We define negation as your intended equivalence class of state transitions

My action in response to your signal does not lie in your intended equivalence class of state transitions, but you realize that it is an accepted or valid state transition [[given the Fregean sense of your signal]].

The Fregean sense of an action i.e. a signal defines an equivalence class of state transitions, but my understanding of the Fregean sense may only be an intended set, the core set, the obvious set of state transitions. And I assume that that is the set that you will perform. If you then do something that is not in that set but I realize that it is in the extension after all, but it wasn’t in the core set, then let us define that as negotiation of the meaning of the signal.

In this world, we now can define goals as (if we have a language for configurations) as pieces on top on of pieces in places.

I am looking for a 2x2 blue square.

We now have a generative language for specifying goals and can run tests against each other to see if we understand each other.

Negotiation ex: I put down the first blue square. I am expecting you to put the piece down in a particular set of places. You do something different, but consistent with my utterance, but not what I expected. Then I realize that it is OK also.

Another term (to decide on term) I say: “the goal is to build a 2x2x2 cube.” I say “put the blue square beside the blue square. For me, beside is different than on-top. You put it on top. What I realize is that on top is not legal with respect to beside, but it is legal with respect to on-at trajectory to the goal.

Has to do with the precedence and importance of subgoals and the global goal.

What we have learned:

* The notion of cooperation is ambiguous between doing exactly what you say and acting in the spirit of what you say. The spirit is “let’s work together to get to the goal.” The exact thing is doing exactly what I said.
* Why do I have to do it your way? This is the distinction between strict cooperation and loose cooperation.

We can do this with a block experiment and ethnographic analysis. Clark style.

Blocks on both sides and curtain between them, etc. Try to build the common thing.

Give a list of terms to use as the signals.

Coordination is a process by which multiple agents work together in some way. Although most classical instances of coordination involve a joint goal, this need not be the case. There need only be mutual benefit that can be gained by the interaction thus yielding a motivation for cooperation. In this paper we aim to analyze coordination as a general phenomenon through an abstract formal model. We will first motivate the need for such a theory and model of coordination, and then we will introduce the model in its most basic and paired-down form. We will then iteratively show how the model applies to various coordination scenarios, show how it does not apply to some others, introduce new machinery to generalize it to those new scenarios, show how it applies to those new scenarios and so on until we reach the general form of our model of coordination. We will then demonstrate the model in comparison to its competitors, concluding that we have provided a general theory of coordination that has specific empirical and theoretical bearing on specific instances of coordination from simple games to full communicative scenarios. As our model is a proxy for a theory, we will conclude by integrating our new theory of coordination to other literature on similar subjects showing that we can capture known empirical phenomena and theory in a more unified framework.

Need to introduce the work by mentioning the problems with current work on coordination

* Various properties of the signaling games
* A shared discrete set of meanings
* Infeasable models of learning
* Explicit cues related to understanding without explaining the mechanisms for inference about understandings
* Unable to capture verification and mind-reading phenomena as well as inference about other minds.
* We tend to think of meanings as things that can be grasped and therefore are associated with the language (talk to David about the sources), but actually they are things to be shared, to be aligned. This change in concept is essential to grasping broader and more accurate understanding of communication and coordiantion.
* They are not philosophically accurate or precise (since Lewis, who is outdated in numerous respects)
* THERE IS DEFINITELY SOMETHING ELSE

Coordination is often modeled via classic signaling systems. However effective such models are at capturing a number of important phenomena, there are an equal number that they are fundamentally unable to represent.

The mechanisms of coordination can be captured by a factorization of the causal processes and relations between the agents and the environments into a particular form. Initially, we will constrain the processes we consider and factor them into a particular Markov process. We will then successively generalize various factors until we have a general model of coordination.

[Perhaps I need to briefly discuss my modeling philosophy: that the analytic model should mirror the synthetic phenomena in parallel, and should provide sufficient flexibility to capture the observed phenomena and avoid unmotivated formal and analytic structures that diverge from the empirical.]

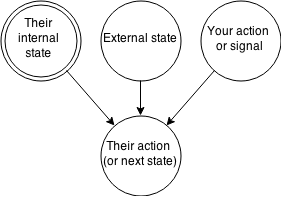
[Also note that we do not wish to derive formal semantics either current or new, but to derive a more accurate model of how language gets meaning in which we can investigate how specific meanings exist and could have arisen.]

The essential model will apply to coordination scenarios where the agents take turns acting on the environment in order to progress towards a joint goal. We do not presently distinguish actions that are on the environment and designed to progress towards the joint goals (pragmatic actions) from those that are designed to enable improved singular or joint understanding or thinking about the situation (epistemic actions) from those that are designed simply to communicate something to the other agent (communicative actions), although there is reason to believe that there are interesting phenomena related to these differences (Some Kirsh paper). What we do distinguish, however, although we will not yet explore the details of this distinction, is the difference between ephemeral and persistent state properties. Ephemeral state properties are those introduced by the action that are part of the current decision between actions for either agent, but that do not persist in the state by themselves. A clear example of this is a speech act, which may change the state, either internal or external to the agent, but does not itself persist in the environment. Persistent state properties, on the other hand, are those that do continue to exist in the environment. We will consider later how different types of actions either introduce persistent state properties (such as the change in position of a block) or ephemeral state properties (such as the flashing of a light).

In addition to the differences between actions and the state changes they induce, we also consider a distinction between the observability of different states. We consider the *internal state* of an agent to be those properties that affect the causality of the situation (most things) which are observable only to that agent and not to others. The prototypical example of this is the canonical mental state. Similarly, we consider the *external state* to be that which is observable to all agents. We simplify the situation by not considering properties that are observable to a subset of the agents involved, though there is no reason to believe that the introduction of these properties would significantly affect our conclusions about simple coordination scenarios. This will become a useful distinction because it separates the features of coordination by their epistemic properties and similarly their computational ones. We will find that these epistemic and computational distinction set us up to provide a very general theory of coordination while minimally sacrificing accuracy and that they will naturally enable us to account for known phenomena essential to coordination such as *common ground* in a principled manner.

[Finish explanation of the basic modeling properties before generalization]

* Assume cooperativity
* Introduce the causal graph
* Explain how the causal graph gets extended iteratively into a turn-wise scenario
* Briefly explain inference mechanisms
* Briefly mention what will be extended later including internal states, generalization, explicit internal state (or belief) inference, and a less structured model of causality that progresses continuously without requiring explicit turn-taking



To begin, we will consider the simplest examples of coordination: those where the actions of agents are determined only by the external state (the jointly observable component) and the last action of the other agent.

Need to assume cooperativity,

No internal state in a reactive system. In this initial case, the other agent is a purely reactive system, a stimulus-response agent.

I have an intended Markov transition, I want to infer whether you conformed to my intentions.

Initial definition of coordination:

**Actions are defined as state transitions.**

**My intentions are defined by an equivalence class of state transitions.**

**Coordination is achieved when your action is contained in my intentions.**

**The intended contribution of a signal is the set of acceptable state transitions for an agent.**

**The meaning of a signal is the set of state transitions that an agent would deem to be consistent with the signal.**

We have the same meaning when your equivalence class is the same as my equivalence class.

But we need to take into account negotiation:

**Acceptable actions with respect to the goal are also defined by an equivalence class of state transitions.**

**Coordination (or cooperation) with me is achieved when your action is contained in my set of acceptable actions.**

**Negotiation occurs when your action is initially outside of my intentions, but is within my set of acceptable actions.**

[Is there any reason to have this initial intentions and negotiation in general model? Are they simply mechanisms by which we reduce the complexity of the computation by separating the prediction and verification mechanisms?]

[Introduce the notion of a coordinating structure. This is what you coordinate on. It should be introduced in a sufficiently general way such that it captures the simple structures used initially, but makes room for more abstract structures.]

An additional and important consideration, especially for later extensions of our model of coordination, is the notion of a coordinating structure. A coordinating structure is what the agents coordinate on. A coordinating structure determines the set of acceptable actions for each agent in the coordination scenario. The set of acceptable actions may or may not be the same for each party, but each agent must agree upon the set of acceptable actions for the other agents including himself. Such a structure must be mutually agreed upon implicitly, as the set of actions you perform must align with the set of actions I find it acceptable for you to perform and vice versa. In the simplest case, the coordination structure is simply the set of sets of acceptable actions for each agent applied to the other agents. [Should graphically illustrate this], but in more complex cases, the coordinating structure can determine these sets of acceptable actions in another way. The coordinating structure captures the essence of the coordination scenario and **coordination is achieved if and only if the each agent’s actions are in accord with those specified by the coordinating structure throughout the coordination scenario.** (Maybe note the similarity to Clark’s notion of common ground and how this will help us to capture common ground and how it is established later in a principled formal manner.)

Furthermore:

**To understand a signal is to intend to perform an action or action sequence that is within the other agent’s intention set for that signal. Understanding a signal and performing this intended action successfully is therefore to successfully coordinate.**

A simple example of non-coordination occurs when, given the mutual goal that we want to build a tower of blocks, I say, “put the blue block on the ground.” There are multiple blue blocks of different sizes so you put the small blue block on the ground. However, I know that a tower needs to be started with the largest block and therefore I intended the largest blue block as the referent of my expression. Your action is consistent with my utterance meaning, but is not in line with my intended contribution, nor does it result in coordination as it is outside my intention set.

However, let us instead imagine that, equal in size to the largest blue block is a red block. In response to my utterance, “put the blue block on the ground,” you put the largest red block on the ground. Though this is inconsistent with the meaning of the utterance, it is within my set of acceptable state transitions, because all I care about is the fact that the largest block gets put on the ground. This is an example of negotiation.

But the inference about coordination success can progress gradually. Let us instead imagine that I tell you “build a tower of blocks alternating between red and blue,” and that all of the blocks are the same size. You begin by placing a blue block on the ground. At this point you appear to in coordination, although I gain very little inference about whether you are in coordination, as your only initial action that could have been inconsistent with this utterance would be to abstain from placing any blocks at all. You then put a second red block on top of the first. You still appear to be in coordination, and now I am more likely to believe that you understood my utterance. You then put a blue block on top and my inferred probability of understanding continues to rise. This is because it is increasingly unlikely, as you continue to perform actions consistent with my intentions, that your conforming to my intentions is a suspicious coincidence. It becomes increasingly likely that it is due to understanding my signal. If you complete the tower with a fourth red block, then you have successfully coordinated with me. If instead, you complete the tower with a fourth blue block, then, despite my earlier predictions, you have failed to conform.

We will see later, both theoretically and empirically, that we interpret utterances, not on their form alone, but on the contingency between the utterance and the state, and more precisely, between the utterances and various important properties of the state such as common ground, beliefs, desires, inferences about the beliefs of the other agent, goals, and more. Though this is equivalent in a mathematical sense to saying that an utterance is interpreted in terms of the context, factoring this relationship into contingencies yields a much clearer picture of the role of the context and on how to incorporate contextual inference into both theoretical and practical coordination and communication systems.

[Discuss the factorization of the contingencies and causation.]

[The action is an ephemeral state property, and the external state is a set of persistent state properties. Any information or interaction caused by ephemeral state properties only have effects beyond the current time step if they are incorporated into the persistent components of the state.]

In addition to these highly simplified experimental tasks, we can also show that this framework applies very naturally to a variety of real-world communication and coordination scenarios.

[Discuss the examples (chess, Chipotle, etc.) from the frame of this formal model]

[Sheep herding as coordination?]

[Do this with Blackjack and a set of alternative point goals. “In their country, in each casino people play blackjack to a different number.”]

We can demonstrate the wide applicability of even this extremely simplified model of coordination by applying it to a set of scenarios. The first scenario is a chess game between two players. Where a Lewis-style model of coordination was applied only to Von Neumann and Morgenstern’s games of pure coordination, we can apply ours to a wider range of more adversarial games such as chess (though it may be possible to do the same with theirs). The key to this lies in the realization that players in a game of chess must still hold to our assumption about cooperativity, which in this case is with respect to the rules of the game. When coordinating in a game such as chess, the coordination is not at the level of the pay-offs of the individual moves, but in playing by the rules. When I make a move, I do not make an arbitrary one, but only those that I believe are deemed legal by the mutually and implicitly agreed upon set of rules defined by chess.

We model chess similarly to how we modeled the example of telling someone to build a tower. By beginning the game, each player begins with a set of acceptable actions that the other player is to follow. It just so happens that this set of acceptable actions is often the same for both players and that for each player, the set of acceptable actions they impose on the other player is the same as the set that they impose on themselves. This is not a necessary feature of the model, but a property of the game of chess and of a larger class of coordination scenarios where the set of rules one imposes on one’s own actions is the same as the set that they expect the other agent to follow. Due to the structure of chess, this property is essential to keep the game fair.

What is different about chess from the block stacking game is that the set of acceptable actions changes at each move. Just like in our more general model, this set is dependent on the external state of the world. However, the set of acceptable actions is defined on a higher level coordinating structure, namely the set of rules of the game. Though, throughout the game, we always expect the other agent to play by the rules, the rules impose a different set of acceptable actions on different states of the game (in chess the game state is almost entirely captured by the state of the board, but not entirely due to castling and en passant). This property of an abstract coordinating structure is but a simple extension to the more basic coordination problem. In the case of an abstract coordinating structure, the set of acceptable actions is simply specified by the abstraction coordinated upon instead of directly.

We can work through an example of a chess game to better illustrate this principle. We will take the perspective of a player A, whose opponent is player B. Though we only take one perspective in this game, it is worth noting that the other perspective plays out in exactly the same way, simply with respect to his own set of intentions and acceptable actions.

[Play out the chess game to illustrate the abstract coordination structure. Additionally, it may be helpful to do a Bayesian analysis of this to capture the agent’s calculated probability of coordination upon a set of rules. It may also be useful to go through how an agent might learn about the transition probabilities in this simple model to make it easier to later add a more complex one.]

…

[This next section is on modeling with a Bayesian approach]

Because of the higher-order coordination structure involved in chess (namely, the rules), a simple observation of the contingencies between states and actions is not sufficient to confirm coordination. This is because the coordination structures and therefore the coordination scenarios do not persist for a single action like in the first block stacking example, but for a series of actions in an extended coordination scenario. Because of this extension into multi-turn coordination scenarios, coordination must be inferred by observing whether agents continue to act within the set of acceptable actions as determined by the coordination structure. In the case of chess, this is equivalent to the observation of whether your opponent continues to act within your understanding of the rules.

Instead of simply observing whether the other agent’s actions were so far consistent with the coordinating structure, we might want to infer the likelihood that they are operating in accordance with the coordinating structure. This is equivalent to inferring the probability that the other agent is in coordination with you: often an important inference to make. In order to do this, we can take a classical Bayesian approach, which fits the scenario quite nicely and is equivalent to an information-theoretic approach that we will not illustrate (I think it is equivalent).

We can begin by imagining the set of acceptable actions as a hypothesis about the rule guiding the behavior of the other agent. If it is indeed the case that the other agent is in coordination with me, his actions will be governed by the set of acceptable actions determined by the coordination structure.

[The other agent’s choice of action with respect to the coordination structure is not uniform in the set of acceptable actions, but guided by other goals. These goals need to be taken into account when doing inference. Because of this, I may wish to introduce this Bayesian framework first with respect to the block stacking scenario and then later extend it to take into account this joint inference between coordination and goal.]

[For the joint inference between goal and coordination: Given a constraint and a goal, we have a probability distribution over actions. Given an action, we have a probability distribution over a constraint and a goal? Given a set of actions, we can do joint inference? It is possible that the action of the agent would have been optimal for a variety of constraints given a goal, but it is also possible that it would not. It is possible that the action of an agent would be optimal for a variety of goals given a constraint, but it is also possible that it would not.]

[Go through the Bayesian approach including the alternate hypotheses, the size principle, the prior probabilities, etc.]

[Discuss the Chipotle game and coordinating on a set of ordering procedures. This game involves an expectation set and a set of acceptable actions as well as an asymmetry between the two players that illustrates asymmetric coordination structures. It also provides an opportunity to further emphasize the importance of contingencies.]

**[Discuss cases where the coordination structures are exchangeable] [Discuss cases where the coordination structures are not exchangeable, but are exchangeable with respect to the observed data] [Discuss how this becomes unlikely when there is complex cross-checking and interdependency] [Philosophically discuss how the cases that are not dissociable and are exchangeable correspond to inference about phenomenological properties such as color phenomena.]**

**Why this model is not general and generalizing the model by introducing the notion of generalization (and the hidden state)**

Though the scenarios so far have been captured in large part by our simplified model of coordination, there is a large class of scenarios that are more complicated and cannot be captured in this way. Moreover, the set of scenarios so far were also analyzable with respect to the classical and generally-accepted model of coordination introduced by David Lewis’ *Convention*. We will show that the ones that follow are not and that only our model can be extended in a simple and natural way to capture the unsofarcaptured aspects.

A simple scenario that defies prediction by our model involves that of a patient in a doctor’s office.

[Need to show later how the state machine model of an agent can be factored into a simple MDP for an agent (I think mostly) and how the addition of an interior state generalizes it to a POMDP (Again I think mostly).]

[Potentially add a note about the perspectives imposed by such a model and how each model captures the perspective of an agent who is a part of a larger system.]

[How are coordination structures set up? They are set up through communication in the same way that they are verified. Go through a set of examples and work through this duality of set-up and use of coordination structures and show how it enables a fluid process by which we can communicate and coordinate. Talk about this process in a discourse and with language. This setting up of coordination structures enables the communication and constant verification of coordination structures. ]

[An utterance sets up a coordination structure by imposing a set of expected and acceptable responses on the other person. These expectations lie solely in the mind of the utterer or first party. The other person then responds to the utterance according to their *interpretation* of its meaning. This interpretation lies entirely within the second party. If the response is within the set of acceptable actions set by the first party, then an instance of coordination has occurred. If all the responses that the second party considered appropriate (and may have performed counterfactually) are in the set of acceptable responses of the first party and if the entire set of acceptable responses of the first party is contained within the set of actions the second party considered appropriate, then there is ***perfect conceptual alignment*** between the first party and the second party with respect to the meaning of the utterance of the first party.]

Perhaps move this to the beginning: [Inferring the coordination of the other party is closely related to asking “did they understand me?” To infer this understanding, there are a number of things that must be taken into account. First is: What do you expect them do understand and what do you expect them to do given this understanding? Are these expectations for a single action? Or for a longer period of time? While they are acting, it is most important to ask: Is what they are doing consistent with an understanding of what I said or attempted to communicate? If so, you can ask: How much evidence does their action give about whether they understand or not? The strength of this evidence is dependent on a number of factors including their goals, the size of the set of the possible actions, the size of the set of consistent actions, and the probabilities of each of the actions independent of the context and goals and given the context and goals, but the key thing to compute is the probability that they would have committed that action given that they did not understand you. If you can compute this, then it is simple to infer whether they did understand you. Of course, if they commit a series of actions, then this evidence accumulates. This can all be elegantly captured using Bayesian probability.]

[We can also use the dual nature of actions in setting up coordination structures and in assessing coordination to argue that the model captures a smooth transition between language acquisition and language use not present in traditional formal accounts. It accounts for phenomena such as discourse alignment and intraconversational allusion where people learn and agree upon usages of words on the fly in a discourse. This is a phenomenon not presently captured in any formal account of communication.]

[How does this relate to theory of mind? Perhaps discuss the with the doctor patient example?]

[What about the case where you do not believe someone that they understand you? Simply ask them to extrapolate beyond what you gave them. If they provide a detailed extrapolation that is a priori unlikely to be generated (not generally applicable) that is consistent with your understanding, then it is extremely unlikely that they generated this without understanding you. This is how we verify understanding, at least in a highly explicit manner.]