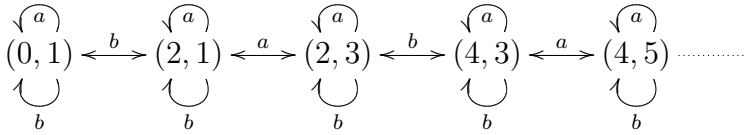
Common knowledge and bounded rationality

Denis Bonnay (IHPST, Paris 1) & Paul Égré (IJN, CNRS)

According to the most received definition, a proposition p is common knowledge between a group of agents provided it is true, everyone knows it, everyone knows that everyone knows it, and so on indefinitely. Common knowledge is a stronger notion than the notion of mutual or shared knowledge, which only requires that everyone know p. An important and controversial feature of this definition of common knowledge is its infinitary character, which is often seen as making too strong an idealization on the logical capacities of the agents. An illustration is provided by the puzzle of consecutive numbers (a.k.a. the Conway Paradox), where two agents each are given a positive number, with the public rule that the numbers are consecutive. A situation in which agent a has a 2 and b a 3 can be depicted by the following Kripke model:



In that situation, the standard epistemic semantics predicts that it is not common knowledge between a and b that their numbers are less than 100000, or even less than any positive number however large it may be, despite the fact that each of a and b knows that the numbers are less than 5. Informally, this corresponds to the fact that a considers it possible that b considers it possible that a has a 4, and so on and so forth. Model-theoretically, this corresponds to the fact that p is common knowledge at w iff p holds at every world in the reflexive transitive closure of the union of the accessibility relations from w.

In this paper, we state a generalization of the standard Kripke semantics (Token Semantics or TS) which makes the metarepresentational resources of the agents explicit, providing a more realistic account of common knowledge in scenarios of this kind. Given a model $\mathcal{M} = \langle W, R_a, R_b, V\rangle$ (the generalization to n agents is straightforward), a sentence is satisfied with respect to a sequence of worlds and a number of tokens available to each agent according to the following rules:

(i) $\mathcal{M}, qw \vDash_{\text{TS}} p\ [m_a, n_b]$ iff $w \in V(p)$.

(ii) $\mathcal{M}, qw \vDash_{\text{TS}} \neg\phi\ [m_a, n_b]$ iff $\mathcal{M}, qw \nvDash_{TS} \phi\ [m_a, n_b]$.

(iii) $\mathcal{M}, qw \vDash_{\text{TS}} (\phi \wedge \psi)\ [m_a, n_b]$ iff $\mathcal{M}, qw \vDash_{\text{CS}} \phi$ and $\mathcal{M}, qw \vDash_{\text{CS}} \psi\ [m_a, n_b]$.

(iv) $\mathcal{M}, qw \vDash_{\text{TS}} K_a\psi\ [m_a, n_b]$ iff

- $m_a \neq 0$ and for all w' such that wR_aw', $\mathcal{M}, qww' \vDash_{\text{TS}} \phi\ [m_a - 1, n_b]$
- Or $m_a = 0$ and $\mathcal{M}, q \vDash_{\text{TS}} \phi\ [1, n_b]$ (and similarly for K_b).

Informally, each token can be seen as the resource that an agent will spend to make a move in the model, until he spends all his tokens, from which iterations of knowledge are made "for free". In the case where two agents have the same number n of tokens, common knowledge will thus follow from n levels of shared knowledge. Thus, adopting the usual syntactic definitions of $C_{a,b}$ (for common knowledge) and $E_{a,b}\phi$ (for mutual knowledge), it is easy to see that: $\vDash_{\text{TS}} ((E_{a,b})^n\phi \rightarrow C_{a,b}\phi)\ [n, n]$, where TS-validity is defined in the expected way. In the case of consecutive numbers, it can be thus be checked that $M, (2, 3) \vDash_{\text{TS}} C_{a,b}$("both numbers are ≤ 5")[1,1].

Three important features of the semantics will be presented: first, the semantics preserves the infinitary definition of common knowledge; second, the standard Kripke semantics is a particular case of TS where the number of tokens is ω for each player; last, with the inclusion of the definition axiom for E and the axiom $E^n\phi \rightarrow C\phi$, the semantics is sound and complete for the system $K4n5n$ (where $4n$ is $K^n\phi \rightarrow K^{n+1}\phi$ and $5n$ is $\langle K\rangle^n\phi \rightarrow \langle K\rangle^{n-1}K\langle K\rangle\phi$), assuming the number of tokens available to each player is n. Thus the framework allows us to reconcile the usual definition of common knowledge with the notion of bounded rationality. Two interpretations of these results will be discussed: in the case of consecutive numbers, we can either see common knowledge as illusorily obtained from a finite amount of shared knowledge, due to the fact that the agents are computationally bounded. Or we can consider that, in such situations, common knowledge may indeed supervene on no more than enough shared knowledge.