Intentional Awareness[1]

Brian Epstein
Tufts University, Medford

Michael D. Ryall
University of Toronto

December 16, 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 $X \backslash Y$ 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]

[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, 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 attitudes* of the individuals acting in that world. Let $S^{\oplus}$ denote the (finite) set of all possible objective states from $t = 0$ to $t = T$, with typical element $s \in S^{\oplus}$.

In addition to elaborating how the world actually exists at t, we also wish to consider counterfactuals and future possibilities—that is, to be able to discuss the way the world is, the way it might have been, and the way it could be. With this in mind, let $S_t \subset S^{\oplus}$ 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 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.

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

In terms of interpretation, it does no harm to imagine that a state, $s \in S^{\oplus}$, elaborates an uncountably infinite number of features of the world. However, we do assume the the number of states in $S^{\oplus}$ 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 Acts

In our approach the acts of Nature and the decision maker jointly cause the system to evolve from a state $s \in S_t$ to a new state $s' \in S_{t+1}$. 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 $s \in S^{\oplus}$, 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, $a_s \equiv (a_s^n, a_s^i)$, one by Nature and one by i. The set of *all objective act profiles at state s* is $A_s \equiv A_s^n \times A_s^i$. 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. Also, keep in mind that the act profiles in A_s are unique. The set of *all possible objective act 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 $A^{\oplus} \equiv \bigcup_{s \in S^{\oplus}} A_s$.

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

[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.

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, let $g^{\oplus} \equiv (S^{\oplus}, \lessdot)$ be the elaboration of all objectively possible sequences of states. Specifically, assume $g^{\oplus}$ is a directed, rooted tree with nodes $S^{\oplus}$ ordered by the precedence relation $\lessdot$ (i.e., the predecessors of each $s \in S^{\oplus}$ are totally ordered by $\lessdot$). Thus, $S_0^{\oplus} = \{s_0\}$, where s_0 is the root node at the beginning of time. Let $\omega : A^{\oplus} \to S^{\oplus}$ be a function mapping act profiles available at a given state to its children. Thus, if $\omega(a_s) = s'$, it means that when the act profile a_s is taken at s, it actualizes the immediately following state s'; e.g., if $s \in S_t$, then $\omega(a_s) \in S_{t+1}$. Assume ω is bijective from A_s to the immediate successors of s (the elements in A_s uniquely label the edges 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.[4] Because the relationship is bijective, we can also write $\omega^{-1}(s') = a_s$.

We assume that at the beginning of time, in state s_0, i is faced choosing from a number of mutually exclusive act-problem. 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, $A_{s_0}^i = \{p_1, \ldots, p_k\}$, the indexed set of k act-problems objectively available to i in state s_0. To refer to the act-problems available in a particular state, s, independent of the other feasible acts available there, we define $P_s \equiv \{p_1, \ldots, p_k\}$.[5]

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 $g^{\oplus}$. 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 which, abusing notation a bit, is $\omega^{-1}(h_{s_0}) \equiv (\omega^{-1}(s_0), \ldots, \omega^{-1}(s))$. Let $H^{\oplus}$ denote the set of objective *terminal histories* (i.e., the full-length, T-period histories).

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 exclusive act-problems: to play indoors or, alternatively, to go outside. If i decides to play indoors, then she must pick

[4]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.

[5]In later periods, $t > 0$, A_s^i may contain numerous acts in addition to choosing a problem from P_s.

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.

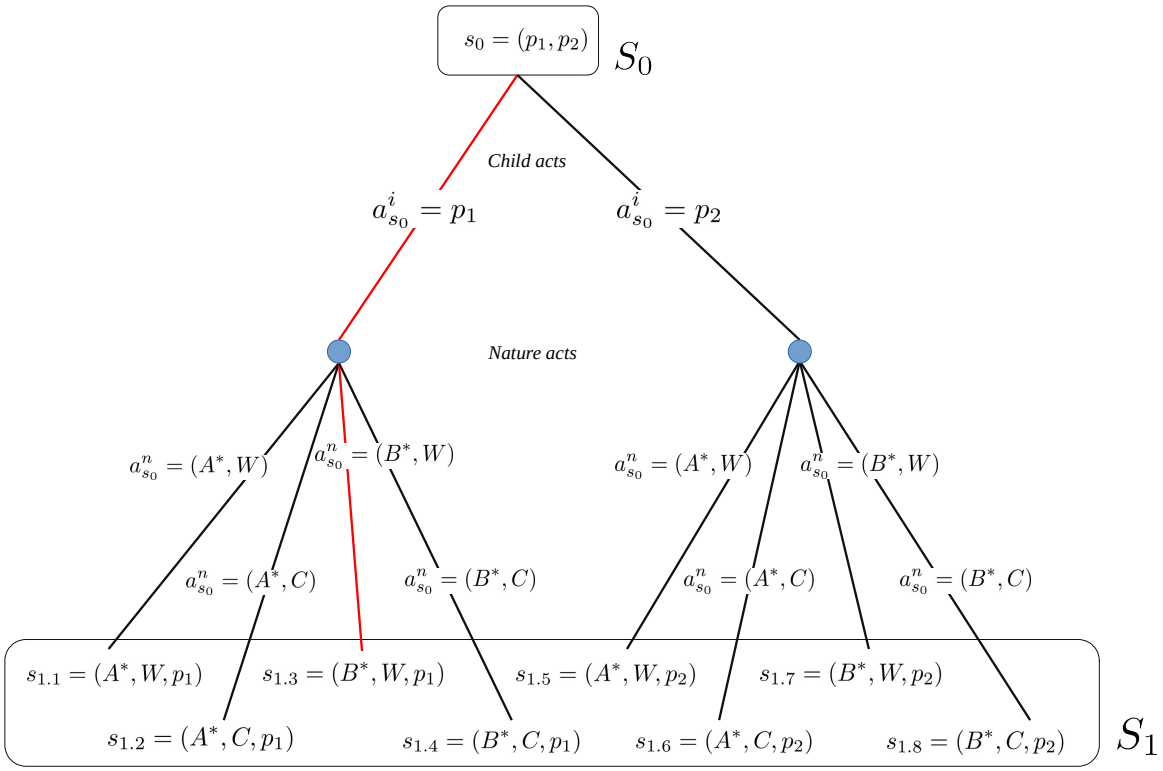

Figure 1: Evolution of Nature's state spaces from $t = 0$ to $t = 1$

At the start of time, the world presents the infant with a list of mutually exclusive act-problems from which to choose. Here, $P_{s_0} = \{p_1, p_2\}$ where p_1 is how to play indoors and p_2 is how to play outside. Therefore, i chooses $A^i_{s_0} = \{p_1, p_2\}$. Behind the scenes, Nature determines which toy is best and what the weather will be at time $t = 1$. Therefore, Nature selects one of among four possible acts: $A^n_{s_0} = \{(A^*, W), (A^*, C), (B^*, W), (B^*, C)\}$.[6]

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

[6]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.

best toy, and X, the weather, 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).

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; the weather is warm; and i is focused playing indoors and deciding which toy to choose. This state was actualized by the act profile $a_0 = ((B^*, W), p_1)$. The mapping 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 unfolding 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 the play-indoors problem," 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 the play-indoors problem, 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, $H^{\oplus}$. These refer to events that unfold over time. For example, the set of terminal histories in the Brian's child example

is

$$H^{\oplus} = \{(s_0, S_{1.1}), \ldots, (s_0, S_{1.8})\}.$$

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

$$\{(s_0, S_{1.1}), \ldots, (s_0, S_{1.4})\} \subset H^{\oplus}.$$

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 $H^{\oplus}$. 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.[7]

4 Awareness

Our basic approach is to assume that, in each objective state of the world, the individual has in mind a subjective version of $g^{\oplus}$. Specifically, for each $s \in S^{\oplus}$, define *i's awareness tree at s*, $g_s \equiv (S_s, \preccurlyeq_s)$, which is also a rooted, directed tree that represents i's subjective awareness of $g^{\oplus}$ when he is in s. Let $S^{\ominus} \equiv \bigcup_{s \in S^{\oplus}} S_s$ be the set of all subjective states in which i could potentially find himself. Note that each set of subjective states, S_s, is distinct from the others; i.e., the sets S_s form a partition of $S^{\ominus}$. The set $S_t^{\ominus} \subset S^{\ominus}$ is the collection of *all* the subjective states that could occur in period t, depending upon which actual state arises. When we know the objective state is s, then $S_{s,t} \subset S_s$ is the set of period t states of which i is aware according to g_s when i is in s.

Most of the notation for objective objects transfers to their subjective counterparts without ambiguity provided we identify which states are objective and which subjective (or understand these from the context). For example, $H_s^{\ominus}$ is the set of full-length histories according to g_s. In some cases, double state references are required, one for a subjective state and one for the objective state with which it is associated. Thus, given $s \in S^{\oplus}$ and $s' \in S_s$, $A_{s',s}$ indicates the feasible act profiles available at subjective state s' according to i's awareness as specified by the objective state s. Similarly, we extend the state-contingent actualization function so that for $s \in S^{\oplus}$ and $s' \in S_s$, $\omega(a_{s'}) = s''$ indicates that, according to i's awareness in objective state s, the subjectively feasible

[7] 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.

act profile $a_{s'} \in A_{s'}$ actualizes the succeeding subjective state s'' according to g_s. Assume that ω meets the same bijectivity conditions for each g_s as for $g^{\oplus}$.

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 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 $s, s' \in S^{\oplus}$, define the surjective *awareness mapping* $r(\cdot|s) : S^{\oplus} \to S_s$ where $s'' = r(s'|s)$ means that—when i is in objective state s—the subjective state s'' represents his awareness of some other objective state s'. In another minor abuse of notation, given an objective history $h_s = (s_0, \ldots, s)$, let $r(h_s) \equiv (r(s_0|s), \ldots, r(s|s))$ denote the history in g_s to which it corresponds.

The idea is that, typically, the individual is aware of only some subset of features of the objective state of the world at any given moment. Thus, r projects all the objective states with those features into a single subjective state that includes only those features. From this perspective, we can think of r as projecting the objective states associated with a particular synchronic event (which would be given by r^{-1}) into a single subjective state that represents the individual's awareness of that event. Importantly, to the individual, this is not an event, but a state of the world as he is aware of it. The subjective state is an impoverished version of the objective states that project into it: the individual is unaware of the additional features of world that would allow him to refine his ability to think about his situation. We impose the following consistency conditions on r:

1. Time consistency: if $s' \in S_t \subset S^{\oplus}$ then $s'' \in S_{t,s}$; e.g., i does not mistake something that happened yesterday with something happening today.

2. Act independence: for all $s \in S^{\oplus}$ and $s' \in S_s$, $A_{s',s} \equiv A^n_{s',s} \times A^i_{s',s}$; i.e., the set of act profiles includes all combinations of the acts by n and i of which i is aware.

3. No imaginary acts for i: if $s'' = r(s'|s)$, then $A^i_{s'',s} \subseteq A^i_{s''}$; i.e., although i may not be aware of all the feasible acts available to him in a particular state, the acts of which he is aware are truly feasible.

4. Dynamic consistency: the terminal histories in g_s correspond to a partition of the terminal histories in $g^{\oplus}$. That is, $\{s \in S_T | r^{-1}(h_s)\}$ is a partition of $H^{\oplus}$.

5. Perfect recall: the individual has perfect recall of his own past acts. Formally, given the objective period t state $s_t \in S_t$, if $r(h_{s_t}) = h_{s'_t} = (s'_0, \ldots, s'_t)$, then for each $s'_j, j > 0$ in

the sequence $h_{s'_t}$, the action profiles $\omega^{-1}(s'_j) = a'_{j-1}$ and $\omega^{-1}(s_j) = a_{j-1}$ agree on the i-act components. That is, if $a'_{j-1} = (a^{n\prime}_{j-1}, a^{i\prime}_{j-1})$ and $a_{j-1} = (a^n_{j-1}, a^i_{j-1})$, then $a^{i\prime}_{j-1} = a^i_{j-1}$.

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 $t = 1$. Suppose that, having decided to solve the play-indoors problem, i becomes aware of which toy is best—to the exclusion of the weather. In this state, i can reason about B being the best toy and the counterfactual that A could have been the best toy. We can also allow i to be aware that she could have chosen to solve p_2, which would have put her in a different state of the world. Then, $S_0^{\ominus} = \{s'_0\}$, such that $P_{s'_0} = P_{s_0}$, and $S_1^{\ominus} = \{(A^*, p_1), (B^*, p_1), (p_2)\}$.[8]

In this state, her awareness tree is $g_{s_{1.3}}$, as depicted in Figure 2. In this diagram, the top half shows $g^{\oplus}$ tipped on its side. The bottom half illustrates $g_{s_{1.3}}$. 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 weather. In particular, it is *not* the case that i is uncertain about what the weather might be. Rather, according to $g_{s_{1.3}}$, i is simply not thinking about the weather at all—it is not in her mind and, therefore, will not be 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 $S_t^{\ominus}$. Condition 2 is met because all combinations of the acts of which i is aware are included as act profiles in $g_{s_{1.3}}$.[9] Condition 3 is clearly met. Condition 4 is met, which can be seen from the fact the terminal states in $S^{\ominus}$ imply a partition of the terminal states in S_1 (and, hence, of $H^{\oplus}$). 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).

[8] Keep in mind, the states in i's awareness trees are mathematically distinct from each other and from the objective states. So, we write $S_0^{\ominus} = \{s'_0\}$ rather than $S_0^{\ominus} = \{s_0\}$ even though $P_{s'_0} = P_{s_0}$; i.e., s'_0 is not the same object as s_0

[9] 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.

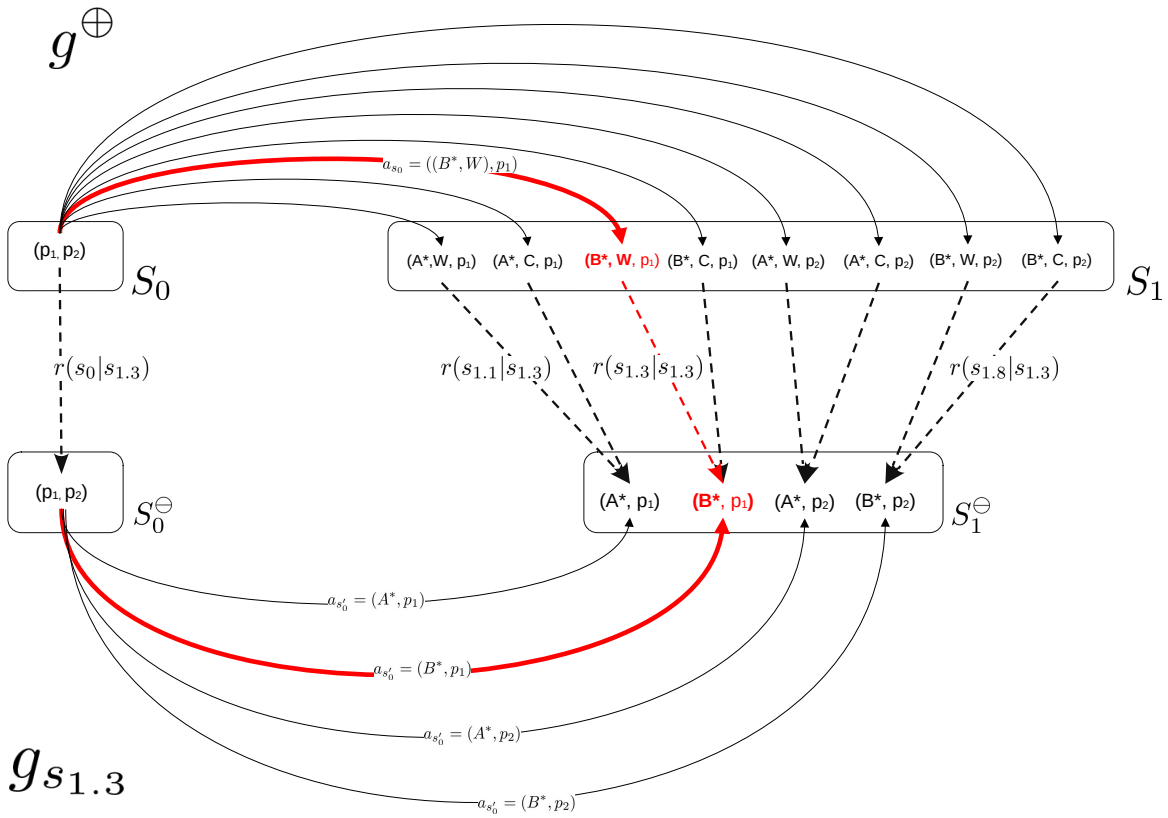

Figure 2: Awareness tree for Brian's Infant when the objective state is $s_{1.3} = ((B^*, W, p_1))$.

5 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 Information sets

For $t = 0, \ldots, T$ and each pair $s \in S_t$, $s' \in S^{\ominus}_{t,s}$, define the *information set at* s', denoted $I_{s',s}$, to be a subset of $S^{\ominus}_{t,s}$. Information sets serve to distinguish states about which i is aware but uncertain. Information sets are assumed to meet the following conditions:

6. No-delusion: $s' \in I_{s',s}$.

7. Introspection: if $s'' \in I_{s',s}$, then $I_{s'',s} = I_{s',s}$.

8. Act consistency: For all $s', s'' \in S_s$, if $A^i_{s',s} \cap A^i_{s'',s} \neq \varnothing$ then $A^i_{s',s} = A^i_{s'',s}$.

9. Distinct acts at disjoint information sets: if $s'' \in S^{\ominus}_{t,s}$ and $A^i_{s',s} = A^i_{s'',s}$, then $I_{s',s} = I_{s'',s}$.

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 must also be uncertain about whether s' or s''' is true). Condition 8 partitions the states in $S^{\ominus}_{t,s}$ into equivalence classes according to the acts available to i. Then, Condition 9 is required because, otherwise, i could discern states based upon differences in their feasible acts. Notice the implication that a state space in a given period is partitioned by its information sets.

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 double-dotted dashed lines.

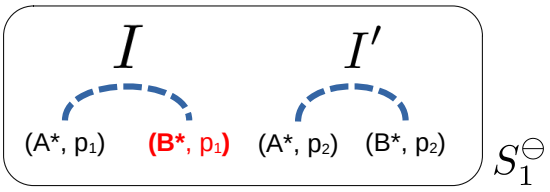

Figure 3: Fully aware child is uncertain about which toy is best

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

5.2 Beliefs

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

For example, consider $S_1^{\ominus}$ in Figure 2. The individual is in objective state $s_{1.3}$. Suppose $\mu_{s_{1.3}}$ assigns equal probabilities to all four histories that terminate there. The information sets in this case are trivial: each state is in its own, singleton set. Therefore, $\mu_{s_{1.3}}((B^*, p_1)|(B^*, p_1)) = 1$. In other words, 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. Now, $\mu_{s_{1.3}}((B^*, p_1)|I) = 0.5$.

Eventually, we will describe how these beliefs can be constructed from the plans of the individual and the behavior of Nature. For now, it is sufficient to understand that they are probability distributions on terminal histories and that they may vary state by objective state.

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

6 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 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.

7 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 $\gamma_i(s) = \varnothing$ 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 $\gamma_i^{s_t} = \varnothing$ 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 $\theta_i \equiv (\mu_i, D_i, \gamma_i)$.[10] 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.

[10] 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).

8 Plans

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.

8.1 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 G^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 G^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 G^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.