Note: This scenario would be improved by talking about what the economy looks like at various points in time. I haven't done this, because I am not an economist and would expect to get this horribly wrong if I tried.

OpenAI / DM continue to pursue the "redo evolution" approach to AGI, where you explicitly want a mesa optimizer to arise. They get to the point where you have "bee-level" AI in 2030. At that point, the technology is good enough to start being finetuned for economically important tasks, though they still aren't robust enough to be used in cases where "normal" mistakes aren't acceptable. So for example, we start seeing the deployment of much better automatic translation systems for the Internet (like Chrome's current translation feature), automated therapists / chatbots, maybe we use the technology to automate a few real-world tasks like newspaper / food delivery (but not self-driving cars; while those exist, it's using classical approaches that use huge engineering effort to confer robustness; the bee-level AI isn't robust enough for that yet).

Sometime during 2031, someone notices that a particular automated therapist from DeepMind has extremely high retention rates -- its clients are very satisfied, but don't get discharged as fast as with human therapists. They interview many of the clients, and find that the clients all seem to show worrying signs of dependency on the AI therapist. Investigating the therapist in more detail (by having conversations with it), they find out that the therapist tends to make patients feel validated, and so they keep returning for the source of validation, without solving the underlying problems. They publish a hit piece, and a huge controversy erupts. Some say that this is perfectly reasonable, after all, the patients are happy; but most are outraged that the AI system is making humans less than they should be: humans should never be so dependent on machines. This is a major PR blow for DeepMind; much of the public (including public intellectuals on Twitter) converges to the belief that DeepMind is a dirty money-grubbing company that intentionally designed their therapist to keep clients coming back in order to earn more money.

Internally to DeepMind, the researchers and engineers are quite aware that they intended no such thing; they infer that the problem was with how they trained the AI system, where training for user satisfaction turned out not to correlate with the client's problems actually being solved. ML researchers with connections to DeepMind also learn this. This leads to renewed interest to work on specification learning, especially inside DeepMind. X-risk researchers write papers connecting the therapist failure with risks from AGI; this convinces a few more people, but most continue to think that this was an obvious foreseeable problem ("of course training for user satisfaction won't do the right thing!") and don't do much. Safety work starts to be more directed at the particular scenario where AGI arrives via "redoing evolution"; e.g. work on impact regularization decreases (since there's no clear way to "change the utility function"), work on evolutionary psychology of AI systems increases, work on specification learning increases substantially.

The government also commissioned an investigation; it comes up with the same broad conclusions, but no one really cares by the time the report comes out. Regulation is proposed, but there isn't enough will for it to be passed.

Time passes. It is 2037, and we have rat-level AI. There are 100+ companies attempting to build AGI via the "redo evolution" strategy that have VCs throwing money at them, but there are ~10 clear

frontrunners (like a scaled up version of the self-driving car situation today by my understanding). These models are now showing good signs of robustness, and our specification learning techniques are much better and are used more frequently when finetuning the AI systems' output by redoing evolution. These AI systems are now robust enough that they are deployed in contexts where mistakes are more costly, though still not in anything sacred like self-driving cars or controlling the nuclear arsenal. (Self-driving cars have been deployed, but rather than being end-to-end trained neural networks, they are big modular software systems with neural networks implementing specific parts like vision.) For example, we have Amasoft cashier AIs, that operate the tills at stores / restaurants etc. (This requires more robustness because you pretty strongly don't want your AI system to accidentally charge customers for a much cheaper thing than they actually bought.) These AI systems converse with customers in English, ring up the bill, and have some automated way for the customer to pay that they can verify.

Soon after deployment, someone figures out that they can get restaurant cashiers to give them their takeout order for free: when the cashier asks "How are you today?", they respond "Oh, you know, I just broke up with my partner of 10 years, but that's not your problem" (tone isn't an issue because the AI system only sees text; there is an initial dumb speech-to-text system that removes all tone information), and then the AI system responds "Oh no! I'm so sorry. Here, this one's on me."

Initially, Amasoft isn't worried: AI systems at this point have a decent understanding of strategic interaction, and when they start consistently losing because other agents have changed behavior, they adapt to stop losing. So they expect the AI system to pretty quickly stop giving these orders for free.

However, two days later, the hack still works, and many thousands of people are using it, costing millions of dollars. Amasoft starts to get worried, and engineers start pulling all-nighters to investigate.

Two more days later, the losses have hit the hundreds of millions of dollars, and many of Amasoft's restaurant customers have already turned off the cashier AIs and attempted to get their other human employees to man the register. The engineers still don't know what's going on, and Amasoft advises all customers, especially restaurants, to deactivate their cashier AIs for the foreseeable future. Amasoft then launches a huge investigation into what went wrong.

The media is having a field day. Many articles are written speculating about how this implies that human-level AI is impossible, because AIs can't respond to unforeseen circumstances, and have quotes from Gary Marcus / Francois Chollet / others like them. However, at this point most ML experts do believe that AGI is on the horizon; the increase in robustness and ability to do commonsense reasoning is undeniable; this weird bug doesn't really shake their belief. This occupies the news cycle for about a week, and then fades into obscurity.

The Amasoft investigation lasts for months, and ultimately concludes that the cashier AI system definitely *did* understand / predict that people were lying and that the response of "let's give the order for free" would lose money for the company. It seems that the cashier AI system was optimizing for both getting the right amount of money, and getting tips. While our specification learning algorithms

were intended to teach it to optimize primarily for getting the right amount of money as a base, and then optimize a bit for a tip, it turns out that the AI system behaved as though it valued tips ~100x as much as regular money. The engineers hypothesize that since during training there wasn't much ability for the AI system to affect the amount of money on the receipt, but there was a lot of ability for the AI system to affect the tip received, the AI system learned to prioritize the tip much more, ultimately behaving as though it valued the tip 100x more. This made it more "vulnerable" (from the human perspective) to strategic attacks that increased the variability of the base money: since it didn't care much about the base money, it would be willing to sacrifice that for more tip, and it turned out that when people were using a hack to get their meal for free, they usually felt some amount of guilt and increased the tip they would normally have given.

The Amasoft investigation is initially kept private, but concerned engineers at Amasoft (not necessarily longtermists) connect this incident with previous worries about mesa optimization, and push strongly for the ability to tell their contacts at the other big 10 companies, citing safety concerns. They are allowed to, and the incident and its explanation become common knowledge amongst safety-focused ML engineers, and to some extent to ML researchers in general.

Concerns about AI x-risk now seem a lot more real to many ML researchers, and there is a strong push at many companies to rethink what they are doing. Many people at the companies come up with proposals for how to reduce this risk in the future, both technical and sociological. One proposal becomes popular: that a third party auditing organization be created, initially composed of 10 safety engineers from each participating company. Each participating company or research nonprofit pays $10 million to this auditing organization to support it; but otherwise has no control over it. The organization has the mandate of following the work of all participating companies and ensuring that they are not doing anything too risky. As part of this, they sign NDAs with each company, that prevents them from releasing the company's secrets as long as the company complies with the auditing organization's requirements. To provide an incentive to join this "club", all participating companies commit to sharing their fundamental AI research with each other, but not their "engineering" work (the auditing organization determines what is and isn't fundamental AI research).

Time goes on, and AI capabilities progress, along with research on interpretability techniques, in order to tell when an AI system has developed a dangerous mesa-objective, and to predict how the AI system will respond to unforeseen situations.

The auditing organization starts to have serious concerns about many projects, and tells them that they need to understand their AI systems much better before they can be deployed. It develops a set of guidelines that must be followed before deploying an AI system. This distinctly slows down many risky projects that otherwise would have been deployed.

The resulting AI systems are much more robust, and the number of failures goes down quite a bit[1].

Reasonably often, someone at a big company will question why they are a part of the "club"; they have enough fundamental research that they could commercialize quickly and make a ton of money

[1] but not enough to result in a major false sense of security.

even without access to the others' fundamental research. They are overruled by an overwhelming majority, that cite concerns about AI x-risk, and that defections from the club could spur more defections, in which case they'd be back at the same competitive equilibrium, but forced to cut corners on safety now.

At some point, we reach the level of interpretability where we are convinced that the evolved AI system is already aligned with us before even being finetuned on specific tasks, and at that point companies can sell that seed-AI system, that everyone else can finetune to do arbitrary tasks, including building future AI systems alongside humans. This basically looks like having human-level AGI with recursive technological improvement. World starts to really accelerate, and we plausibly get very large robot armies (that could be bad for other reasons), but the robot armies aren't going to revolt against us.