The last CAPTCHA

Paul Christiano Follow

Apr 5, 2016 · 6 min read

I think we might be nearing the last moment in history when it is possible for a human to beat a machine at *any precisely defined game*.

To formalize this, consider the following meta-game:

- The **CAPTCHA designers** produce a CAPTCHA and publish it.
- The CAPTCHA is built to interact with two agents at once, and has to guess which is a human. The interaction should be through a keyboard/mouse/monitor, take place over the course of at most 8 hours, and not involve the outside world. For example, the CAPTCHA can organize a game of Go between the two agents.
- We pick the single human who is most likely to be able to win the CAPTCHA (the **champion**). A team of coaches can spend a year trying to help the champion train to win at the CAPTCHA.
- A sophisticated research team spends a year designing an AI to beat the CAPTCHA (the **machine**). The designers can use any resources at their disposal, and as much computing power as they can get. They have access to the source of the CAPTCHA and can run it as many times as they like.
- We play a sequence of matches between the champion and the machine, starting one year after the publication of the CAPTCHA.
- The CAPTCHA designers and the champion win if the CAPTCHA can identify the human in 2/3 of trials.
- The AI designers win if the machine can pass as human at least 1/3 of the time. (We could replace 1/3 with any constant less than 1/2.)

Until very recently it was clear that the champion would win, since Go is a simple counterexample and it seemed unlikely to be solvable with a year of effort.

My view of the current state of affairs

I don't think that Go itself is the last CAPTCHA. But I do think that we are rapidly running out of good contenders:

- There are board games and strategic video games which humans can currently win, but there are none which I would expect to stand up to a year of serious attention.
- Similarly, there are perceptual problems where humans can still outperform machines. But given access to any distribution over instances, I wouldn't count on that particular distribution lasting a year. Also, we can't really collect a big enough data set of natural images that the AI team couldn't just label them all by hand.
- There are motor control problems with well-defined success criteria where a human will beat a machine. But things get easier when the task occurs in a simulation. And they get much easier when the distribution over tasks is narrow enough that it can be algorithmically generated, and when the AI designers are willing to use arbitrarily special-cased software. I don't know if there any real contenders here.
- There are tasks involving symbolic reasoning that seem hard for AI systems. But it is usually very difficult to generate hard instances. For example, a "random" math problem is more likely to be easy for a machine than a human—humans excel at a narrow class of problems that we find interesting. And we can't algorithmically sample instances from that distribution.

My best single guess is that we will run out of CAPTCHAs within ~5 years. What do you think? What are the strongest candidates for robust CAPTCHAs, and how long do you expect them to last?

The range of Turing tests

We can consider three increasingly difficult tests. In each, the judge has some moderate length of time, say a day, to make their determination:

- **General CAPTCHA**. An automated judge interacts with two agents, and guesses which is a human and which is a machine.
- **Adaptive CAPTCHA.** Rather than publishing a CAPTCHA and then designing an AI, the AI designers *first* publish their AI, and

then the CAPTCHA designers choose a CAPTCHA to beat that AI. The AI can train against the CAPTCHA for a year, but that process needs to be entirely automated.

- **Turing test**. A human judge interacts with two agents, and guesses which is a human and which is a machine.

Each of these tests seems significantly more difficult than the last. I expect to see significant gaps in time between these three milestones, though I wouldn't be too surprised if two of them (or even all of them) were accomplished essentially at the same time.

I don't know where "radically transforming the world" fits in the list. My guess is that it comes between adaptive CAPTCHA and the Turing test.

We can vary the tests in other ways. In particular:

- We could restrict the communication to be entirely text.
- We could compete with the 50th percentile human, or the 99th or the 99.9999999th.
- We could give the human champion access to an arbitrary AI assistant, designed over the course of the year (in the same way as the machine). Once AI systems are sufficiently powerful, the machine should still be able to win.
- In the non-adaptive CAPTCHA cases, we could allow a human to produce a single input which is given as advice to the judge and also given to each player. For example, the human might pose a chess puzzle that would be particularly easy for a human to solve, even if we couldn't algorithmically sample from any distribution that is especially easy for a human to solve.
- We could allow the judge to take 5 minutes, 5 hours, 5 weeks, or 5 years.
- We could give the human AI designers no time (as in adaptive CAPTCHA), 1 week, 1 month, 1 year, or 1 decade.

Testing with markets

Here is a simple procedure for trying to figure out whether we can design a CAPTCHA. It can be easily modified to cover the adaptive CAPTCHA and Turing test.

We offer a $1M prize for producing a winning CAPTCHA. We sell the right to produce the CAPTCHA to the highest bidder (for the second highest bidder's price).

We announce the winning CAPTCHA, and then auction off the role of both the champion and the AI designer. The winner of the game will be given a $1M prize, whether machine or champion. We sell the right to participate in the game to the highest bidder on each side (for the second highest bidder's price).

We don't expect either the champion or the AI company to actually pay their own way, but instead anticipate that they will be backed by investors in return for a cut of their profits if they win.

The proceeds from both auctions are used to increase the prize pool for the final match. If the final match is won by the human champion, then we also pay $1M to the CAPTCHA designer.

If the human wins, we conclude that we can build an adequate CAPTCHA. If the human loses, we conclude that we can't.

For robustness we would really pick several CAPTCHAs, and several champions and machines for each one. I think the design of market structures for this kind of exercise is quite interesting and subtle (and there are problems with this simple version), but I don't want to dwell on it.

Cheating

The CAPTCHA designer should not be allowed to communicate privately with the human champion. For example, they could design a CAPTCHA requiring inverting a trapdoor OWF, and give the human champion the secret key. I don't have any proposal for ruling this out other than looking at the behavior of the CAPTCHA/champion to see if they seem to cheat in this way.

Unless the market is carefully set up there may also be incentives for someone to buy the right to design the CAPTCHA in order to rig the subsequent match. I think this is probably OK if you have a sufficiently robust market structure and can rule out the kind of cheating described in last paragraph.

Conclusion

I think there is a good chance that we will run out of CAPTCHAs quite soon, even if we use a very liberal definition. I think this is a really interesting and exciting milestone. I'd be quite interested to see a serious effort to design a hard CAPTCHA, both from the perspective of forecasting and from the perspective of more clearly understanding the nature of the AI problem.