Methodology: I input probabilities into the "Generate Future" widget on the "Relevant Pre-AGI Possibilities" blog post. The numbers I input were for the thing happening in the next 5 years, rather than for the thing happening prior to AGI. I then generated a 5-year future, and wrote a vignette for it. Then I rolled some dice to see whether AGI would be created in the next 5 years; turns out it is due to happen in 2029. Then I went through the widget again and adjusted upwards and downwards various probabilities, in light of how the world changed in the first 5 years, and also in light of the fact that AGI would be coming in 2029. Then I generated another 5-year future and recorded the results. I would have continued this process until AGI appeared; it just happened to appear pretty quickly, in 2029. Finally, once AGI did appear, I wrote a new section for how I think it would happen given the prior events and developments, and then I wrote a final section for what I think would happen after AGI appears.

I had tons of fun with this, and highly recommend this methodology.

2020 - 2025:

Dramatically improved computing hardware (TML, TAS, POL)
Prediction tools (TML, TAS, POL, CHA, MIS)
Shift in favor of counterespionage (POL, CHA, MIS)
Credible commitment mechanisms (POL, CHA)
Deterioration of collective epistemology (TML, TAS, POL, CHA, MIS)
Automation of human labor (TML, TAS, POL, CHA, MIS)

Several new hardware designs come online that make the cost of compute for AI fall off a cliff; whereas Moore's Law had been faltering before, now we are actually slightly ahead of schedule.

Tesla completes their self-driving software in 2021 and wins the legal battle to have autonomous robo-taxis by 2023. By 2025, other competitors have caught up, and about ten million Americans alone have lost their driving jobs (ubers, taxis, truckers, bus drivers, etc.) Meanwhile, various other industries (Delivery? Call centers? Retail?) continue to lose jobs due to automation, and the pace seems to be accelerating.

It turns out you can use loads of compute to make prediction markets for AIs. Sorta like an efficient but less capable version of Logical Induction. Sorta like population-based training for prediction bots. It works with loads of data. This technology develops rapidly over a few years, just as language tech (e.g. GPT-2) developed in the five years prior. It doesn't change the stock market much, because the market was already pretty efficient, but it starts to revolutionize politics, war, and project management. (Big companies start to require their employees to play in internal prediction markets or prediction tournaments, partly to promote epistemic best practices but partly to provide data for the algorithms.) As a side-effect of this sort of thing, it becomes very easy to automatically detect hostile or unauthorized intrusions into a network--the AIs can tell what sort of traffic is normal and what isn't. Counterespionage wins the day, for now.

Collective epistemology continues to deteriorate. The traditional news media either goes bankrupt or surrenders to the new propaganda-clickbait-and-outrage-porn normal. Most Americans get their news from Reddit, Facebook, or other such sites, and most of the "news" they get is a combination of literal propaganda planted by various world governments, clickbait, and outrage porn. Sadly, this starts to significantly influence what academics, civil servants and wonks, politicians, and other strategically relevant individuals spend their time and energy thinking about. Tribalism and polarization increases; both the Democratic and Republican parties start to fracture into smaller sub-cultures that hate each other as much as they hate the republicans/democrats.

Blockchain sees another resurgence akin to the crypto boom of the late 2010's. This time, "smart contracts" are all the rage. Enforcement of some contracts can be automated now, which serves as a credible commitment mechanism for some kinds of actors. Corporations use this technology to make deals with each other sometimes, deals for which it is doubtful whether the government would enforce them.

2025-2030:

Dramatically improved computing hardware (TML, TAS, POL)
Advanced science automation and research tools (TML, TAS, CHA, MIS)
Prediction tools (TML, TAS, POL, CHA, MIS)
Shift in favor of counterespionage (POL, CHA, MIS)
AI interpretability tools (TML, TAS, POL)
Manufacturing consolidation (POL)
Shift in level of public attention on AGI (TML, POL, CHA, MIS)
Change in investment in AGI development (TML, TAS, POL)
New social movements concerning AI (TML, TAS, POL, MIS)

Let's start with society. The wave of automation we saw in the previous 5 years continues, though the pace no longer seems to be accelerating; it's the new normal. But many people hate the new normal. There is now massive public attention on AI and AGI. There is also a strong social movement to ban or restrict AI, mostly a reaction to all the job losses from automation. Public discourse is terrible; at least it hasn't gotten any worse since the early 2020's. But still, it's pretty bad — the debate is basically "Techie elitists are making robots that oppress us by taking our jobs" vs. "Ignorant neo-luddites hate progress and wish to reinstate the old racial and class hierarchies." Cooler and wiser heads still exist, of course, but they have much less influence. Concerns about AI safety and the future of humanity are raised much more often than in the past, but these concerns are now politicized; e.g. "I'm not just bitter because I lost my trucking job; I really do think AI is going to kill us all, and you fuckers don't care because you think you can merge with it anyways, you creeps" vs. "Those doomsayers are just an alliance between people bitter about automation and people who love sci-fi apocalyse fantasies"

Also as a consequence of automation, there is a massive push for AI interpretability tools. When a study indicates that Teslas are slightly more likely to run over black or hispanic people than white or asian people, the politicians and academics demand to know why. Enough mainstream academic and industry researchers devote themselves to this that substantial progress is made. Modern AI systems are composed of parts, some of which are admittedly black boxes, but still, a surprising amount can be said about how the parts relate to each other and there are tools for guessing what the black boxes are likely to do.

Also as a consequence of automation, there is a huge amount of investment into AI hardware and AGI research. Fixed costs explode relative to variable costs; this kills off or merges most of the hardware startups, until there are just three corporations in two locations producing most of the world's AGI-relevant hardware in enormous factory complexes. This research continues to pay off, however. Surprising many, hardware continues to improve even faster than Moore's Law would predict.

Counterespionage gets even better, since the defender has the data about their network and so they can easily detect intruders who don't know how to "act normal" since they don't have the data. This even includes detecting "potential snowdens" before they act.

Prediction tools continue to improve, faster than expected. Turns out there's a lot of cross-pollination between the science of forecasting, AI science, and the science of science. These three fields start to merge into one, and the algorithms used start to resemble each other more and more. Prediction tools lead to science automation and research tools: Project management software for R&D projects, superhuman intuitions (at least in some domains, some of the time) about what sorts of research proposals are promising and which are likely to be dead ends, really efficient simulations which allow for rapid prototyping and optimization, and really good automatic design, training, and testing of new AIs.

The AI safety community uses prediction tools (plus common sense, I mean look at what's been happening!) to become fairly convinced that some sort of singularity will happen by 2030. However, there is debate over how it will happen. Agents, or tools? Fast, or slow? Outside the AI safety community, various "in the know" actors (e.g. hedge funds) came to similar conclusions, but think through the implications less well. The prediction tools, though good, aren't so good at seeing past the singularity. Not surprising, especially since they depended to some extent on human collective epistemology.

2029: AGI

In retrospect, it was obvious that these conditions would swiftly lead to superhuman AGI. Here's how it happened:

Better hardware led to better R&D tools and more automation. Better R&D tools designed better hardware, and more automation built more of it more quickly. Repeat. Hardware technologies like optical, quantum, and neuromorphic computing — which only began to become mainstream in the early 20's--were advanced to very mature levels by 2028.

Thanks to interpretability, these tools really did resemble tools more than agents. As a whole, the loop of hardware —> better AI tools —> better hardware designs —> more hardware was basically all done by non-humans by 2029, but human overseers, assisted by powerful interpretation tools, watched over it and guided it at every step. Within the AI safety community, people began to say that Drexler's CAIS was looking more and more correct every day, whereas more agent-focused predictions like those of Bostrom, and Yudkowsky were incorrect. Christiano was judged to be somewhat in between, since we seemed to be in a slow takeoff but it seemed to be less agenty than he expected. Outside the AI safety community, people didn't pay much attention to these thinkers--or at least, not much more attention than they paid in the 2010's.

The general consensus of the market was that the big money wasn't in AGI, but rather in R&D tools and task-specific automation. No need to create a new species; it's much quicker and easier to keep doing what we are doing, automating away specific tasks in the economy and making a profit by replacing entire industries. But nevertheless, there was enough general investment in AI that quite a lot of it went to AGI specifically. Besides, DeepMind and OpenAI were still around, and that was their founding goal.

It being only 2029, and compute being abundant, AGI was achieved not by finally understanding intelligence and rationality and using that understanding to build an agent, but rather by leveraging R&D tools. R&D tools (themselves by this point pretty byzantine) were tasked with producing a system that met a list of criteria deemed sufficient to count as AGI; they were given mind-boggling amounts of compute to experiment with; they delivered, mostly by trying many things and iterating quickly on what seemed to be working.

There wasn't really a period of weak or expensive or slow AGI prior to advanced AGI. The reason was that the problem had more to do with specifying the correct criteria than with slowly getting better at meeting them. Early attempts to build AGI in the 2026-2028 era failed in that they produced systems that were really good at the set of games and tasks used in training, but not good at generalizing to other things. We could give our R&D tools a dataset and a metric and they would design an architecture that would perform very well at the dataset and metric, but we had trouble designing a dataset and metric that would result in a system that could perform well on importantly different tasks. What does generalization mean, exactly? How do we instruct our R&D tools to produce an architecture that generalizes well? There were other conceptual difficulties too. But once they were solved (or more accurately, muddled through) the R&D tools churned out agent AGI even faster than DeepMind churned out Go engines back in

the year of AlphaGo. After all, they were pretty powerful R&D tools at this point; they sure knew how to make metrics go up. AGI blew past the human range of intelligence in no time.

Because of manufacturing consolidation and counterespionage, there were only three (and basically only two) corporations in the running for creating AGI, and there was basically no involuntary information transfer between them in 2029. Being rivals, there was no voluntary transfer either--they were in a race. This, combined with the tricky nature of the problem described in the previous paragraph, meant that one project achieved superhuman agent AGI with a four month lead; in particular, the other project had just figured out the path to victory when the first project completed it.

***2030:** Struggle for control of the future*

On the one hand, superintelligent agent AGI created by a process several steps removed from human understanding. Alignment failures, here we come! On the other hand, decently good AI interpretation tools.

On the one hand, a single powerful corporation with a 4-month window in which they have the only agent AGI in the world. World takeover, here we come! On the other hand, pretty good prediction and R&D tools held by other powerful actors in the world, which is already heavily automated. Maybe there are low-hanging fruit for the AGI to pick, but scientific research, technology development, and doing narrow tasks better and faster than humans are not among them.

Not sure what will happen. I'll think about it more, maybe write multiple endings. Suggestions welcome.