Everything I know: a syntactical approach to only-knowing

Consider a group of children who have been playing in the garden. The father of one of them announces "At least one of you has a muddy forehead", and then repeatedly asks them to put their hands up if they know they have a muddy forehead. (The children can see each others' foreheads, but not their own.) Suppose that exactly n of the children in fact have muddy foreheads: at what point do they put their hands up? If $n = 1$, it is easy to see that the muddy child will put his hand up the first time the father asks: seeing that all the other children are clean, and bearing in mind the father's announcement, he knows he must be the only muddy child. If $n = 2$, then the first time the father asks, nothing will happen; but the next time he asks, both muddy children will put their hands up: seeing that there is exactly one other muddy child, and deducing that, since none of them put his hand up the first time, there must be at least two muddy children, they will know that they are muddy. By repeating the same reasoning, in the general case, the n muddy children will all put their hands up after the father asks for the nth time.

This well-known puzzle illustrates just some of the intricacies of reasoning about the knowledge states of idealized agents given limited amounts of information. An agent may represent a human being, a player in a game, a computer, or a robot. The enterprise of formalizing and analysing this kind of reasoning is known as *epistemic logic*, and has been the object of considerable interest among computer scientists in recent decades, for whom the problem of reasoning about the information possessed by communicating agents is of obvious interest. In its simplest form, an *epistemic logic* is a formal language featuring an operator K with the interpretation "The agent knows that ...". Thus, if p stands for the statement "The password has been reset", then $\neg Kp$ is read "The agent does not knows that the password has been reset", and $p \to Kp$ states "If the password has been reset, then the agent knows this". (In formal logic, the symbol $\neg$ is read "It is not the case that ...", and $\to$ is read "If ..., then ...".) Using such a language, principles of reasoning about knowledge and ignorance can be formulated in such a way as to allow the derivation of, for example, the solution to the muddy children puzzle with which we started. Of course, in examples such as this, we need to talk about the knowledge of different agents, or even the knowledge of different agents at different times. But this in general poses no problem, we decorate the K-operator with indices indicating the agent or time concerned. Thus $K_a p$ states that agent a knows that p, with $K_t p$ and $K_{a,t} p$ interpreted analogously. The principles of such logics are well-understood.

Returning, for simplicity, to the single-agent, single time case, consider again the formula Kp, interpreted as "The agent knows that p". This statement imposes a *lower bound* on our agent's knowledge, namely, that includes at least p, but possibly other things as well. But what of *upper bounds*? How can one say that our agent a knows *only* that p, i.e. that he knows p and *nothing else*? It is not hard to think of situations in which such a facility might be desirable. And while we can list some of the things such an agent does not know by writing $\neg Kq \wedge \neg K(p \to q) \wedge \cdots$, such a list may be very long or even infinite. More attractive is the possibility of an operator O with the reading "The agent *only knows* that ...", or perhaps "The agent's *total knowledge* is ...".

But what, exactly, does this mean? In fact, the problem of defining an only-knowing operator is trickier than one might think, and has come to form a sub-field of epistemic logic in its own right. Despite repeated efforts over the past thirty years, there is no completely satisfying account of this concept, especially when dealing with multiple agents and times. The purpose of the research proposed here is to correct this situation. We conjecture that mainstream approaches to the logic of total knowledge are pursuing the wrong strategy. Our point of departure in our research will be an existing, alternative strategy to the logic of only-knowing that can be shown to work well for the single-agent, single-time case, but which has not been developed beyond that point. Our hypothesis is, that, by combining techniques from the two strategies, we can construct a convincing and useful logic of only-knowing able to account for, among other things, puzzles such as the one above. The research will analyse the mathematical and computational properties of the developed formalism.