1 Perfect Imitation

In principle you can perfectly imitate any process with a giant lookup table, or by simulating the physics of the system on a sufficiently powerful computer. However this is impractical. The imitation algorithm has to learn from only inputs and outputs, leaving the inner workings as a black box. It also has to do this while using a reasonable amount of computing power. If the human mind works as we think it does, with most of the computation consisting of 10^{14} synapses firing at 100 Hz then the number of flops needed to simulate a human brain real time is around 10^{16} to 10^{17}, or the capacity of top current supercomputers. Unless the human brain is making use of quantum effects or something else weird, this serves as an upper bound on the amount of processing power needed to imitate a human, as a more abstract or higher level model may be more efficient. However, this is the amount of processing power needed to run a model of a human, not to build that model from data. Also note that human thought may contain a practically stochastic element that comes from thermal noise in neurons. The Imitator is sampling from the same random distribution as the human, not predicting their behaviour.

There are also ethical concerns that if these models are sufficiently human, they are deserving of ethical value and could be suffering.

1.1 Proof about perfect imitating

You can make a perfect imitator of a black box if you have transfinite computing power and know what to ask. This algorithm is based on Kolmogorov complexity and is related to AIXI. Model the human as a function h in $F = \{f : I \to R\}$, this is unrealistic as asking one question may effect their answers to later questions. Here I and R are sets, with their elements stored in a simple format. Let $kk(f)$ be the Kolmogorov complexity of the function and $w(f) = 2^{-kk(f)}$. Let $\Sigma(G \subseteq F) = \sum_{g \in G} w(g)$ and let $G \subseteq F$ with the humans function $h \in G$

Either $\forall i \in I$

$$\Sigma(\{g \in G | g(i) = h(i)\}) > \frac{1}{2}\Sigma(G)$$

So given the set G and an input i, $h(i)$ can be calculated by taking

$$\max_{r \in R}(\Sigma(\{g \in G | g(i) = r\}))$$

Alternatively $\exists i \in I$

$$\Sigma(\{g \in G | g(i) = h(i)\}) \leq \frac{1}{2}\Sigma(G)$$

In this case, if G was your previous hypothesis set, learning $h(i)$ lets you construct a new hypotheses set $G' = \{g \in G | g(i) = h(i)\}$ with $\sigma(G') \leq \frac{1}{2}\sigma(G)$ so if you know what to measure, it takes at most $kk(h)$ measurements to get the function h.

1.2 Behavioural lotteries

Let L be a large number, $I = \{1, 2, \cdots, L\}$ and $p \in I$ be random member of the set. Let $R = \{0, 1\}$ and

$$h(i) = \begin{cases} 1 & i = p \\ 0 & i \neq p \end{cases}$$

then the function predicted by an imitator will be $f(i) = 0$ if p is not an example. If the human has some secret, even if it is just bank details, they would probably react in a similar way to a failed guess regardless of what was guessed, but react differently if the algorithm guessed right first time. So long as the imitator program does not need to act surprised and ask how you knew when told the original humans secret, this might not be a problem. The only way around this is to take apart the system and examine the components, mind uploading.

2 Intelligence Amplification

Suppose the imitation program is perfect. We could, given plenty of resources, replace it with the original. If we had an uploaded human mind, no concerns about upload rights and unlimited computing power, we wouldn't need the imitator. How could we produce superhumanly smart behaviour by arranging a very large number of identical superfast human minds into some sort of organization? The number of hypothetical minds grows exponentially with the number of steps, so could be far larger than the number of atoms in the universe. It is almost certainly possible to get some good answers out of such a system if it is well designed, however there are factors that would seem to limit this. For instance the task needs to be subdividable. Whether or not there are any tasks that can't be subdivided at all is unknown.

2.1 Subgoals

If the problem is too complex for a single mind to easily understand the whole purpose and how they contribute to it, the delegation may include subgoals. These are likely to be simple, approximate measures. The solutions produced may land in the region where the subgoal is not a good measure of performance. Given the goal of designing a national transport system, a subgoal might be to make a plane that is as fast as possible. The final solution might pull lethal G forces or need its engines replaced after every flight or spew polution etc. The performance metric asked for was not actually what was wanted, implicit assumptions about the range of solutions considered were made. The result is that the sub-agents pursues a lost purpose.

2.2 Cognititive Biasies

Ideas are only ever being selected based on what sounds good, or even what is fun to think about, not what is good. If the mind is fundamentally biased or broken in some key way, you might get an endless repetition of fallacious arguments. An agent with bad priors and no meta rule that would correct those priors will keep them, even with unlimited thought time. Fortunately the human can be told about the Kolmogorov complexity version of occams razor, which would provide good priors if they were able to apply it accurately. If they are independently searching for good solutions to an optimization problem, and one of them makes a subtle mistake, it is likely that the mistaken answer comes out far better than any real solution could. The one mistake is then selected for by a maximization. If a million people design a million engines, and they all try to calculate the power output, someone will make a mistake and get an engine they think is 300% efficient. If you just pick the best engine, you get the mistake. In this case it would be obvious there was a mistake, but only because we know conservation of energy provides a hard upper bound on how good an engine can be. If you optimize an imperfect measure of performance, then maximizing the measure can maximize the error. Of course, with a lot of people looking for errors, someone might find one, but these are copies of a single mind and so are likely to have similar thought patterns. Given a huge number of copies making many errors, might one be so subtle and counter intuitive, that it is not spotted. Even if an error is spotted, with so many thinkers, someone will think they found an error in the genuine cases as well. Given a large number of unreliable parts, you need an error checking and handling system.

2.3 Memes

If the people have freeform communication, especially if they have easy copying, counterproductive memes become a big concern. Compared to the modern internet, this may be better from having a smart and serious people that are there to work. Unfortunately it's endless copies of the same person, so the memes evolving will be highly targeted to their particular psychology. The number of people might also be far larger. So for a seemingly sensible serious person, is there some idea that they will want to share with everyone? The idea could be literally anything people look at online and more. Anything from panicing about a potential problem, a funny lolcat, a puzzle, a philosophical concern, a plausible reason why the task they are doing is pointless and more. A misunderstanding could produce plausible and important fake news. The sharers may think they are helping with the task, or they may be overwhelmed by some emotion or more. Human minds are complex and operate on heuristics, they are not resistant to adverse optimization, especially when applied to the particular quirks of that individual mind. If they had to remember and repeat things, long memes would propogate less than if they could just press share. However this create an enormous game of chinese whispers for real messages.

2.4 Total Insanity

Some tasks can't be split into many small pieces to accomplish in parallel. If the person needed to become an expert in a new field for some task, this could reguire subjective years of study. Years without any company besides other yous, possibly doing nothing but studying. The effects of isolation on mental wellbeing are quite severe. How long will it take for a mind to go bonkers.

2.5 Limited Solutions

If the system contained many different minds, the problems with extreme optimization fitting memes to the nieches of a particular psyce disapears. If they are given an opportunity to socialize and play, most of them won't go mad on the timescale of human lifetimes. (leave them for long enough and they might) This is a virtual scociety of human minds, working and relaxing togeather. Whether such a scociety would be more productive when organized as a democracy, dictatorship or something else is left to political science.

Another alternative is if there is no opertunity for the memetic propegation. Each mind is a mathmetition. The top of a hirarchy is given a formally specified problem, and breaks it down into smaller problems. This proceeds recursively until the parts are small enougth to solve, with everyones work being checked by an automated theorem prover. This only produces theorems, you are at best using this because virtual mathematicians use less processing power than an exhaustive search through all theorems. It seems likely that some heuristic or reinforcement learning agent would take less processing power, as well as not having virtual mind rights issues. In any case, this does not solve the human values problem.

3 Conclusion

Suppose the technical problems relating to effective copying can be solved adequately and the ethical problems are solved or ignored. This system would be able to best a single human in some areas, by imitating a large group of humans, complete with all the organizational and memetic evolution problems that entails.