Joint Behavior and Common Belief

Meir Friedenberg
Department of Computer Science
Cornell University
meir@cs.cornell.edu

Joseph Y. Halpern
Department of Computer Science
Cornell University
halpern@cs.cornell.edu

For over 25 years, common belief has been widely viewed as necessary for joint behavior. But this is not quite correct. We show by example that what can naturally be thought of as joint behavior can occur without common belief. We then present two variants of common belief that can lead to joint behavior, even without standard common belief ever being achieved, and show that one of them, *action-stamped* common belief, is in a sense necessary and sufficient for joint behavior. These observations are significant because, as is well known, common belief is quite difficult to achieve in practice, whereas these variants are more easily achievable.

1 Introduction

The past few years have seen an uptick of interest in studying cooperative AI, that is, AI systems that are designed to be effective at cooperating. Indeed, a number of influential researchers recently argued that "[w]e need to build a science of cooperative AI ... progress towards socially valuable AI will be stunted unless we put the problem of cooperation at the centre of our research" [6].

One type of cooperative behavior is *joint behavior*, that is, collaboration scenarios where the success of the joint action is dependent on all agents doing their parts; one agent deviating can cause the efforts of others to be ineffective. The notion of joint behavior has been studied (in much detail) under various names such as "acting together", "teamwork", "collaborative plans", and "shared plans", and highly influential models of it were developed (see, e.g., [2, 4, 10, 11, 15, 22]). Efforts were also made to engineer some of these theories into real-world joint planning systems [21, 18]. Examples of the types of scenarios these works considered include drivers in a caravan, where if any agent deviates it might lead the entire caravan to get derailed, and a company of military helicopters, where deviation on the part of some agents can lead to the remaining agents being stranded or put in unnecessarily high-risk scenarios.

All the earlier work agrees on the importance of beliefs for this type of cooperation. In particular, because each agent would do her part only if she believed that all of the other agents would do their part as well, there is a widespread claim that *common belief* (often called *mutual belief*) of how the agents would behave was necessary. That is, not only did everyone have to believe all of the agents would act as desired, but everyone had to believe everyone believed it, and everyone had to believe that everyone believed everyone believed it, etc. This, they argued, followed from the fact that everyone acts only if they believe everyone else will. (See, e.g., [2, 4, 10, 11, 15, 22] for examples of this claim.)

As we show in this paper, this conclusion is not quite right. We do not need common belief for joint behavior; weaker variants suffice. Indeed, we provide a variant of common belief that we call *action-stamped* common belief that we show is, in a sense, necessary and sufficient for joint behavior. The key insight is that agents do not have to act simultaneously for there to be joint behavior. If agent 2 acts after agent 1, agent 1 does not have to believe, when he acts, that agent 2 currently believes that all agents will carry out their part of the joint behavior. Indeed, at the point that agent 1 acts, agent 2 might not even be aware of the joint action. It suffices that agent 2 believes *at the point that she carries out her part*

To appear in EPTCS.

of the joint behavior that all the other agents will believe at the points where they are carrying out their parts of the joint behavior ... that everyone will act as desired at the appropriate time. If actions must occur simultaneously, then common belief is necessary [9]; the fact that we do not require simultaneous actions is what allows us to consider weaker variants of common belief.

Why does this matter? Common belief may be hard to obtain (see [9]); it may be possible to obtain action-stamped common belief in circumstances where common belief cannot be obtained. Thus, if we assume that we need common belief for joint behavior, we may end up mistakenly giving up on cooperative behavior when it is in fact quite feasible.

The rest of the paper is organized as follows. In the next section, we provide the background for the formal (Kripke-structure based) framework that we use throughout the paper. In Section 3, we give our first example showing that agents can have joint behavior without common belief, and define a variant of common belief that we call *time-stamped common belief* which enables it to happen. In Section 4, we give a modified version of the example where time-stamped common belief does not suffice for joint behavior, but *action-stamped common belief*, which is yet more general, does. In general, the group of agents involved in a joint behavior need not be static; it may change over time. For example, we would like to view the firefighters at the scene of a fire as acting jointly, but this group might change over time as additional firefighters arrive and some firefighters leave. In Section 5, we show how action-stamped (and time-stamped) common belief can be extended to deal with the group of agents changing over time. In Section 6, we go into more detail regarding the significance of these results. In Section 7, we show that there is a sense in which action-stamped common belief is necessary and sufficient for joint behavior. Finally, in Section 8, we conclude.

2 Background

To make our claims precise, we need to be able to talk formally about beliefs and time. To do so, we draw on standard ideas from modal logics and the runs-and-systems framework of Fagin et al. [9].

Our models have the form $M = (R, \Phi, \pi, \mathscr{B}_1, \ldots, \mathscr{B}_n)$. R is a *system*, which, by definition, is a set of *runs*, each of which describes a way the system might develop over time. Given a run $r \in R$ and a time $n \in \mathbb{N}_{\geq 0}$ (for simplicity, we assume that time ranges over the natural numbers), we call (r,n) a *point* in the model; that is, it describes a point in time in one way the system might develop. Φ is the set of variables. In general, we will denote variables in Φ with uppercase letters (e.g., P) and values of those variables with lowercase ones (e.g., p). π is an *interpretation* that maps each point in the model and each variable $P \in \Phi$ to a value, denoting the value of P at that point. (Thus, the analogue of a primitive proposition for us is a formula of the form $P = p$: variable P takes on value p.) Finally, for each agent i, there is a *binary relation* $\mathscr{B}_i$ over the points in the model. Two points (r_1, n_1) and (r_2, n_2) are related by $\mathscr{B}_i$ (i.e., $(r_1,n_1),(r_2,n_2)) \in \mathscr{B}_i$) if the two points are indistinguishable to agent i; that is, if, at the point (r_1,n_1), agent i cannot tell if the true point is (r_1,n_1) or (r_2,n_2). We assume throughout that the $\mathscr{B}_i$ relations satisfy the standard properties of a belief relation: specifically, they are *serial* (for all points (r,n), there exists a point (r',n') such that $((r,n),(r',n')) \in \mathscr{B}_i$), *Euclidean* (if $((r_1,n_1),(r_2,n_2))$ and $((r_1,n_1),(r_3,n_3))$ are in $\mathscr{B}_i$, then so is $((r_2,n_2),(r_3,n_3))$), and transitive. These assumptions ensure that the standard axioms for belief hold; see [9] for further discussion of these issues.

To talk about these models, we use the language generated by the following context-free grammar:

$$\varphi := P = p \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid B_i\psi \mid E_G\psi \mid C_G\psi,$$

where P is a variable in Φ, p is a possible value of P, and G is a non-empty subset of the agents. The intended reading of $B_i\psi$ is that agent i believes ψ; for $E_G\psi$ it is that ψ is believed by everyone in the group G; and for $C_G\psi$ it is that ψ is common belief among the group G.

We can inductively give semantics to formulas in this language relative to points in the above models. The propositional operators $\neg$ and $\wedge$ have the standard propositional semantics. The other operators are given semantics as follows:

- $(M,r,n) \models P = p$ if $\pi((r,n),P) = p$,
- $(M,r,n) \models B_i\psi$ if $(M,r',n') \models \psi$ for all points (r',n') such that $((r,n),(r',n')) \in \mathscr{B}_i$,
- $(M,r,n) \models E_G\psi$ if $(M,r,n) \models B_i\psi$ for all $i \in G$
- $(M,r,n) \models C_G\psi$ if $(M,r,n) \models E^k_G\psi$ for all $k \geq 1$, where $E^1_G\psi := E_G\psi$ and $E^{k+1}_G\psi := E_G(E^k_G\psi)$.

There are a number of axioms that are valid in these models. Since they are not relevant for the points we want to make here, we refer the reader to [9] for a discussion of them.

3 Time-Stamped Common Belief

We now give our first example showing that joint behavior does not require common belief. We do not define joint behavior here; indeed, as we said, there are a number of competing definitions in the literature [15, 4, 10, 11]. But we hope the reader will agree that, however we define it, the example gives an instance of it.

> General Y and her forces are standing on the top of a hill. Below them in the valley, the enemy is encamped. General Y knows that her forces are not strong enough to defeat the enemy on their own. She also knows that General Z and his troops, though knowing nothing of the encamped enemy, will arrive on the hill the next day at noon on the way back from a training exercise. Unfortunately though, General Y and her troops must move on before then. Thankfully, all generals are trained for how to deal with this situation. Just as her training recommends, General Y sets up traps that will delay the enemy's retreat, and leaves one soldier behind to inform General Z of the traps upon his arrival. At 11:30 the next morning, General Y receives a (false) message informing her that General Z and his troops have been captured, and thus (incorrectly) surmises that the enemy will live to fight another day. What in fact happens is that General Z's troops arrive at noon and attack the enemy, the enemy attempts to retreat and is stopped by General Y's traps, and the enemy is successfully defeated.

Clearly, Generals Y and Z jointly defeated the enemy. Yet they never achieved common belief of what they were doing. Before noon, General Z didn't even think that the enemy was there, and from 11:30 on, General Y thought that General Z would never arrive. It follows that there was no point at which they could have had common belief. So what is going on here?

What this example suggests is that there are times when a type of *time-stamped common belief* (cf., [9, 12]) suffices to enable joint behavior. Intuitively, on the first day, General Y believed that at noon on the second day General Z would act, attacking the enemy. Similarly, at noon on the second day, General Z believed that General Y had acted the day before, setting up the necessary traps. They also hold higher-order beliefs; for example, at the time she set the traps, general Y believed that at noon the next day general Z would believe that she had set the traps, otherwise she wouldn't have wasted the resources to set them, and so on. Much as in the usual case of common belief, these nested beliefs extend to arbitrary depths. What sets this example apart from those considered by earlier work is that, whereas in the earlier work agents needed to believe others would act as desired *at the same point*, here the agents need to believe only that others will act as desired *at the points where they're supposed to act for the joint behavior*. This suggests that time-stamped common belief can suffice for joint behavior.

We can capture this type of time-stamped common belief formally with the following additions to the logic and semantic models above. Syntactically, we add two more operators to the language, $E^t_G\psi$ and $C^t_G\psi$, where G is a set of agents. We then add to the semantic model a function t that maps each agent and run to a non-negative integer. The intended reading of these is "each agent $i \in G$ believes at the time $t(i,r)$ that ψ" and "it is time-stamped-by-t common belief among the agents in G that ψ", respectively. We give semantics to these operators as follows:

- $(M,r,n) \models E^t_G\psi$ if $(M,r,t(i,r)) \models B_i\psi$ for all $i \in G$
- $(M,r,n) \models C^t_G\psi$ if $(M,r,n) \models E^{t,k}_G\psi$ for all $k \geq 1$, where $E^{t,1}_G\psi := E^t_G\psi$ and $E^{t,k+1}_G\psi := E^t_G(E^{t,k}_G\psi)$.

These definitions are clearly very similar to the (standard) definitions given above for $E_G\psi$ and $C_G\psi$, except that the beliefs of each agent $i \in G$ in run r is considered at the time $t(i,r)$. It follows from the semantic definitions that $E^t_G\psi$ and $C^t_G\psi$ hold at either all points in a run or none of them.

In the example above, this notion of time-stamped common belief *is* achieved if we take $t(Y,r)$ to be the time in run r that Y laid the traps (which may be different times in different runs) and take $t(Z,r)$ to be the time that Z arrived in run r (which was noon in the actual run, but again, may be different times in different runs), provided that it is (time-stamped) common belief that both Y and Z will follow their training. That is, when Y lays the traps, Y must believe that Z will believe when he arrives that Y laid the traps, Z will believe when he arrives that Y believed when she laid the traps that he would believe when he arrived that Y laid the traps, and so on. The key point here is that time-stamped common belief can sometimes suffice for achieving cooperative behavior, even without standard common knowledge.

Our notion of time-stamped common belief is a generalization of (and was inspired by) Halpern and Moses' notion of (time-T) *time-stamped common knowledge*. Roughly speaking, for them, time-T time-stamped common knowledge of ϕ holds among the agents in a group G if every agent i in G knows ϕ at time T on her clock, all agents in G know at time T on their clock that all agents in G know ϕ at time T on their clock, and so on (where T is a fixed, specific time). If it is common knowledge that clocks are synchronized, then time-stamped common knowledge reduces to common knowledge. If we take $t(i,r)$ to be the time in run r that i's clock reads time T (and assume that it is commonly believed that each agent's clock reads time T at some point in every run), then their notion of time-stamped common knowledge becomes a special case of our time-stamped common belief. But note that with time-stamped common belief, we have the flexibility of referring to different times for different agents, and the time does not have to be a clock reading; it can be, for example, the time that an event like laying traps occurs.

4 Action-Stamped Common Belief

There is an even more general variant of common belief that can suffice for joint behavior. What really mattered in the previous example is that everyone had the requisite beliefs at the times that they were acting. But there need not necessarily be only one such point per agent per run; an agent might act multiple times as part of the plan, as the following modified version of the story illustrates:

> General Y and her forces arrive to the south of the town where the enemy forces are encamped. General Y knows that her forces are not strong enough to defeat the enemy on their own. She also knows that General Z and his troops are expected to arrive to the north of the city some time in the near future, though she and her troops must move on

> before then. The swiftly-coursing river prevents the enemy from escaping to the east. But unfortunately, they can still escape inland to the west. Thankfully, all generals are trained for how to deal with this situation as well. Just as her training recommends, General Y sets up traps that will delay the enemy's southward retreat and then, as she heads inland, also sets up traps to the west, finally leaving one soldier behind to go north and inform General Z of the traps upon his arrival. The next morning, General Y receives a (false) message informing her that General Z and his troops have been captured, and thus (incorrectly) surmises that the enemy will live to fight another day. What in fact happens is that General Z's troops arrive later that day and are informed by the remaining soldier that, not too long ago, General Y's troops set traps to the south and west. They attack the enemy, the enemy attempts to retreat and is stopped by General Y's traps, and the enemy is successfully defeated.

Again, Generals Y and Z jointly and collaboratively defeated the enemy, but time-stamped common belief doesn't suffice for this version of the story, because we cannot specify a single time for General Y's actions. Instead, what really matters is that when they are acting as part of a joint plan they hold the requisite (common) beliefs. The joint plan need not be known upfront; General Z does not know what he will need to do to achieve the common goal until he arrives at the scene. To capture this new requirement, we define a notion of *action-stamped common belief*.

We begin by adding a special Boolean variable $ACTING_{i,G}$ for any group G and agent $i \in G$. This variable is true (i.e., takes value 1, as opposed to 0) at a point (r,n) if the agent i is acting towards the group plan of G at (r,n) and false otherwise. So for the generals, $ACTING_{Y,G} = 1$ would be true when she lays the traps, $ACTING_{Z,G} = 1$ would be true at the point when he attacks, and they'd both be false otherwise (where $G = \{Y,Z\}$). We often write $ACTING_{i,G}$ and $\neg ACTING_{i,G}$ instead of $ACTING_{i,G} = 1$ and $ACTING_{i,G} = 0$, and similarly for other Boolean variables. By using $ACTING_{i,G}$, we can abstract away from what actions are performed; we just care that some action is performed by agent i towards the group plan, without worrying about what that action is.

As in the case of time-stamped common belief, we add two modal operators to the language (in addition to the variables $ACTING_{i,G}$). Let G be a set of agents. $E^{\mathbf{a}}_G\psi$ then expresses that, for each agent $i \in G$, whenever $ACTING_{i,G}$ holds (it may hold several times in a run, or never), i believes ψ. $C^{\mathbf{a}}_G\psi$ then defines the corresponding notion of common belief for the points at which agents act at part of the group.

We give semantics to these modal operators as follows:

- $(M,r,n) \models E^{\mathbf{a}}_G\psi$ if for all n' and all $i \in G$ such that $(M,r,n') \models ACTING_{i,G} = 1$, it is also the case that $(M,r,n') \models B_i\psi$.
- $(M,r,n) \models C^{\mathbf{a}}_G\psi$ if $(M,r,n) \models E^{\mathbf{a},k}_G\psi$ for all $k \geq 1$, where $E^{\mathbf{a},1}_G\psi := E^{\mathbf{a}}_G\psi$ and $E^{\mathbf{a},k+1}_G\psi := E^{\mathbf{a}}_G(E^{\mathbf{a},k}_G\psi)$.

Returning to the example, although the agents do not have time-stamped common belief at all the points when they act, they do have action-stamped common belief. General Z acted believing that General Y had acted as expected, and also believing that General Y acted believing that he would act as expected, and so on.

It is easy to see that time-stamped common belief can be viewed as a special case of action-stamped common belief: Given a time-stamping function t, we simply take $ACTING_{i,G}$ to be true at those points (r,n) such that $t(i,r)$ holds.

It is worth noting that, in both this and the previous section, the agents having a protocol in advance for how to deal with the situation is not really necessary for them to succeed. In the examples, consider a scenario where generals are in fact not trained for how to handle the situation, but instead General Y has the brilliant idea to lay traps and send a messenger to meet General Z upon arrival. As long as message delivery is reliable, action-stamped common belief can be achieved and they can successfully defeat the enemy.

5 Joint Behaviors Among Changing Groups

In practice, the members of groups change over time. For example, a group of firefighters may work together to safely clear a burning building, but (thankfully!) they don't need to wait until all the firefighters are on the scene, or even until it is known which firefighters are coming, in order for the first firefighters to begin. Instead, structures and guidelines allow the set of firefighters who are on the scene to act cooperatively, even without each firefighter knowing who else will show up.

The formalisms of the two previous sections assumed a fixed group G, so cannot capture this kind of scenario. But the changes necessary to do so are not complicated. Rather than considering (some variant of) common belief with respect to a fixed set G of agents, we consider it with respect to an *indexical* set S, one whose interpretation depends on the point. More precisely, an indexical set S is a function from points to sets of agents; intuitively, $S(r,n)$ denotes the members of the indexical group S at the point (r,n). We assume that a model is extended so as to provide the interpretation of S as a function.

Our semantics for action-stamped common belief with indexical sets are now a straightforward generalization of the semantics for rigid (non-indexical) sets:

- $(M,r,n) \models E_S^{\mathbf{a}}\psi$ if for all n' and all $i \in S(r,n')$ such that $(M,r,n') \models ACTING_{i,S}$, it is also the case that $(M,r,n') \models B_i\psi$.
- $(M,r,n) \models C_S^{\mathbf{a}}\psi$ if $(M,r,n) \models E_S^{\mathbf{a},k}\psi$ for all $k \geq 1$, where $E_S^{\mathbf{a},1}\psi := E_S^{\mathbf{a}}\psi$ and $E_S^{\mathbf{a},k+1}\psi := E_G^{\mathbf{a}}(E_S^{\mathbf{a},k}\psi)$.

The only change here is that in the semantics of $E_S^{\mathbf{a}}$, we need to check the agents in $S(r,n)$ for each point.

Of course, we can also allow indexical sets in time-stamped common belief in essentially the same way. Whereas in the semantics of $E_G^t\psi$, we required that $(M,r,n) \models E_G^t\psi$ if, for all $i \in G$, $(M,r,t(i,r)) \models B_i\psi$, now we require that $(M,r,n) \models E_S^t\psi$ if, for all agents i, if $i \in S(r,t(i,r))$, then $(M,r,t(i,r)) \models B_i\psi$. We care about what agent i believes at $(M,r,t(i,r))$ only if i is actually in group S at the point $(r,t(i,r))$.

6 Significance

In Sections 3-5 we showed that action-stamped common belief can suffice to enable joint behavior, whereas the prior work on the topic had assumed common belief was necessary. Why does this matter? We argue that it is important for two reasons: 1) misunderstanding the type of belief necessary can lead to mis-evaluation of cooperative capabilities, and 2) requiring common belief can unnecessarily make cooperation impossible in scenarios where it is in fact possible and could be quite beneficial.

As part of the recent push for more research on cooperative AI, some have argued that we should "construct more unified theory and vocabulary related to problems of cooperation" [7]. One important step in this program is (in our opinion) formalizing the requirements for various types of cooperation, including joint behavior. This, in turn, requires understanding the level and type of (common) belief needed for joint behavior. As our examples have shown, full-blown common belief is not necessary; weaker variants that are often easier to achieve can suffice. Relatedly, there has been a push to develop methods for *evaluating* the cooperative capabilities of agents, as a way of developing targets and guideposts for the community [5]. Again, this will require understanding (among other things) what type of beliefs are necessary for cooperation. Incorrect assumptions about the types of beliefs necessary can lead to incorrect conclusions about the feasibility of cooperation. For example, if an evaluation system takes as given the assumption that it is impossible for agents that cannot achieve common belief to behave cooperatively, it may in fact lead to effective cooperative agents being scored badly, leading to misdirected research.

A second reason that it is important to clarify the types of beliefs necessary for joint behavior is that misunderstanding them can lead to systems unnecessarily aborting important cooperative tasks. As is well known, achieving true common knowledge can be remarkably difficult in real-world systems, often requiring either a communication system that guarantees truly synchronous delivery or guaranteed bounded delivery time together with truly synchronized clocks [9]. Action-stamped common belief can sometimes be achieved when common belief cannot. To demonstrate the importance of this, we consider an example from the domain of urban search and rescue, a domain where 1) the use of multi-agent systems consisting of humans and AI agents has long been considered and advocated for, 2) the types of teamwork necessary can be complex, and 3) there is some evidence of potential adoption, having been used, for example, at a small scale in the aftermath of September 11th [3, 14, 13, 16, 17, 20]. Though the example we give is a simple, stylized case, the domain is sufficiently complex that we would expect these types of issues to arise in practice if systems were deployed at scale.

Example 1. *An earthquake occurs, causing a large building to collapse. The nearest search and rescue team arrives on scene, and the incident commander has to decide how to proceed. The team has determined that the structure is stable and will not collapse, and so is safe to enter. However, attempting to exit the building may disrupt the structure and cause harm. The incident commander determines that there are two reasonable options:*

1. *Wait a week for a heavy piece of machinery that will certainly be able to safely lift the roof of the collapsed building on its own and allow rescuers safe access to the building.*
2. *A team can enter the building and restabilize parts of the roof. The restabilization would not be enough to make it safe to exit—in fact, it would require adjusting the structure in ways that would make an attempt to exit even more risky—but it would allow a more easily accessible robotic system to safely remove the roof piece by piece, allowing the rescuers and anyone trapped inside to safely escape.*

The incident commander decides it is best not to wait, and so takes the second, joint-behavior-based approach. He sends the team of rescuers in to begin the necessary process, and tells them the full plan and that he expects it will be 2-3 hours before the robot arrives on scene. The group enters the wreckage and secures it in the necessary ways, as planned. But it turns out that the earthquake affected many buildings, so the robot is in high demand. It ends up taking close to 8 hours for the robot to arrive on scene. When the robot arrives, the incident commander enters the relevant information in the robot's system—namely, the full plan and that the restabilization has been carried out—so the robot can carry out its part of the specified cooperative plan.

If the robot's model of joint behavior requires common belief, a problem will arise. At no point is there ever common belief of the joint behavior. Before the robot arrives, the robot certainly has no belief about the joint behavior. And when the robot arrives, it must consider the possibility that, because of the delay, the rescuers have given up hope of the robot arriving and concluded that they may have to wait a full week until the larger piece of machinery is available. Even if this isn't actually the case, because the robot considers it possible, common belief will not be achieved. So if the robot assumes that joint behavior requires common belief, it will determine that the joint behavior cannot be carried out. Thus, everyone will have to wait a week for the heavier machinery, risking the lives of anyone trapped inside.

If, on the other hand, the robot's theory of joint behavior is based on action-stamped common belief, the task will be properly and safely carried out as soon as the robot arrives: When the rescuers perform *their part, they believe that the robot will arrive soon and perform its part of the task. Similarly, when the robot arrives and the incident commander enters the relevant information, the robot believes that the rescuers held those beliefs when acting (and therefore performed the required adjustments). The rescuers believed that the robot would hold these beliefs when it arrived, the robot believed they would, and so on. The fact that the robot arrived later than expected and the rescuers may have started to have uncertainty about the plan doesn't affect the requisite beliefs because all that matters are the beliefs of the agents at the points where they act.*

This example highlights the value of getting the types of beliefs necessary right; getting the theory right, and basing it on action-stamped common belief instead of standard common-belief, can enable cooperation in a range of important scenarios where standard common belief is impossible or difficult to achieve, whereas action-stamped common belief may be easily attainable.

7 On the Necessity and Sufficiency of Action-Stamped Common Belief for Joint Behavior

We've argued in this paper that the prior work was incorrect in asserting that common belief was necessary for joint behavior, and shown by example that action-stamped common belief can suffice. We now argue that an even stronger statement is true: there is a sense in which action-stamped common belief is necessary and sufficient for joint behavior. We say "in a sense" here, because much depends on the conception of joint behavior being considered. So what we do in this section is give a property that we would argue is one that we would want to hold of joint behavior, and then show that action-stamped common belief is necessary and sufficient for this property to hold.

What does it take to go from a collection of individual behaviors to a joint behavior? The following example may help illuminate some of the relevant issues.

> Jasper and Horace are both crooks, though neither is an evil genius by any stretch of the imagination. They both independently decide to rob the Great Bank of London on exactly the same day. As it turns out, neither of them did a good job preparing, and they each knew about only half of the bank's security systems, and so made plans to bypass only that half. By sheer dumb luck, between them they know about all the bank's security systems. So when each bypasses the part that they know about (at roughly the same time), the bank's security systems go down. They each make it in, steal a small fortune, and escape, none the wiser as to the other's behavior or that their plan was doomed to fail on its own.

Is Jasper and Horace robbing the bank an instance of joint behavior? We think not. One critical component that distinguishes this from a joint behavior is the beliefs of the agents. Joint behaviors are collective actions where people do their part because they believe that everyone else will do their part as well. Here, Jasper and Horace have no inkling that the other will help disable the system.

We now want to capture these intuitions more formally. We start by adding another special Boolean variable $SHOULD\_ACT_{i,S}$ for each agent i and indexical group S, specifying the points in each run where agent i is supposed to act towards the plan of group S. We then add a special formula χ_S to the language:[1]

- $(M,r,n) \models \chi_S$ if for all n' and all agents $i \in S(r,n')$, $(M,r,n') \models SHOULD\_ACT_{i,S} \rightarrow ACTING_{i,S}$

[1] As long as the set of agents is finite (which we implicitly assume it is), we can express χ_S in a language that includes a standard modal operator $\boxdot$, where $\boxdot\varphi$ is true at a point (r,n) iff φ is true at all points (r,n') in the run. For ease of exposition, we do not introduce the richer modal logic here.

The formula χ_S is thus true at a point (r,n) if, at all points in run r, each agent i in the indexical group S plays its part in the group plan whenever it is supposed to.

If we think of $ACTING_{i,S}$ as "i is taking part in the joint behavior of the group S", then the property JB_S (for "Joint Behavior, group S") that we now specify essentially says that to have truly joint behavior, each agent in S must believe when she acts that all of the members of the (indexical) group S will do what they're supposed to; if they don't all have that belief, then it's not really joint behavior. Formally, JB_S is a property of an indexical group S in a model M:

[JB_S:] For all points (r,n) and agents $i \in S(r,n)$, $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\chi_S$.

Requiring JB_S for joint behavior makes action-stamped common belief of χ_S necessary for joint behavior.

Theorem 7.1. *If JB_S holds in a model M, then $(M,r,n) \models C^a_S\chi_S$ for all points (r,n).*

Proof. We begin by defining a notion of *a-reachability*: A point (r',n') is S-a-reachable from (r,n) in k steps if there exists a sequence $(r_0,n_0),\ldots,(r_k,n_k)$ of points such that $(r_0,n_0) = (r,n)$, $(r_k,n_k) = (r',n')$, and for all $0 \leq l < k$, there exists a point (r_l,n'_l) and an agent $i \in S(r_l,n'_l)$ such that $(M,r_l,n'_l) \models ACTING_{i,S}$ and $((r_l,n'_l),(r_{l+1},n_{l+1})) \in \mathscr{B}_i$.

By the semantics of C^a_S, $C^a_S\chi_S$ holds at (r,n) iff χ_S holds at every point (r',n') that is S-a-reachable from (r,n) in 1 or more steps. Consider any such point (r',n'). Then, by the definition of reachability, there exists some point (r'',n'') and some agent $i \in S(r'',n'')$ such that $(M,r'',n'') \models ACTING_{i,S}$ and $((r'',n''),(r',n')) \in \mathscr{B}_i$. Because $(M,r'',n'') \models ACTING_{i,S}$, we get by JB_S that $(M,r'',n'') \models B_i\chi_S$. Then by the semantics of B_i and the fact that $((r'',n''),(r',n')) \in \mathscr{B}_i$ we get that $(M,r',n') \models \chi_S$. But (r',n') was an arbitrary point S-a-reachable from (r,n) in 1 or more steps, so χ_S holds at all such points, and we have that $(M,r,n) \models C^a_S\chi_S$. But (r,n) was also arbitrary, so $C^a_S\chi_S$ holds at all points. □

The converse to Theorem 7.1 also holds; that is, action-stamped common belief of χ_S suffices for JB_S to hold. Put another way, action-stamped common belief is exactly the ingredient that we need to meet the belief requirements of the property that we used to characterize joint behavior.

Theorem 7.2. *If $(M,r,n) \models C^a_S\chi_S$ for all points (r,n), then JB_S holds in M.*

Proof. Consider an arbitrary point (r,n) and agent $i \in S(r,n)$ such that $(M,r,n) \models ACTING_{i,S}$. By assumption, $(M,r,n) \models C^a_S\chi_S$. So, by the semantics of C^a_S, it follows that $(M,r,n) \models E^a_S\chi_S$. In turn, it follows from the semantics of E^a_S that $(M,r,n) \models B_i\chi_S$ (because $(M,r,n) \models ACTING_{i,S}$). But r, n, and i were arbitrary, so we have that $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\chi_S$ for all such points and agents. Thus, JB_S holds in M. □

The astute reader will have noticed that the proofs of Theorem 7.1 and 7.2 did not depend in any way on χ_S. The formula χ_S in these theorems can be replaced by an arbitrary formula φ. In other words, if all the agents in S believe φ at the point when they act, then φ is action-stamped common belief, and if φ is action-stamped common belief, then all agents in S must believe φ at the point when they act. Formally, the proofs of Theorem 7.1 and 7.2 also show the following:

Theorem 7.3. *If $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\varphi$ for all points (r,n) and agents $i \in S(r,n)$, then $(M,r,n) \models C^a_S\varphi$ for all points (r,n).*

Theorem 7.4. *If $(M,r,n) \models C^a_S\varphi$ for all points (r,n), then $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\varphi$ for all points (r,n) and agents $i \in S(r,n)$.*

8 Conclusion and Future Work

We have argued here that, contrary to what was suggested in earlier work, common belief is not necessary for joint behavior. We have presented a new notion, *action-stamped* common belief, and shown that it is, in a sense, necessary and sufficient for joint behavior, and can be achieved in scenarios where standard common belief cannot. This is important because modelling the conditions needed for joint behavior correctly can enable cooperation in important scenarios, such as search and rescue, where it might not otherwise be possible. We chose to use the term *joint behavior* in this paper because it sounded to our ears like it most accurately captured the notion we were considering; no doubt to some readers other terms will sound like a better fit. As we showed in Section 7, action-stamped common belief characterizes scenarios where individuals do their part only if they believe others will do the same, whatever terminology we use.

We suspect that, for some readers, the idea that action-stamped common belief is sufficient for joint behavior will seem obvious. In a certain sense, we agree; in retrospect, it *does* feel like the obviously correct notion for joint behaviors. That said, while action-stamped common belief seems quite natural, it does not seem to have been studied in any prior literature.

With that in mind, it is worth briefly discussing the connection between the ideas in this paper and some of the prior work that has been done. First, note that action-stamped common belief is in some ways the natural variant of common belief for extensive-form games. Because an agent i's information sets are usually specified only at nodes at which agent i moves, it is possible to reason about agent i's beliefs only at points where agent i acts. This makes it all the more surprising that action-stamped common belief has not been formalized and studied in its own right; in some sense, it captures what epistemic game theorists have been implicitly considering in the case of extensive-form games.

In this paper, we have considered the types of *beliefs* necessary for joint behavior, but that may not be the only factor involved (nor do we claim it is; we are just focused in this paper on the belief component). For example, in much of the literature, *intent* is taken to play an important role in various cooperative behaviors. Dunin-Keplicz and Verbrugge [8] proposed a three-part notion of "collective commitment", with the levels of belief (e.g., no one believes, everyone believes, it is common belief) held at each of the three parts leading to various types of collective commitment. Their work is in some ways orthogonal to ours; it can be thought of as considering various types of cooperative behaviors that can occur, while ours just focuses on one particularly strict form, joint behavior. One way of interpreting our work in the context of theirs is as saying that the top level of belief to consider for cooperation should in fact be action-stamped common belief.

Blomberg [1] gives an insightful argument that common belief (and variants thereof) of intentions is not necessary for a joint intentional act. Roughly speaking, an agent may (incorrectly) believe that other agents do not share his intent, as long as what they intend would still lead them to act in the manner conducive to his goals. We find his counterexample and arguments compelling. But this is perfectly consistent with our results. Theorems 7.1 and 7.2 show that action-stamped common belief (or in the case of simultaneous acts, standard common belief) that agents will do the necessary acts is required for joint behaviors. We place no requirements on what agents have to believe about other agents' intentions. Put a different way, Blomberg makes a compelling case that, when characterizing joint behavior, it is a mistake to instantiate the φ in Theorems 7.3 and 7.4 with formulas about shared intents. That is to say, it is not a necessary property of cooperative behavior that agents act only if they are sure others are acting with the same purpose.

Lastly, Roy and Schwenkenbecher [19] consider a novel notion of belief that they call "pooled knowledge", which is related to distributed knowledge, and argue that it is both weaker than common knowledge and sufficient for shared intentions. The basic idea behind the argument is that if agents are rational, then pooled knowledge would induce some agent to act as a coordinator to guide the behavior of the group. It's certainly an interesting proposal, and one that deserves further study. From the perspective of this paper, it would be interesting to try to formally analyze under what conditions pooled knowledge/belief would lead to action-stamped common knowledge/belief.

The present work suggests two areas that are ripe for future work. The first is to more fully explore the logical aspects of action-stamped common belief. Can a sound and complete axiomatization be provided? What is the complexity of the model-checking and validity problems for a language involving action-stamped common knowledge? How can we practically engineer systems that rely on action-stamped common belief? The second area we think worth exploring is that of understanding better what levels of group knowledge are required for other aspects of joint behavior and other types of cooperation. We focused on one aspect, revealing a nuanced but important error in earlier thinking. We think that there may well be other aspects of cooperation that are worth digging into in this fine-grained way. Given the importance of cooperative AI, we hope that others will join us in exploring these questions.

References

[1] O. Blomberg (2016): *Common knowledge and reductionism about shared agency*. *Austalasian Journal of Philosophy* 94(2), pp. 315–326.

[2] M. E. Bratman (1992): *Shared cooperative activity*. *The Philosophical Review* 101(2), pp. 327–341.

[3] J. Casper & R. R. Murphy (2003): *Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center*. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 33(3), pp. 367–385.

[4] P. R. Cohen & H. J. Levesque (1991): *Teamwork*. *Nous* 25(4), pp. 487–512.

[5] Cooperative AI Foundation (CAIF) (2022): *Evaluation for Cooperative AI: Call for Proposals*. https://www.cooperativeai.com/calls-for-proposals/evaluation-for-cooperative-ai; accessed October. 25, 2022.

[6] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson & T. Graepel (2021): *Cooperative AI: machines must learn to find common ground*.

[7] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson & T. Graepel (2020): *Open problems in cooperative AI*. Available at https://arxiv.org/pdf/2012.08630.pdf.

[8] B. Dunin-Kęplicz & R. Verbrugge (2004): *A tuning machine for cooperative problem solving*. *Fundamenta Informaticae* 63(2–3), pp. 283–307.

[9] R. Fagin, J. Y. Halpern, Y. Moses & M. Y. Vardi (1995): *Reasoning About Knowledge*. MIT Press, Cambridge, MA. A slightly revised paperback version was published in 2003.

[10] B. J. Grosz & S. Kraus (1996): *Collaborative plans for complex group action*. *Artificial Intelligence* 86(2), pp. 269–357.

[11] B. J. Grosz & C. L. Sidner (1990): *Plans for discourse*. In P. R. Cohen, J. L. Morgan & M. E. Pollack, editors: *Intentions in Communication*, chapter 20, MIT Press.

[12] J. Y. Halpern & Y. Moses (1990): *Knowledge and common knowledge in a distributed environment*. *Journal of the ACM* 37(3), pp. 549–587.

[13] H. Kitano & S. Tadokoro (2001): *Robocup rescue: A grand challenge for multiagent and intelligent systems*. *AI magazine* 22(1), pp. 39–39.

[14] H. Kitano, S. Tadokoro, I. Noda, H. Matsubara, T. Takahashi, A. Shinjou & S. Shimada (1999): *Robocup rescue: Search and rescue in large-scale disasters as a domain for autonomous agents research*. In: *1999 IEEE International Conference on Systems, Man, and Cybernetics*, 6, pp. 739–743.

[15] H. J. Levesque, P. R. Cohen & J. H. T. Nunes (1990): *On acting together*. In: *Proc. Eighth National Conference on Artificial Intelligence (AAAI '90)*, pp. 94–99.

[16] Y. Liu & G. Nejat (2013): *Robotic urban search and rescue: A survey from the control perspective*. *Journal of Intelligent & Robotic Systems* 72(2), pp. 147–165.

[17] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju & T. Westerlund (2020): *Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision*. *IEEE Access* 8, pp. 191617–191643.

[18] C. Rich & C. L. Sidner (1997): *COLLAGEN: When agents collaborate with people*. In: *Proceedings of the First International Conference on Autonomous Agents*, pp. 284–291.

[19] O. Roy & A. Schwenkenbecher (2021): *Shared intentions, loose groups, and pooled knowledge*. *Synthese* 198(5), pp. 4523–4541.

[20] J. Scholtz, J. Young, J. L. Drury & H. A. Yanco (2004): *Evaluation of human-robot interaction awareness in search and rescue*. In: *Proc. IEEE International Conference on Robotics and Automation, 2004 (ICRA'04)*, 3, pp. 2327–2332.

[21] M. Tambe (1997): *Towards flexible teamwork*. *Journal of A.I. Research* 7, pp. 83–124.

[22] R. Tuomela (2005): *We-intentions revisited*. *Philosophical Studies* 125(3), pp. 327–369.