1. I like subagents

I very much like multi-agent theories of the mind. Here are two reasons not in the Lesswrong post.

1.1. Fast evolution

Humans evolved their intelligence advantage quickly with a sharp acceleration around 80,000 years ago (wikipedia). So any theory that explains human intelligence should require few moving parts. If we assume our most complex cognition relies on multiple subagents that can use the same generic optimization hardware, we explain how we get such bangs for little evolutionary time bucks. A small change would have allowed an organization of the mind that benefits from rushing for more compute.

1.2. Introspection

It intuitively and introspectively makes sense. We sometimes feel inner conflict, as though being pushed in opposite directions by different parts of ourselves.

2. How smart are the subagents?

I am not sure we need to assume subagents do anything like long-term planning for their own goals or taking a lot of context into account. It seems to me the theory still holds with subagents that are closer to simple optimizers with some contextual information (like the immediate emotional context or ideas at the forefront of the mind) and no ability to plot. To fix ideas, we could imagine them as big single-pass neural networks trained with gradient descent (with rich inputs and outputs).

3. How good is the coordination?

In the same spirit, there might be no need for an elegant negotiation procedure that incentivizes agents to come together in some approximation of the best interest of the human they compose. As an alternative, we could consider something like office politics: subagents can delegate to other subagents and come to learn which subagents are trustworthy and which are not. Being trusted is the main way for subagents to accomplish their own goals, so they learn to please their boss. Overall the negotiation procedure would simply be "each subagent controls some resources (acting on the body or communicating with other agents) and can listen to other subagents, they do what's in their best immediate interest and learn what works". In such a system the emergence of system-wide intelligence would rely on the organization chart: "who wants what, who controls what".

4. Concepts and complexity

One question this begs is "how do the subagents communicate?". We could assume that each pair of subagents that needs to communicate will learn an ad hoc "language" by reinforcement learning. But that does not feel right to me. Intuitively it seems to me multiple subagents can communicate with many others. If so there must be some "global encoding", a common tongue they all speak.

Maybe Kolmogorov complexity / algorithmic compression has something to say on the topic. The idea that there is some equivalence between understanding and compression seems to be floating around in computational epistemology. An important aspect is that what is simple (low Kolmogorov complexity) is considered relevant. I would say Solomonoff's induction and simplicity theory are both examples of this assumption. The layers of active inference sounds quite a lot like multiple passes of compression. If some part of the whole process of understanding something can be reduced to compression it would help in knowing what properties the target "language" must exhibit. This would in turn help knowing how well the subagents can communicate.

5. What can we do?

5.1. I listed 3 main ideas on the topic

1. The subagents can be somewhat simple optimizers
2. Their organization can be simple office politics
3. Kolmogorov complexity seems a promising idea with regard to how they communicate

5.2. If we do follow up on these ideas here is what I would probably do first

1. A mathematical toy model for how to organize the mind with small optimizers. What powers do I need to give to different agents and what do they need to want? Maybe we could build a toy prototype with LLMs using natural language to communicate.
2. Try to see how such a model can allow for the mix between epistemology and planning, allowing information to flow both ways and the same systems to do task execution and understanding.
3. Find some properties the mind should exhibit under the toy model then look for them in the literature. These would probably be predictions about what kind of failure state is possible to reach with brain damage. I have a handful of cognitive scientist friends I can ask about this. We can also try to fit the model to more usual failure states (cognitive dissonance, akrasia, trauma, ...)