# Relating the Entropy of Joint Beliefs to Multi-Agent Coordination

Mikhail Prokopenko, Peter Wang

Intelligent Interactive Technology
CSIRO Mathematical and Information Sciences
Locked Bag 17, North Ryde, NSW 1670, Australia
{mikhail.prokopenko, peter.wang}@csiro.au

**Abstract.** Current approaches to activity coordination in multi-agent systems (teams) range from strictly top down (plan-based coordination) to purely emergent (reactive coordination), with many hybrid variants, each having its specific advantages and disadvantages. It appears to be extremely difficult to rigorously compare various hybrid approaches to multi-agent coordination (and communication), given the lack of a generic semantics or some guidelines. In this paper, we studied some intuitive inter-agent communication policies and characterised them in terms of generic information-theoretic properties. In particular, the relative entropy of joint beliefs was suggested as an indicator of teams coordination potential. Our novel behaviour-based agent architecture (based on the Deep Behaviour Projection framework) enabled consistent reasoning about belief change, including beliefs about other agents. This allowed us to examine some of the identified communication policies empirically. The obtained results confirmed that there are certain interesting invariants — in particular, a change in team coordination (and overall performance) was shown to be within the boundaries indicated by the relative information entropy.

## 1 On entropy and multi-agent agreements

The primary objective of this work is a formal characterisation of certain classes of multi-agent agreements. In achieving this goal, we tried to make as few assumptions as possible about the choice of inter-agent communication variables and periods of team synchronisation (extensively analysed by Stone and Veloso [8]). In particular, we studied *selfish* agreements covering "selfish" agents that communicate data about themselves only, *transitively-selfish* agreements ensuring that each "cooperative" agent always communicates the data about some other agent, and *mixed* agreements, where a team composition parameter determines the precise split between selfish and cooperative agents.

In order to capture the agreements in a formal information-theoretic setting we analysed the joint "output" of inter-agent communication after each period of team synchronisation. Then we estimated the *relative entropy* as a precise measure of the amount of freedom of choice (the degree of randomness) [7] contained in the resultant joint beliefs. Our intention was to use the relative entropy of joint beliefs in multi-agent teams as a generic indicator of the team coordination potential. Clearly, the team following an agreement with near-zero entropy (almost no "misunderstanding" in joint beliefs) has a higher coordination potential than the team adherent to an agreement with near-maximal entropy (joint beliefs are almost random).

We start our analysis with a simple protocol $\mathcal{P}_1$ that allows an agent to communicate data about only one agent precisely. In other words, each agent is able to encode either the data about itself or about the other agent. Without loss of generality, we may assume that the protocol $\mathcal{P}_1$ has enough symbols to encode $n$ distinguishable objects and a single-object capacity for each communication message. We introduce a binary relation $S(a_i, a_j)$ to denote that the agent $a_i$ *sends* a message containing the object $a_j$. Let $S^*$ denote the transitive closure of the relation $S$. Arguably, on of the most intuitive agreements is an agreement among selfish agents — since the data about themselves is, arguably, more readily available, the selfish agents choose this data as their content. In fact, we may assume for our analysis that each agent is always "self-aware". Formally, $K(a_i, a_i) = \mathit{true}$ for a Boolean (belief-)function $K$ defined for each agent pair. Generally, we propose the following definition.

**Definition 1.** *A locker-room agreement is called* selfish *if and only if* $S(a_i, a_i)$ *for all agents* $a_i$*,* $1 \leq i \leq n$*.*

*A locker-room agreement is called* transitively-selfish *if and only if* $S^*(a_i, a_i)$ *for all agents* $a_i$*,* $1 \leq i \leq n$*.*

*A non transitively-selfish agreement is called* mixed*.*

One might argue that the transitively-selfish agreement is an agreement among more "cooperative" agents choosing to communicate the data about the other agent (when available). Notice, however, that (given a successful team synchronisation) everyone is in the "loop". Of course, by definition, a selfish locker-room agreement is always transitively-selfish. In a mixed agreement, there are $(\alpha n)$ agents such that $S(a_i, a_i)$, and $(1 - \alpha)n$ agents such that $S(a_i, a_j)$ where $i \neq j$. Basically, the value of $\alpha$ determines the team composition (and we sometimes refer to $\alpha$ as the team composition parameter).

In order to formally capture the distinction among selfish, transitively-selfish and mixed agreements, we consider the joint "output" of inter-agent communication at the end of each period of team synchronisation. More precisely, we analyse joint beliefs represented by the sequence of individual beliefs $K_t = K(a_1, a_1), \ldots, K(a_i, a_j), \ldots, K(a_n, a_n)$, where $1 \leq i \leq n$ and $1 \leq j \leq n$, at the time $t$. In other words, rather than compute the amount of information contained in each message we attempt to estimate how much information is contained in the whole *team* after a period of team synchronisation.

In the simplest cases, the amount of information can be measured by the logarithm (to the base 2) of the number of available choices. The *entropy* is a precise measure of the amount of freedom of choice (or of the degree of randomness) contained in the object — an object with many possible states has high entropy. Formally, the entropy of a probability distribution $P = \{p_1; p_2; \ldots; p_m\}$ is defined by

$$H(P) = \sum_{i=1}^{m} p_i * \log{(1/p_i)}.$$

Having calculated the entropy $H(P)$ of a certain information source (such as a joint result of inter-agent communication) with the probability distribution $P$, one can compare this to the maximum value $H_{max}$ this entropy could have, assuming that the source employs the same symbols. The ratio of the actual to the maximum entropy is called the *relative entropy* of the source [7]. Therefore, if we calculate the relative entropy $H_r$ of $K_{t+p}$ we can characterise the multi-agent agreement employed between $t$ and $t + p$. The following representation results were obtained[1].

**Theorem 1.** *Selfish agreements attain minimal entropy.*

*Transitively-selfish agreements without the selfish agents attain maximal entropy asymptotically when the number of agents* $n \to \infty$.

*The trajectory of the relative entropy in multi-agent teams* ($n > 2$) *following mixed agreements does not have a fixed-point as a function* $H_r(\alpha)$ *of the team composition parameter:* $H_r(\alpha) \neq \alpha$.

This theorem basically states that whenever team agents agree to communicate the data about themselves only, they eventually leave nothing to choice. In other words, they always maximise their joint beliefs upon successful synchronisations. The obvious drawback is that while using single channels this saturation of joint beliefs requires that every agent takes turns in communication according to some schedule, and hence, large teams may take a while to minimise the entropy. The clear benefit, on the other hand, is that this minimisation is shown to be theoretically possible.

On the other hand, the "organisation" or "order" brought about by the transitively-selfish agreements is not sufficient to combat the entropy. Intuitively, the pair-wise "ignorance" of agents grows faster than the transitively-selfish agreement can cope with. Clearly, with the number of agents approaching infinity (and the entropy reaching its maximum asymptotically) the time to synchronise the team becomes infinite as well.

Obviously, the entropy of joint beliefs in multi-agent systems following mixed agreements exhibits some properties of both selfish and transitively-selfish configurations. We might expect that the selfish agents will bring in some order (as the compensation for potentially redundant information about themselves), while the cooperative (transitively-selfish) agents will lead to a higher degree of randomness (providing sometimes potentially non-trivial information about other agents). Formally, the relative entropy produced by mixed agreements asymptotically approaches 1 with growth in the number of agents. In other words, the selfish agents "loose" the battle for order (asymptotically) when the number of agents is infinitely large. Interestingly, however, the lower limit is not

[1] The proofs are omitted due to the lack of space.

zero, meaning that absolute order is never achievable regardless of the team split or the number of agents. In fact, our results showed that the joint beliefs obtainable in multi-agent teams with mixed agreements exhibit information-theoretic *complexity* in terms of the team composition. It has been recently pointed out in the literature (eg., by Suzudo [9]) that the entropy trajectory is a useful descriptor for a variety of self-organised patterns: eg., non-complex cellular automata (CA) have a fixed-point entropy trajectory and converge quickly to either very low or very high values. It should be noted that Suzudo considered the entropy of CA associated with the temporal pattern, while our analysis is focused on entropy of joint beliefs associated with the team composition parameter.

Our analysis was carried out for the protocol $\mathcal{P}_1$. However, it can be easily shown that protocols with higher capacities can be analysed in already presented terms. For example, consider the protocol $\mathcal{P}_2$ allowing an agent to communicate data about precisely two agents (including the data about itself). In other words, in the case of $n$ agents the protocol $\mathcal{P}_2$ has enough symbols to encode $n$ agents and two-objects capacity for each communication message. It is, nevertheless, possible to consider every message $S(a_i, a_j + a_k)$, where $+$ denotes the concatenation of the symbols corresponding to two objects, as two consecutive separate messages $S(a_i, a_j)$ and $S(a_i, a_k)$. This decomposition can be applied if $i = j$ or $i = k$ as well. Therefore, in order to analyse resultant joint beliefs one can double the synchronisation period in length and consider as a result the union of two sets of joint beliefs — the first set obtained after all messages with the first object are communicated, and the second set obtained after all messages with the second object are communicated. In other words, the decomposition allows to reduce the analysis of the protocol $\mathcal{P}_2$ (or any $k$-object capacity protocol $\mathcal{P}_k$) to that of the protocol $\mathcal{P}_1$ — simply because each divided message conforms to $\mathcal{P}_1$. That is, the resultant joint beliefs will be a combination of beliefs obtained by some selfish, transitively-selfish or mixed agreement in $\mathcal{P}_1$.

Another interesting reduction can be obtained in cases when agents intend to communicate the data about other objects in the environments (eg., the ball vectors in the RoboCup environment). In this case we just consider the ball to be a silent agent in the $(n+1)$-agent team. More precisely, denoting by $b$ the ball object, the messages $S(a_i, a_j + \ldots + b)$ would be possible while $S(b, a_j + \ldots + b)$ would be ruled out, again reducing the consideration to the protocol $\mathcal{P}_1$.

In summary, the advantage of higher-capacity protocols is in the shorter periods of required synchronisation but not in some exceptional information-theoretic properties of resultant joint beliefs.

## 2 Agents situated in time and relativity of behaviours

The strength of the presented analysis, we believe, is in its generic nature. The results lay down some general guidelines in terms of team composition and suggest definite boundaries on the team coordination potential.

In this section we focus on the agents ability to dynamically change their beliefs under different scenaria. We assumed previously that (during any synchronisation period) joint and individual beliefs can only expand, while obviously some of them should be discarded with time and some should be reconciled with new observations. At this stage, we shall describe some design and implementation details required to verify maximal and minimal limits of the entropy contained in the agents' dynamic beliefs.

In general, the agent's capability to maintain dynamic beliefs is based on another very important cognitive skill — the ability to remove itself from the current context. This ability is sometimes informally referred to as "possession of a reality simulator" [2]. Running a reality simulator or "imagining" allows the agent to reflect on past behaviour and project the outcome of future behaviour. For example, Joordens [2] makes a conjecture that higher mental states emerge as a result of a reality simulator: "an animal with no reality simulator basically lives in the present tense, and sees the world through only its eyes, at all times", while "the possession of a reality simulator may also allow an organism to experience many of the high-level cognitive processes that we identify with being human". Moreover, there is a possibility that an organism with a reality simulator is more likely to engage in cooperative behaviour because of its ability to conceptualise rewards to others, and long-term rewards to itself.

We maintain that "world model" should appear in the architecture incrementally. In our previous work [4–6] we described the Deep Behaviour Projection (DBP) hierarchical framework. The DBP framework formally represents *increasing* levels of agent reasoning abilities, where every new level can be projected onto a deeper (more basic) behaviour. Put simply, a DBP behaviour can be present in the architecture in two forms: implicit (emergent) and explicit (embedded).

It is interesting at this stage to compare such behaviour duplication in DBP with the distinction between *automatic processes* and *controlled processes* in cognitive psychology. It is well-known that certain processes become highly automatic through repetition and are unconsciously triggered in the presence of certain stimuli, while controlled processes are mostly goal-oriented rather than reactive. With time and/or practice newly learned behaviours often shift from being controlled to automatic.

What the DBP approach suggests in addition, is that *the reactive/cognitive distinction is always relative* in a hierarchical architecture. The behaviour produced by the level $l_k$ may appear reactive with respect to the level $l_{k+1}$ but, at the same time, may look deliberate with respect to the level $l_{k-1}$. Let us exemplify this with the following three levels of the DBP agents:

| | |
|---|---|
| – tropistic behaviour: | Sensors → Effectors |
| – hysteretic behaviour: | Sensors & Memory → Effectors |
| – tactical behaviour: | Sensors & Memory & Task → Effectors. |

The hysteretic behaviour is definitely more reactive when compared with the tactical behaviour, because the latter uses the task states in choosing the effectors. However, contrasted with a very basic tropistic behaviour, the hysteresis provided by (internal) memory states ensures a *degree* of cognition. More precisely, the hysteretic behaviour addresses some lagging of an effect behind its cause, providing a (temporary) resistance to change that occurred previously. For instance, in order to intercept a fast moving ball the agent needs to observe the shift in the ball positions and estimate its velocity before activating the effectors. Thus, the hysteretic behaviour is slightly more deliberate than the tropistic one (exemplified by a simple chase after the ball) — it better situates the agent in time (not only in space) and allows it to better respond to changes. Continuing with the example we re-iterate that the hysteretic intercept is a behaviour embedded explicitly, while the tropistic intercept is only possible as an emergent result of the recurring chase.

Thus, a reality simulator appears *incrementally* — starting from a basic ability to detect a change (eg., in direction) and moving towards a more and more comprehensive incorporation of the temporal asymmetry or "time's arrow" (eg., from direction-sensitive cells to a measurement of a shift in observed positions, to the notion of velocity emerging after a series of measurements, etc.). This means that an emergence of essentially new behavioural patterns always indicates a need for new elements in the agent architecture. At some stage, increasing levels of reasoning about change require an ability to consistently maintain the agent's beliefs — expand, contract or revise them according to some rational principles, such as the principle of minimal change (information economy) [1].

In order to address this requirement, we explicitly introduced a domain model into the DBP architecture, resulting in the following hierarchy (a refinement of the architecture reported in [6]):

$$\begin{array}{rll}
\langle S, E, & \mathit{tropistic\_behaviour} : S \to E, & \\
I, & \mathit{hysteretic\_behaviour} : I \times S \to E, & \mathit{update} : I \times S \to I, \\
T, & \mathit{tactical\_behaviour} : I \times S \times T \to E, & \\
 & \mathit{tactics} : I \times S \times T \to 2^T, & \mathit{decision} : I \times S \times T \to T, \\
D, & \mathit{domain\_update} : I \times S \times D \to D, & \\
 & \mathit{domain\_revision} : I \times S \times D \to D, & \\
 & \mathit{domain\_projection} : I \times S \times D \to S \ \rangle &
\end{array}$$

where $S$ is a set of agent sensory states, $E$ is a set of agent effectors, $I$ is a set of internal agent states, $T$ is a set of agent task states, and $D$ is a set of domain model states. The DBP agents extrapolate their domain model each simulation cycle with the *domain_update* function, and revise it with the *domain_revise* function whenever new information becomes available. The partition between update and revision corresponds to the well-known distinction between belief update and belief revision [3]. In particular, the belief *update* is appropriate when the world has changed and the agents need to accommodate this change into the previously correct beliefs. The belief *revision* should be used to incorporate new information about the same state of the world, in order to correct potential inconsistencies.

In the absence of new observations, the updated domain model $d^* = \mathit{domain\_update}\ (i, s, d)$ is the best approximation of the domain. In these cases, the domain model $d^*$ is transformed by the *domain_projection* function into the agent's sensory state $s^* = \mathit{domain\_projection}(i, s, d^*)$. Very importantly, all the choices made by the agent based on $s^*$ are not distinguishable from the choices it could have made if the same sensory state $s^*$ was a result of the direct sensory input. Intuitively, the *domain_projection* function projects the results of the reality simulator and the agent *imagines* that these results have been observed directly. The projection function is needed only in the absence of new observations, and should not be invoked at other times — the imaginative side of the agent is not needed when "live" information is available anyway.

## 3 Experimental results and conclusion

In order to support our analysis of boundaries on the team coordination potential, we varied communication policies while leaving all other factors (agents skills and tactics) unchanged. The factors beyond our control (eg., a possible change in the opponent strategy) were minimised by repeated runs. This focused the experiment on the dependency (if any) between communication policies (and therefore, resultant joint beliefs) and the team coordination potential.

Our benchmark opponent was selected from the top five teams of the RoboCup-2001 championship. The baseline test team ("Full Communications") was our team running with standard communication messages (512 symbols $\approx$ 100-objects capacity) — we used the protocol of the Soccer Server 7.10. This enabled the full use of the benchmark as well. Then we investigated three communication policies with the protocol $\mathcal{P}_1$. The first policy ("Ball") was to communicate only the ball object, if the data were accurate enough. This mixed variant is quite similar to the transitively-selfish agreement, with high relative entropy and very local coordination, enabling a pressing aggressive game (simply because the players close to the ball might be unaware of each other). The second policy ("Ball | Self") allowed, in addition, each agent to communicate the data about itself according to a schedule, but possibly at times when some other agent communicated the ball object. This mix is much closer to the selfish agreement, with low relative entropy and very global coordination, enabling a passing non-aggressive game (now the players within the ball neighbourhood are often aware of each other, and in addition more team-mates can be considered for a pass). The third policy ("Ball | Self | Wait") prevented self-messages when a team-mate was likely to say ball. This implicit synchronisation is aimed at some mixture of local and global coordination, balancing predominantly pressing game with some passing chances — truly a mixed agreement with (anticipated) bounded relative entropy. The results are presented in the table below.

| Team | Goals For | Goals Against | Wins | Draws | Losses | Points |
|---|---|---|---|---|---|---|
| Full Communications | 111 | 101 | 38 | 29 | 33 | 143 |
| Ball | 123 | 125 | 34 | 31 | 35 | 133 |
| Ball \| Self | 102 | 124 | 26 | 27 | 47 | 105 |
| Ball \| Self \| Wait | 114 | 112 | 35 | 27 | 38 | 132 |

**Table 1.** Results against the benchmark after 100 games for each test.

All the tests have performed, as expected, worse than the baseline. The "Ball" policy achieved almost a parity with the benchmark, while the "Ball | Self" policy was clearly worse. Obviously, this just indicates that the pressing game (emerging as a result of the high entropy of joint beliefs and the ensuing local coordination) is more suitable against this particular benchmark. This conjecture was supported by performance of the "Ball | Self | Wait" policy, achieving an equality against the benchmark as well. Apparently, the information contained in the self-messages and communicated fairly infrequently was not enough to create statistically significant passing chances, and therefore, the emergent coordination was more local than global. Importantly, the third (mixed) policy was *within the boundaries* marked by the first two variants (and closer to the first one), as suggested by the relative entropy of joint beliefs. Similar encouraging results were obtained for extensions of all three policies to the protocol $\mathcal{P}_2$.

These empirical results illustrate the dependency between communication policies, the information entropy of joint beliefs and the team coordination potential. Identification of this relation is a main contribution of the presented analysis, opening a new general perspective on reasoning about belief dynamics in multi-agent scenaria.

**Acknowledgements**

The authors are grateful to Marc Butler, Thomas Howard and Ryszard Kowalczyk for their exceptionally valuable prior contributions to our RoboCup Simulation efforts and studies of multi-agent systems.

## References

1. Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* Bradford Books, MIT Press, Cambridge Massachusetts, 1988.
2. Steve Joordens. Project Cetacea: A Study of High-Level Cognition In Toothed Whales. `http://www.psych.utoronto.ca/~joordens/courses/PsyD58/Cetacea.html`.
3. Pavlos Peppas, Abhaya Nayak, Maurice Pagnucco, Norman Foo, Rex Kwok and Mikhail Prokopenko. Revision vs. Update: Taking a Closer Look. In Proceedings of the 12th European Conference on Artificial Intelligence, 1996.
4. Mikhail Prokopenko and Marc Butler. Tactical Reasoning in Synthetic Multi-Agent Systems: a Case Study. In Proceedings of the IJCAI-99 Workshop on Nonmonotonic Reasoning, Action and Change, Stockholm, 1999.
5. Mikhail Prokopenko, Marc Butler and Thomas Howard. On Emergence of Scalable Tactical and Strategic Behaviour. In RoboCup-2000: Robot Soccer World Cup IV, Springer, 2000.
6. Mikhail Prokopenko, Peter Wang and Thomas Howard. Cyberoos'2001: 'Deep Behaviour Projection' Agent Architecture. In RoboCup-2001: Robot Soccer World Cup V, Springer, 2001.
7. Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1949.
8. Peter Stone and Manuela Veloso. Task Decomposition, Dynamic Role Assignment, and Low-Bandwidth Communication for Real-Time Strategic Teamwork. In Artificial Intelligence, volume 100, number 2, June 1999.
9. Tomoaki Suzudo. The entropy trajectory: A perspective to classify complex systems. In Proceedings of the International Symposium on Frontier of Time Series Modeling, Tokyo, 2000.