted $h_t^{\ominus}$ and the set of all subjective complete histories denoted $H_T^{\ominus}$.

This brings us to the issue of how objective states and actions correspond to i's awareness of them. To that end, we assume that i's awareness is a limited but accurate version of reality. In

[5] Again, the pattern is that "$\oplus$" objects—like $A^{\oplus}$—all have "$\ominus$" counterparts—like $A^{\ominus}$. Functions and correspondences—like $A^j(\cdot)$—return objective or subjective objects depending upon whether the argument is objective or subjective, respectively.

particular, we operate from the assumption that, although i is not aware of all the features of reality, those of which he *is* aware are correct. To this end, for each $s_t^{\oplus} \in S^{\oplus}$, define the surjective *projection* $r(\cdot|s_t^{\oplus}) : S^{\oplus} \to S(s_t^{\oplus})$ and assume that if $s_j^{\ominus} = r(s_k^{\oplus}|s_t^{\oplus})$ then $j = k$. Thus, for example, $s_{t-3}^{\ominus} = r(s_{t-3}^{\oplus}|s_t^{\oplus})$ is the impoverished version of the historical $s_{t-3}^{\oplus}$ about which i is aware upon finding himself state $s_t^{\oplus}$. Making $r(\cdot|s_t^{\oplus})$ surjective means that every objective state always maps to some state in i's awareness space and, possibly, many to one (reality is typically more refined than i's awareness of reality). The $j = k$ requirement ensures that i's awareness is not time-inconsistent; e.g., i does not mistake something that happened yesterday with something happening today.

We also impose certain consistency conditions on i's awareness over time. If $h_t^{\oplus} = (s_0^{\oplus}, \ldots, s_t^{\oplus})$ is the objective history at t, then i is aware of the corresponding history as he experienced it, namely, $(s_0^{\ominus}, \ldots, s_t^{\ominus})$ where $s_0^{\ominus} = r(s_0^{\oplus}|s_0^{\oplus}), \ldots, s_t^{\ominus} = r(s_t^{\oplus}|s_t^{\oplus})$. We assume that, at $s_t^{\oplus}$, i has perfect recall of his own actions corresponding to $h_t^{\oplus}$. We do not require i to be aware of Nature's acts (though, he may well be so). Instead, i may speculate or hypothesize about what Nature's acts are that, in combination with his own, get from $s_0^{\ominus}$ to $s_t^{\ominus}$.

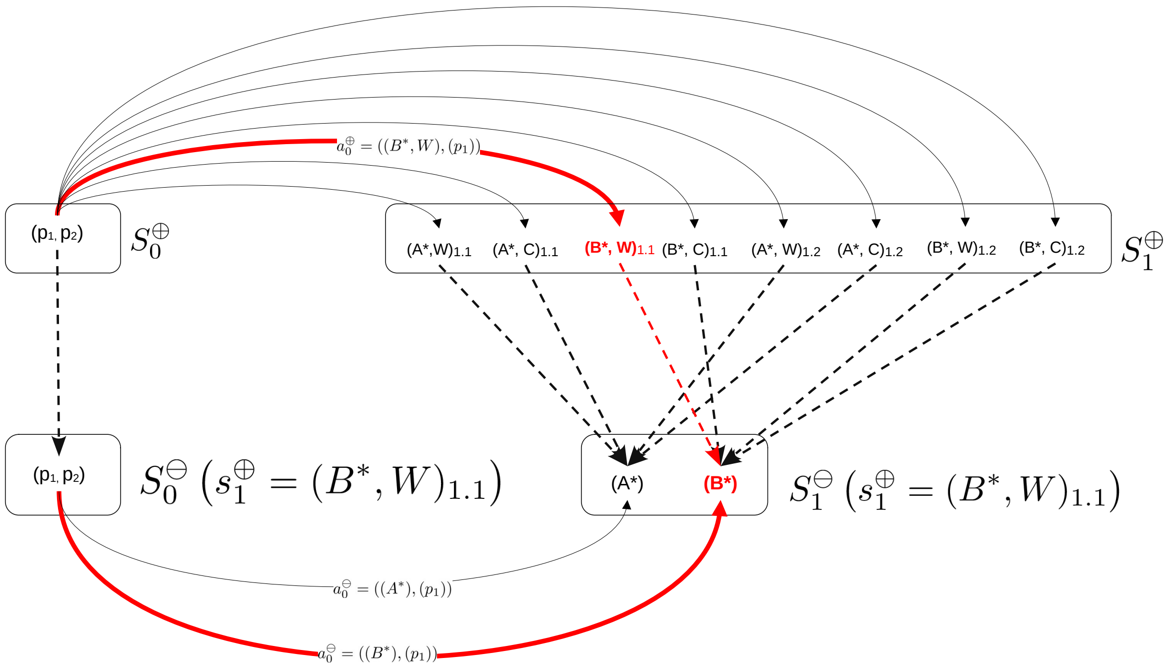
4.2 Example: Brian's Infant, Part 2

Picking up the example of Brian's Infant where we left off, let us map the preceding formalism to the infant's perceptions of reality. At the beginning of time $s_0^{\oplus} = (p_1, p_2)$. Suppose i is fully aware of this. Then, $S(s_0^{\oplus}) = \{s_0^{\ominus}\}$ where $s_0^{\ominus} = (p_1, p_2)$. Moreover, since i is aware of the available act-problems, she correctly perceives $A(s_0^{\ominus}) = \{p_1, p_2\}$. Now, suppose the act profile at $t = 0$ is $a_0^{\oplus} = (B^*, W, p_1)$: Nature's act is $a_0^n = (B^*, W)$ and i chooses to work on p_1 (how to play outside).

According to Figure 1, this brings the world to objective state $s_1^{\oplus} = (B^*, W)_{1.1}$ in period $t = 1$. Suppose that, having decided to solve the play-indoors problem, i becomes aware of which toy is best—to the exclusion of other features of the world. Then, i finds herself in awareness state $s_1^{\ominus} = (B^*)$. Being aware that B is the best toy, presumably she can reason about the counterfactual that A could have been best instead. Then, $S_1^{\ominus} = \{(A^*), (B^*)\}$. Furthermore, i is able to recall her awareness from $t = 0$, which makes her aware of her experienced history $h_1^{\ominus} = ((p_1, p_2), (B^*, W)_{1.1})$.

Notice that we can keep track of how each objective state maps to i's states of awareness when i is in $(B^*, W)_{1.1}$. According to the preceding discussion, $(B^*) = r((B^*, W)_{1.1}|(B^*, W)_{1.1})$. However, having chosen p_1, i is only aware of the quality of the toys. This can be represented by setting $(B^*) = r((B^*, \cdot)_{1.\#}|(B^*, W)_{1.1})$ and $(A^*) = r((A^*, \cdot)_{1.\#}|(B^*, W)_{1.1})$: that is, all objective states in which B is best map to awareness state $s_1^{\ominus} = (B^*)$ and those in which A is best map to $s_1^{\ominus} = (A^*)$.

Figure 2 illustrates these ideas. In this diagram, the top half is a reproduction of Figure 1 tipped on its side. The bottom half presents the corresponding awareness tree $\Gamma^{\oplus}$ at $s_1^{\oplus} = (B^*, W)_{1.1}$. Notice that, in this tree, i recalls her experience of $s_0^{\ominus}$ from the perspective of $s_1^{\ominus} = (B^*)$. From this point, i must deliberate about what goal to seek, set up an action plan to obtain it, and then execute the plan.



(p1, p2)
$S_0^{\oplus}$
$a_0^{\oplus} = ((B^*, W), (p_1))$
(A*,W)1.1
(A*, C)1.1
(B*, W)1.1
(B*, C)1.1
(A*, W)1.2
(A*, C)1.2
(B*, W)1.2
(B*, C)1.2
$S_1^{\oplus}$
(p1, p2)
$S_0^{\ominus}\ (s_1^{\oplus} = (B^*, W)_{1.1})$
(A*)
(B*)
$S_1^{\ominus}\ (s_1^{\oplus} = (B^*, W)_{1.1})$
$a_0^{\ominus} = ((A^*), (p_1))$
$a_0^{\ominus} = ((B^*), (p_1))$

Figure 2: Detailed awareness space for Brian's Infant when $s_1^{\oplus} = (B^*, W)_{1.1}$ with state projections

*******STOP HERE*******

4.3 Synchronic 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 game theory usage) as *synchronic events*, and those associated with properties unfolding through time (the philosophy usage) as *diachronic events.* We begin with synchronic events.

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. Our usage will be consistent with the standard probability theoretic one.

In the Brian's child example in Figure 2, (A^*) is a state in i's awareness when i is in Nature's state (C, A^*). This state corresponds to the event in Nature's state space $\{(R, A^*), (C, A^*)\}$. To be precise, it is the event "A is the best toy." If someone of impeccable integrity tells a fully aware individual (one whose awareness space is equivalent to Nature's state space), "A is the best toy," then that individual would know that (R, A^*) or (C, A^*) is true. However, when i in state (C, A^*) is told, "A is the best toy," she knows only that (A^*) is true because she is unaware of the weather—she is not thinking about it.

Notice that, because r is surjective, i's awareness *states* in S^i always correspond to *events* in $S^\oplus$. Specifically, for all $s_t^\oplus \in S^\oplus$, $r^{-1}(s_k^i|s_t^\oplus) \subseteq S_k^\oplus$ is the event in Nature's state space associated with s_k^i. Nature's states elaborate all the features of the world. Therefore, Nature's state space contains all the states required to account for all the potentially realizable instantiations of features. In any given moment, individuals are typically aware neither of all the possible features nor of all the realizable instantiations of those features of which they are aware.

Because individual i's state spaces are related to fully elaborated reality, it will be helpful to define a kind of "event" that accounts for states of awareness *as well as* their associated states of Nature. For an event of which i is aware in state $s_t^\oplus$, say $B \subseteq S^i(s_t^\oplus)$, let $B^\uparrow = B \bigcup r^{-1}(B|s_t^\oplus)$ be the extension of B to include all states in Nature's state space that project into B. Then, a *synchronic event in state* $s_t^\oplus$ is a set of the form $B^\uparrow$ for some $B \subseteq S^i(s_t^\oplus)$. We refer to B as the *basis* of the sychronic event $E = B^\uparrow$. Let $\Sigma(s_t^\oplus)$ be the collection of all synchronic events in state $s_t^\oplus$. If $E \in \Sigma(s_t^\oplus)$ with basis B, then $\neg E \equiv (S^i(s_t^\oplus) \setminus B) \uparrow$. By this definition, not every subset of $S^\oplus$ is the extension of some basis in $S^i(s_t^\oplus)$. Nevertheless, by these definitions, $\neg\neg B^\downarrow = B^\downarrow$.

To see how this definition works, consider the illustration in Figure 3. This shows Nature's

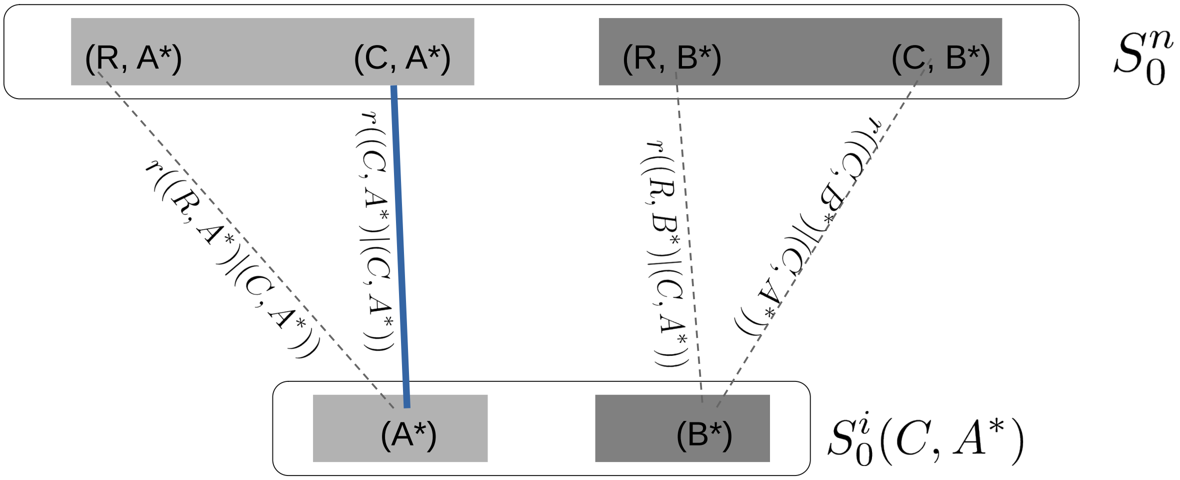


Figure 3: Synchronic events in the Brian's Infant example

state space and the child's awareness space when the actualized state is (C, A^*). The potential basis events in $S_0^i(C, A^*)$ are $\{(A^*)\}, \{(B^*)\}, \{(A^*), (B^*)\}$, and $\emptyset^i$ (where we use the "i" superscript to indicate the reference to S_0^i, i.e., as opposed to $S_0^{\oplus}$). Then, $\{(A^*)\} \uparrow = \{(A^*), (R, A^*), (C, A^*)\}$. Working through all four possibilities, we see that $\Sigma(C, A^*)$ contains the following synchronic events:

$$\{(A^*), (R, A^*), (C, A^*)\},$$

$$\{(B^*), (R, B^*), (C, B^*)\},$$

$$\{(A^*), (B^*), (R, A^*), (C, A^*), (R, B^*), (C, B^*)\},$$

$$\{\emptyset^i, \emptyset^n\}.$$

It is worth noting that there are several subsets in $S_0^{\oplus}$ that are not implied by any basis and, therefore, are not included as part of a synchronic event in $\Sigma(C, A^*)$. Examples include $\{(R, A^*), (R, B^*)\}$, $\{(C, A^*), (R, B^*)\}$, $\{(R, A^*), (C, A^*), (R, B^*)\}$ and so on.

4.4 Beliefs and uncertainty

In addition to the states about which i can think about, consistent with S^i, we wish to account for uncertainty on the part of the individual decision maker. To begin, for each state of Nature $s_t^\oplus \in S^\oplus$, define the *information partition at* $s_t^\oplus$, denoted $\mathcal{I}^i(s_t^\oplus)$, to be a partition of $S^i(s_t^\oplus)$. We refer to an event in $I^i \in \mathcal{I}^i(s_t^\oplus)$ as an *information set.* Information sets serve to distinguish states about which i is aware but uncertain. In keeping with time consistency, we require that for each $I^i \in \mathcal{I}^i(s_t^\oplus)$, there exists a $S_t^i(s_t^i)$ that contains I^i; i.e., information sets do not span time periods. The information set I^i corresponds to the synchronous event $I^i \uparrow$.

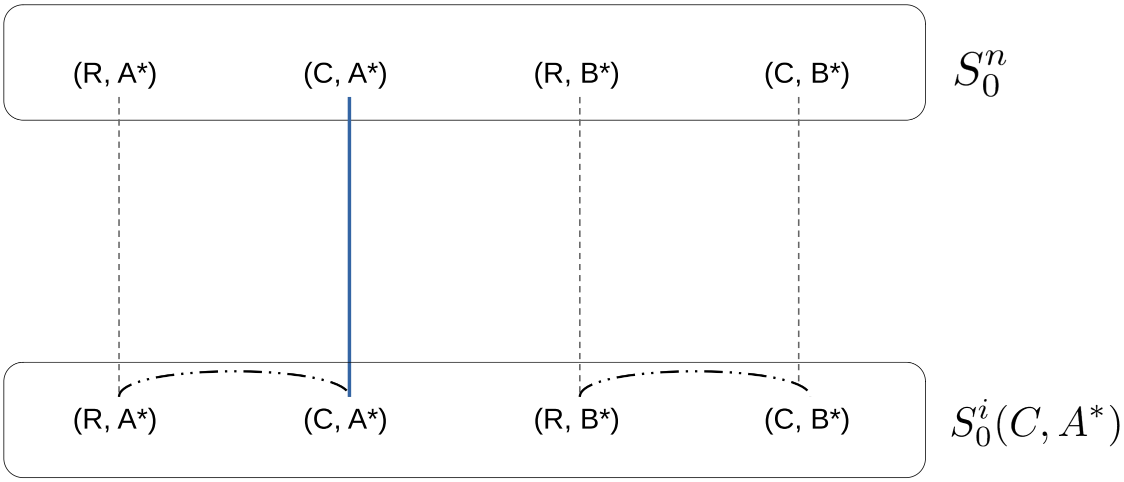


Figure 4: Fully aware child is uncertain about the weather

Information sets are illustrated for the Brian's Infant example in Figure 4. In this scenario, i is fully aware of all the states of Nature (her awareness space is as refined as Nature's). The information sets are shown by the double-dotted dashed lines. The interpretation is as follows. The true state of the world is (C, A^*) and knows that A is the best toy. However, although i is aware that the weather could be cloudy or raining, she is unable to discern which is true. She knows the best toy but is uncertain about the state of the weather. At the same time, she can reason about counterfactuals: B could have been the best toy and, had it been, she still would have been uncertain about the weather. In this case, $\mathcal{I}^i(C, A^*) = \{I^i_{0.1}, I^i_{0.2}\}$, such that

$$I^i_{0.1} = \{(R, A^*), (C, A^*)\},$$
$$I^i_{0.2} = \{(R, B^*), (C, B^*)\},$$

where we indicate the time period and numeric label in the subscript.

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 $s^{\oplus}_t = (C, A^*)$ 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^*).

4.5 Mental attitudes

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 μ^s_i 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 μ^s_i 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 μ_i, we will slightly abuse notation and write, e.g., $\mu^s_i(Z) = \mu^s_i(H)$, 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 ω 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 μ_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_μ denote the space on which μ is a probability measure. Whenever for some event $E \in S$ we have $S_\mu \succeq S(E)$ (i.e., the event E can be expressed in space S_μ) then we abuse notation slightly and write

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

If $S(E) \npreceq S_\mu$ (i.e, the event E is not expressible in the space S_μ because either S_μ is strictly poorer than $S(E)$ or S_μ 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 ω: 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_μ is the space on which μ is a probability distribution. Since $t_i(\omega)$ now refers to agent i's probabilistic belief in state ω, 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 ω. This terminology is intuitive because at ω 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 ω is weakly less expressive than the space contains that ω. 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 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

[6] The example is taken from Schipper (2016) 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 Heifetz et al. (2013)).

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 θ^s without ambiguity.

[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 Harsanyi, 1967; Mertens and Zamir, 1985).

4.6 Planning

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.

4.7 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., Geanakoplos (1989), Samet (1990)). 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, μ_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 Γ^n:

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 σ_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 Γ^n that begin at s and terminate in X. Then σ_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 Γ^n that begin at s and terminate in X. Then σ_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]

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 unawareness. *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.