Notes on Group Agency[1]

Brian Epstein
Tufts University, Medford

Michael DD. Ryall
University of Toronto

December 9, 2020

[1]

1 Overview

This version begins with a model of individual agency in Section 3, then moves on to groups and group agency in the remaining sections. We are aiming for a formal framework that is fairly general, thereby allowing for a substantial degree of flexibility in the sorts of phenomena it can represent. The formalism for groups builds on the individual setup.

Our primary concern is to model groups and group agency. We begin with individuals and individual agency first of all because it is helpful to understand key aspects of group agency by way of analogy to the agency of individual people. Equally important is to be clear where group agency differs from individual agency. With formal models of both individual and group agency, these differences can be highlighted.

2 Notational conventions

Capital letters (G, N, etc.) refer to sets. Small Arabic and Greek letters refer variously to elements of sets (e.g., $i \in N$), functions (e.g., $\sigma : N \to \mathcal{N}$), and indexed lists (e.g., $x \equiv (x_1, \ldots, x_n)$). Terms are *italicized* at the point of definition. A *profile* is a placeholder for a list of elements; e.g., x where $x \equiv (x_1, \ldots, x_n)$. The "$\equiv$" symbol indicates the definition of a mathematical object. If X is a set, then 2^X is the notation denoting 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.

3 Individuals

Begin with a *population of individuals*, indexed by $N \equiv \{1, \ldots, n\}$ with typical element $i \in N$ and $\mathcal{N} \equiv 2^N$. When a group is formed, its members will be from this population. As we explain below, we assume the population is sufficiently large to accommodate many potential groups, not all of which will form. Moreover, not all individuals in the population need be in a group. This allows us, for example, to consider groups whose existence depends on individuals being in certain mental

states.

3.1 States

An important element of our analysis is a *state space*, denoted S, with typical element $s \in S$.[1] In our analysis, each element in S elaborates the status *of all features of the world that are relevant to all individuals and groups.* This includes all the relevant "mind-independent" features of a particular situation as well as of such "mind-dependent" features of the individuals in those states.

A note on terminology: the term 'event', which we define in the next paragraph, is used differently by philosophers than it is by game theorists. We will use the game-theorist's use of the term, but since we are writing to both audiences, it is important to clarify this difference. In game 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. But in game theory (and in our usage), an event is a set of states. Because an individual state encodes all relevant features of the world, events provide a way of identifying all the states consistent with a particular feature of interest.

The canonical example is rolling a six-sided die. The state space is $S \equiv \{1, 2, 3, 4, 5, 6\}$. Then, for example, the event "the die roll is even" is represented by $X \equiv \{s \in S | s = 2, 4 \text{ or } 6\}$ or, more simply, $X \equiv \{2, 4, 6\}$. In our context, 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. I THINK THIS STILL NEEDS WORK. IN PARTICULAR, I'M NOT SURE WHETHER A STATE IS REALLY A MOMENT OR TIME-SLICE OF THE WORLD. IF SO, I CAN TALK MORE PRECISELY ABOUT HOW 'EVENTS' ARE EXPRESSIBLE AS PROPERTIES FOR THE PHILOSOPHER. THE PROBLEM I'M HAVING WITH THE IDEA OF A STATE AS A TIME-SLICE OF THE WORLD IS THAT SOME PROPERTIES OF THE WORLD DEPEND ON HISTORICAL FACTS. FOR INSTANCE, A FOOTPRINT IS ONLY A FOOTPRINT IF IT WAS MARKED IN THE PAST BY A FOOT. SIMILARLY FOR LOTS OF INSTITUTIONS AND SOCIAL FACTS. SO IT'S A LITTLE PROBLEMATIC TO THINK OF STATES AS ACTUALIZED AT A PARTICULAR TIME. EVEN IF WE INCLUDE MEMORIES OF HISTORY

[1] Again, this setup can be generalized to include countably or uncountably infinite state spaces. Limiting attention to finite sets allows us to sidestep the mathematical complexities associated with measurability concerns.

AMONG THE PRESENT STATES, THAT IS NOT REALLY ADEQUATE TO CAPTURE THE CONSTITUTIVE ROLE OF HISTORICAL FACTS IN THE PRESENT.

An *event*, then, is a subset of S. Let $\mathcal{S} \equiv 2^S$ be the set of *events*, with typical element $E \in \mathcal{S}$.

States are assumed to be actualized through time in a fashion to be discussed below. At time t, the individuals in N find themselves in some *actualized* state of the world, s_t. An event $X \in \mathcal{S}$ is said to be *actualized* at time t if $s_t \in X$. For example, if s_0 is the state prior to the roll of the die and s_1 is the state in which the die roll is revealed, the event "the die roll is even" is actualized in $t = 1$ if $s_1 = 4$ (as is the state $s_1 = 4$). Except under conditions of extreme rationality, individuals have an imperfect grasp of the details of their actual state. For example, they generally will not know the mental states of the other individuals and may have an imperfect understanding of the consequences of their actions. Hence, they will know events but not the true state itself (like knowing the die roll was, in fact, even without knowing the exact number that made it so).

3.2 Acts

The sequence of states actualized over the period of analysis is effected by the acts of the individuals in the population in conjunction with acts of Nature (i.e., all the changes that, in addition to the acts of the individuals, determine actualizations of one state from a previous state).

*** I'M FINDING THIS TERMINOLOGY A BIT AMBIGUOUS, BECAUSE ACTS ARE DIFFERENT FROM ACTIONS. (ACTIONS SEEM TO BE A NARROWER CATEGORY.)ARE WE TO UNDERSTAND ACTS AS BEHAVIORS? THAT SEEMS TOO NARROW IN A DIFFERENT WAY. I'M ALSO UNCLEAR WHETHER AN INVOLUNTARY CHANGE IN AN INDIVIDUAL (LIKE HAIR GROWING) IS AN ACT OF AN INDIVIDUAL OR OF NATURE.

For all individuals i, the set of available acts is a state-contingent set: A_i^s indicates the set of *s-feasible acts available to* i with typical element $a_i \in A_i^s$.[2]

We adopt the convention that $A_i^s = \varnothing$ indicates individual i has no available acts in state s. An *act profile* is a list of acts, one for each individual, denoted $a \equiv (a_0, a_1, \ldots, a_n)$. Here, we add Nature as "Individual 0" so that a_0 summarizes the relevant contingencies that, in conjunction with the individuals' acts, determine which state is actualized following s. The set of *all act profiles at state*

[2]Since A_i^s is a set-valued function on S, we could use a more standard convention such as $A_i(s)$. Conceptually, the actions available to i comprise part of the information encoded in the description of the world represented by s. The function A_i^s can be thought of as "pulling out" this subset of information from s. We have several functions that serve this purpose for various state-related features. Therefore, we use the state superscript convention to highlight these special funcions (and, simultaneiously, reduce notational clutter).

s is $A^s \equiv \times_{i=0}^{n} A_i^s$. The set of *all possible act profiles* is $A \equiv \cup_{s \in S} A^s$.

3.3 Dynamics

Let $S_0 \subset S$ be the subset of potential period-0 states; i.e., those candidate states from which individuals believe the action *might* begin. Objectively, the action starts at $s_0^* \in S_0$, where the asterisk indicates the objective starting state. Beginning in s_0^*, agents choose some profile of acts $a \in A^{s_0^*}$ which then actualizes the next state s_1.

I THINK THE PRECEDING PARAGRAPH NEEDS MORE INTRODUCTION/CONTEXT. WHICH ACTION? ALSO NEED TO INTRODUCE THAT THERE'S GOING TO BE A SBUJECTIVE/OBJECTIVE DISTINCTION HERE AND WHY. AND WHY ARE THE "POTENTIAL PERIOD-0 STATES" ONES THAT PERTAIN TO INDIVIDUALS' BELIEFS? "POTENTIAL" SEEMS LIKE AN OBJECTIVE FEATURE. AND IS IT JUST ANY INDIVIDUAL, OR ALL INDIVIDUALS OF SOME SORT? (THAT IS, I DON'T REALLY UNDERSTAND WHAT S0 IS SUPPOSED TO BE.)

As indicated above, a summarizes all the contingencies(???TERMINOLOGY) required to determine the next state. Let $\omega : S \times A \to S$ be the *state-contingent actualization function*, where $\omega^{s_t}(a) = s_{t+1}$ indicates that if the act profile at state $s_t \in S$ is $a \in A^s \subset A$, then the next state actualized is s_{t+1}. Given the true starting state s_0^* and a sequence of action profiles, this setup allows us to identify the sequences of states that actually happen according to ω. Given a state s_t, if there exists an act profile $a \in A^s$ such that $\omega^{s_t}(a) = s_{t+1}$, we use a directed edge $s_t \to s_{t+1}$ to indicate that s_{t+1} is actualized by *some* feasible act profile available at s_t.

A *feasible path* is any sequence of nodes $p = (s_t, s_{t+1}, \ldots, s_{t+x})$ such that $s_t \to s_{t+1} \to \cdots \to s_{t+x}$ according to ω. We refer to a feasbile path that starts at some initial state $s_0 \in S_0$ as a *history*. The *history at time t* is denoted $h_t = (s_0, \ldots, s_t)$. An *objective history at t* is a path consistent with ω starting with s_0^*, denoted h_t^*. At the beginning of time, we start with some *null history* $h_0 = (s_0)$. Let H be the set of all possible histories, with $H^* \subseteq H$ denoting the set of all possible objective histories.

This setup allows us to represent all possible histories as a *directed graph*, denoted $\Gamma = (S, \omega)$, with nodes S and edges as determined by ω. The root nodes of Γ are the states contained in S_0. Given that states encode mental features, such as memory of previous states, *it is reasonable to assume that, whereas one state can branch into many children, it is never the case that a state has*

multiple parents. For example, suppose it were that case that $s_1 \to s_3 \leftarrow s_2$. Then, being at s_3 would imply the possiblity of some individuals experiencing an inconsistent mental state due to knowing two mutually exclusive histories simultaneously. I WONDER WHY THIS IS REQUIRED. IT DOES SEEM PROBLEMATIC TO THINK THAT NECESSARILY THERE WAS ONLY ONE HISTORY THAT COULD BRING US TO THE STATE OF THE WORLD TODAY. SURELY THERE ARE AT LEAST SOME STATES IN THE PAST THAT ARE IRRELEVANT TO THE PRESENT STATE OF THE WORLD. TO SAY THAT s1 -¿ s3 ¡- s2 IS TO SAY THAT THERE IS SOME FEASIBLE PATH FROM S1 TO S3 AND SOME FEASIBLE PATH FROM S2 TO S3. WHAT IF S1 AND S2 JUST DIFFER WITH RESPECT TO SOME PARTICLE THAT FLEETINGLY APPEARS IN S2 AND THEN DISAPPEARS? SO S1 AND S2 ARE DIFFERENT STATES, BUT BOTH ARE EQUALLY ELIGIBLE TO BE PARENTS OF S3.

The preceding implies that Γ is a collection of trees, one for each $s_0 \in S_0$. Let $Z \subset S$ denote the subset of terminal states (the leaves of the trees) in S, with typical element z. Then, S can be partitioned according to subsets of states corresponding to time periods: $S = S_0 \cup S_1 \cup \cdots Z$. Now, we represent a *complete history* in Γ as a sequence starting at some s_0 and ending in a terminal state, $h_t = (s_0^*, \ldots, z_t)$ where t is the t^{th} time increment required to reach z. This notation allows us to distinguish full histories from partial histories from paths. For example, $h_t = (s_0^*, \ldots, s_t)$ is a partial history. Alternatively, $p = (s_3, \ldots, s_9)$ is a path; it is not a history since it does not start with an s_0, though it is part of *some* complete history in Γ.

3.4 Mental attitudes

Let us consider the following mental features of individuals, which are encoded in each state. First, WE MODEL *beliefs* AS subjective probabilistic conjectures about the likelihood of the objective history, one's present state, and future events, represented by a probability distribution on states. In this approach, an individual is *certain* of an event when her belief assigns probability = 1 to it.[3] Second, *desires* are the individual's attitudes toward events, represented as a partial order relation on the set of paths in Γ, P. IN PHILOSOPHICAL TERMINOLOGY, ALL OF THESE ARE ATTITUDES, SO IT ISN'T ENOUGH TO SAY THAT DESIRES ARE ATTITUDES. CAN WE

[3]Some scholars equate "knowledge" and certainty in the probabilistic sense defined here. We leave open the possibility of defining knowledge according to some other criteria but maintain that, however one defines it, certainty of an event is an implication of knowing it. I THINK WE SHOULD SKIP THIS LAST SENTENCE. I DON'T THINK KNOWLEDGE DOES IMPLY CERTAINTY, AND ANYWAY EVEN IF IT DID, THAT WOULD NOT IMPLY THAT PROB=1 IMPLIES KNOWLEDGE.

JUST SAY THAT DESIRES ARE MODELED OR REPRESENTED AS A PARTIAL ORDER RELATION, RATHER THAN SAYING THAT THEY ARE ATTITUDES TOWARD EVENTS? Third, *intentions* represent an agent's commitment to undertake a plan of action designed to actualize an event. HERE YOU USE THE TERM 'REPRESENT', WHICH IS DIFFERENT THAN YOUR USE OF 'REPRESENT' IN THE PREVIOUS SENTENCE. HOW ABOUT: WE UNDERSTAND AN INTENTION TO BE AN AGENT'S COMMITMENT TO UNDERTAKE A PLAN OF ACTION DESIGNED TO ACTUALIZE AN EVENT. BELOW WE WILL DISCUSS HOW THESE ARE TO BE REPRESENTED. Thus, the object of an intention is an event. Individual consistency conditions are required to align beliefs, desires, intentions and action plans.

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_i^s to indicate i's subjective probability distribution on H at state s. This distribution induces a distribution on history events, $\mathcal{H} \equiv 2^H$. Note that each μ_i^s induces a probability distribution on S. For example, the probabilities of the elements of Z (terminal nodes) are equal to the probabilities of the complete histories they terminate. I'M FINDING THIS RATHER DIFFICULT TO FOLLOW. 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., $\mu_i^s(Z) = \mu_i^s(H)$, even though $Z \in \mathcal{S}$ while $H \in \mathcal{H}$.

It is important to note that the existence of more than one element in S_0 means that individuals may be uncertain about which tree is the objective one and, hence, the true history they have experienced. DOES THE DEFINITION OF BELIEF IMPLY THAT WE HAVE BELIEFS ABOUT EVERY PAST STATE OF THE WORLD? WE SHOULD DISCUSS THIS. 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.

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. WOULDN'T THIS BE TAKEN CARE OF BY THE HISTORY ENCODED IN THE PRESENT STATE, WITH THE CURRENT MODEL? I AGREE THAT IT WOULD BE BETTER TO TREAT THIS AS PATH PREFERENCE RATHER THAN PREFERENCE OF ONE STATE OVER ANOTHER, BUT IT SEEMS AS THOUGH WE MAY HAVE TWO REDUNDANT WAYS OF TREATING THIS PREFERENCE. 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. INTERESTINGLY, GABRIEL IS NOW GOING THROUGH A STAGE WHERE HE INDIVIDUATES THE THINGS HE DESIRES NOT BY THE END STATE BUT BY THE PAIR OF THE END-STATE AND THE AGENT DOING IT. I CLOSE THE BLINDS, AND HE SAYS, "NO, MOMMY DO IT!" SO SHE HAS TO OPEN THE BLINDS AND THEN RE-CLOSE THEM. 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. CAN'T AN INDIVIDUAL HAVE MULTIPLE INTENTIONS IN A GIVEN STATE? We assume that individuals have desires and beliefs in all states, but not necessarily intentions. I'M NOT CLEAR ON WHETHER YOU'RE

SAYING THAT INDIVIDUALS NEED NOT HAVE INTENTIONS AT ALL IN A GIVEN STATE, OR WHETHER YOU'RE TALKING ABOUT DESIRES, BELIEFS, AND INTENTIONS AS APPLIED TO ALL STATES. IT SEEMS THAT IN THE MODEL, EVERYONE HAS A COMPLETE SET OF DESIRES AND BELIEFS REGARDING ALL STATES, RIGHT? 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 $\theta_i \equiv (\mu_i, D_i, \gamma_i)$.[4] 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 θ_i^s and θ^s without ambiguity.

3.5 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 $i \in N$, $s_t \in S$, $\mu_i^s(h_t^*) > 0$.*

[4] 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).

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:

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 posessing 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 that begin at s and terminate in X. Then σ_i is said to be weak means-ends consistent with γ_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 that begin at s and terminate in X. Then σ_i is said to be strong means-ends consistent with γ_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]

4 Groups

To begin, let $M \equiv \{1, \ldots, m\}$ be a finite set indexing all possible *groups*.[5] Let $\mathcal{M} \equiv 2^M$. All the indices correspond to entities that could exist as social groups. Therefore, when k does not exist we refer to it as a *potential group*; when k does exist, we simply refer to it as an *existing group*.

4.1 Group composition and existence

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

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

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

[5] Later, we can make all the sets infinite if necessary. In the meantime, remember that M can be very large indeed.

5 Initial conditions

5.1 Modest social groups

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

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

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

5.2 Analytical sequence

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

Therefore, assume that the initial state of the world is $s_0^* \in E_\varnothing$, a state in which no groups exist. The profile of mental features is a primitive of the model. Therefore, everyone begins with mental states $\theta(s_0^*)$. These imply a profile of intended actions $a(s_0^*)$. According to these primitives, in a fashion not yet described, some new state of the world, s, obtains in which the groups $e(s)$ come into existence along with the updated mental features $\theta(s)$. Our task is to identify how these all hang together in a coherent metaphysics.

5.3 Human acts

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

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

5.4 Discussion

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

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

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

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

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

5. The *event* that $i \in N$ knows that the individuals G intend to join k: let $\bar{E}_i(s)$ denote the support of $\mu_i(s)$. Then, $K_i(E_{G \to k}) \equiv \{s \mid \bar{E}_i(s) \subseteq E_{G \to k}\}$, where K_i denotes events determined by what i knows in their states. Thus, $K_i(E_{G \to k})$ is the collection of states in which, given μ_i, i knows $E_{G \to k}$.

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

7. $E_{G\to k}$ is *common knowledge* at $s \in S$ if and only if there exists an event E such that: $s \in E$ and, for all $i \in N, E \subseteq K_i(E)$ and $E \subseteq K_i(E_{G\to k})$. This is the Monderer and Samet (1989) formulation, which is a restatement of Aumann (1976) in terms of evident events. For example, E can be the event "The individuals G publicly and credibly announce their intention to join k." This announcement is evident to everyone (for all $i \in N, E \subseteq K_i(E)$) and, once it occurs, it implies that everyone knows the individuals G will act to join k, knows that they know, that they know that they know that they know, etc. (for all $i \in N, E \subseteq K_i(E_{G\to k})$). Note that $E_{G\to k}$ is not necessarily evident knowledge: it is possible to have some state $s \in E_{G\to k}$ in which not everyone knows $E_{G\to k}$.

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

6 Group formation

Since we only have in mind such simple group activities as"we take a walk to NYC" we can think of a fairly simple sequence of acts and consequences that appear to be implied by them. Let us roughly follow (Bratman, 2014, Ch. 2) to see how this setup relates.

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

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

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

Then, Bratman suggests that the key is framing the group intention as "we each intend that we go to NYC." This is where we run into problems because what is being "intended" is vague and, in any event seems to be doing too much lifting. In our framework, an individual can intend his or her own acts – full stop. They cannot intend the intentions or actions of others. In our construction, Bratman's sentence of intention is nonsensical.

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

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

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

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

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

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

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

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

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

6.1 Unawareness Structures

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*.

Geanakoplos, J. (1989). Game theory without partitions, and applications to speculation and consensus. Technical report, Cowles Foundation Discussion Paper 914.

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.

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.

Monderer, D. and D. Samet (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior 1*(2), 170–190.

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