The absentee billionaire

Paul Christiano Follow
Feb 20, 2015 · 5 min read

Once each day, Hugh wakes for 10 minutes. During these 10 minutes, he spends 10 million dollars. The other 1430 minutes, he slumbers.

Hugh has a goal. Maybe he wants humanity to flourish. Maybe he wants to build the world's tallest building. Maybe he wants to preserve the rainforest. Whatever Hugh wants, our question is: how can he get it?

The desideratum

Hugh has commissioned us to come up with a system for him to use—to tell him how to spend his ten minutes, what to look up, who to talk to, and how to decide where to send his money. We can give Hugh an instruction manual and spend a few hours explaining the system to him. After that, he's on his own.

Hugh wishes he could pay people to adopt the goal "Hugh's goals are satisfied" wholesale, to pursue it as effectively as they can. In the best case, he could do this with no more expense—and no more difficulty—than would be required to hire them to do similar work. Let's call this ideal arrangement *perfect delegation*.

We'll consider our system a success if it lets Hugh achieve his goals nearly as well as he could using perfect delegation.

With access to perfect delegation Hugh could find the best team money could buy and hire them to administer his budget in the service of his goals. They would in turn identify and evaluate opportunities, and use perfect delegation to implement them. Doing the same using only 10 minutes a day, without perfect delegation, would be a tall order.

The details

Some details of Hugh's situation:

1. Hugh has access to the internet. He can send and receive emails, can browse the internet, and so on. He can attach money to emails painlessly.

2. In principle, the people of Earth could all agree not to accept Hugh's money. On average, this would make them better off. But it's not going to happen.

3. The world isn't on Hugh's side, either. He doesn't have any super-trustworthy assistants, much less thousands of them. There are just a lot of people who want their share of $10M / day. In the easy version of the problem, there are also some people who share Hugh's vision—but even then, who can tell the difference?

4. Hugh has written a lot about his goals and his outlook. Anyone who's curious can go learn about Hugh on the internet; lots of people have.

5. If someone doesn't want Hugh to learn something, they can pay a news site not to cover it. They can launch a DDoS against Google. They could even go to more extreme lengths. But we'll assume that Hugh's connection to the internet is itself tamper-proof. And just like the world won't coordinate to turn down Hugh's money, they won't coordinate to set up a parallel version of the internet just to delude him.

6. Hugh has no prospect of fixing his debilitating sleep disorder. He could leave his room if he wanted, but what is he going to do in 10 minutes, anyway? He's already arranged for his room to be secure and well-stocked, and for his infrastructure to be reliable.

7. Hugh's goals do not require the cooperation of any specific person, or any unverifiable private information. Everything that Hugh wants to do could be done by one of several people (and, as per assumption #2, we assume that these people will not all collude). All of the information that Hugh needs could in principle be verified by Hugh, if he had enough time to spend veryfing it.

The prognosis

Hugh's situation would not be interesting if it were hopeless. I'm posting this puzzle because I think it pushes the boundaries of what is possible without going beyond.

I have a few ideas and a partial solution. I'll present some of them in upcoming posts. I'd also love to see other approaches to the problem!

The metaphor

Hugh's situation is a caricature, but the basic problem seems ubiquitous. If available, perfect delegation would be useful in many domains: from philanthropists making grants, to voters electing representatives, to executives managing companies. A lot can be lost between intention and implementation.

The deeper reason I care about Hugh's predicament is that I think it illustrates a much more fundamental difficulty. As a society, I don't think we yet have the hang of building organizations that can effectively pursue an on-paper mandate, or that can make decisions explicitly and rationally. Instead we get bureaucracies, we get coalitional politics, we get crude rules applied bluntly, we get inertia, we get strong cultures and personal networks that can only grow so far before they decay or drift. In many cases, the best we can do is to leave decisions in the hands of the best-equipped individuals we can find and trust.

Setting aside this more abstract difficulty, there are two concrete cases I find particularly interesting.

Philanthropy. A philanthropist with a pure heart seeks the most effective way to use their money to improve the world. They are surrounded by people pursuing their own projects for their own reasons, uninterested and sometimes unable to think about their expertise in the context of the philanthropist's broader goal. How can the philanthropist make funding decisions in a domain without implicitly adopting its practitioner's values? How can the philanthropist best leverage the knowledge of many diverse experts?

I think we have a lot of room to improve our basic institutions for this setting—Hugh's extreme situation makes the problem especially obvious, but I think it is always there. The Open Philanthropy Project is making a noble effort to attack this problem head on, and I expect they will do a lot of good. So far they are relying on a combination of small scale and staff with closely aligned values, and it will be interesting to see how these efforts scale. (I expect that the practical issues the OPP is struggling with are more pressing than the theoretical question; but I think that theoretical question has still received less attention than it deserves.)

AI Control. Unless we go out of our way to avoid it, we humans will eventually live in a world that outpaces us as much as our world outpaces Hugh. If things go well, we may also be comparably rich (in absolute terms). One way of understanding the AI control problem is: how can humans continue to influence the overall direction of society when decisions are made automatically, much more quickly and much more frequently than humans can directly oversee?

The analogy is not entirely straightforward, but I expect that a solution to Hugh's problem would also prove useful for attacking the AI control problem.