Self-driving as a case study for AGI

Jan 21, 2024

Sparked by progress in Large Language Models (LLMs), there's a lot of chatter recently about AGI, its timelines, and what it might look like. Some of it is hopeful and optimistic, but a lot of it is fearful and doomy, to put it mildly. Unfortunately, a lot of it is also very abstract, which causes people to speak past each other in circles. Therefore, I'm always on a lookout for concrete analogies and historical precedents that can help explore the topic in more grounded terms. In particular, when I am asked about what I think AGI will look like, I personally like to point to self-driving. In this post, I'd like to explain why. Let's start with one common definition of AGI:

> AGI: An autonomous system that surpasses human capabilities in the majority of economically valuable work.

Note that there are two specific requirements in this definition. First, it is a system that has full autonomy, i.e. it operates on its own with very little to no human supervision. Second, it operates autonomously across the majority of economically valuable work. To make this part concrete, I personally like to refer to U.S. Bureau of labor statistics index of occupations. A system that has both of these properties we would call an AGI.

What I would like to suggest in this post is that recent developments in our ability to automate driving is a very good early case study of the societal dynamics of increasing automation, and by extension what AGI in general will look and feel like. I think this is because of a few features of this space that loosely just say that "it is a big deal": Self-driving is very accessible and visible to society (cars with no drivers on the streets!), it is a large part of the economy by size, it presently employs a large human workforce (e.g. think Uber/Lyft drivers), and driving is a sufficiently difficult problem to automate, but automate it we did (ahead of many other sectors of the economy), and society has noticed and is responding to it. There are of course other industries that have also been dramatically automated, but either I am personally less familiar with them, or they fall short of some of the properties above.

partial automation

As a "sufficiently difficult" problem in AI, automation of driving did not pop into existence out of nowhere; It is a result of a gradual process of automating the driving task, with a lot of "tool AI" intermediates. In vehicle autonomy, many cars are now manufactured with a "Level 2" driver assist - an AI that collaborates with a human to get from point A to point B. It is not fully autonomous but it handles a lot of the low-level details of driving. Sometimes it automates entire maneuvers, e.g. the car might park for you. The human primarily acts as the supervisor of this activity, but can in principle take over at any time and perform the driving task, or issue a high-level command (e.g. request a lane change). In some cases (e.g. lane following and quick decision making), the AI outperforms human capability, but it can still fall short of it in rare scenarios. This is

analogous to a lot of tool AIs that we are starting to see deployed in other industries, especially with the recent capability unlock due to Large Language Models (LLMs). For example, as a programmer, when I use GitHub Copilot to auto-complete a block of code, or GPT-4 to write a bigger function, I am handing off low-level details to the automation, but in the exact same way, I can also step in with an “intervention” should the need arise. That is, Copilot and GPT-4 are Level 2 programming. There are many Level 2 automations across the industry, not all of them necessarily based on LLMs - from TurboTax, to robots in Amazon warehouses, to many other “tool AIs” in translation, writing, art, legal, marketing, etc.

full automation

At some point, these systems cross the threshold of reliability and become what looks like Waymo today. They creep into the realm of full autonomy. In San Francisco today, you can open up an app and call a Waymo instead of an Uber. A driverless car will pull up and take you, a paying customer, to your destination. This is amazing. You need not know how to drive, you need not pay attention, you can lean back and take a nap, while the system transports you from A to B. Like many others I’ve talked to, I personally prefer to take a Waymo over Uber and I’ve switched to it almost exclusively for within-city transportation. You get a lot more low-variance, reproducible experience, the driving is smooth, you can play music, and you can chat with friends without spending mental resources thinking about what the driver is thinking listening to you.

the mixed economy of full automation

And yet, even though autonomous driving technology now exists, there are still plenty of people calling an Uber alongside. How come? Well first, many people simply don’t even know that you can call a Waymo. But even if they do, many people don’t fully trust the automated system just yet and prefer to have a human drive them. But even if they did, many people might just prefer a human driver, and e.g. enjoy the talk and banter and getting to know other people. Beyond just preferences alone, judging by the increasing wait times in the app today, Waymo is supply constrained. There are not enough cars to meet the demand. A part of this may be that Waymo is being very careful to manage and monitor risk and public opinion. Another part is that Waymo, I believe (?), has a quota of how many cars they are allowed to have deployed on the streets, coming from regulators. Another rate-limiter is that Waymos can’t just replace all the Ubers right away in a snap of a finger. They have to build out the infrastructure, build the cars, scale their operations. I posit that all kinds of automations in other sectors of the economy will look identical - some people/companies will use them immediately, but a lot of people 1) won’t know about them, 2) if they do, won’t trust them, 3) if they did, they still prefer to employ and work with a human. But on top of that, demand is greater than supply and AGI would be constrained in exactly all of these ways, for exactly all of the same reasons - some amount of self-restraint from the developers, some amount of regulation, and some amount of simple, straight-up resource shortage, e.g. needing to build out more GPU datacenters.

the globalization of full automation

As I already hinted on with resource constraints, the full globalization of this technology is still very expensive, work-intensive, and rate-limiting. Today, Waymo can only drive in San Francisco and Phoenix, but the approach itself is fairly general and scalable, so the company might e.g. soon expand to LA, Austin or

etc. The product may also still be constrained by other environmental factors, e.g. driving in heavy snow. And in some rare cases, it might even need rescue from a human operator. The expansion of capability does not come “for free”. For example, Waymo has to expend resources to enter a new city. They have to establish a presence, map the streets, adjust the perception and planner/controller to some unique situations, or to local rules or regulations specific to that area. In our working analogy, many jobs may have full autonomy only in some settings or conditions, and expanding the coverage will require work and effort. In both cases, the approach itself is general and scalable and the frontier will expand, but can only do so over time.

society reacts

Another aspect that I find fascinating about the ongoing introduction of self-driving to society is that just a few years ago, there was a ton of commentary and FUD everywhere about oh “will it”, “won’t it” work, is it even possible or not, and it was a whole thing. And now self-driving is actually here. Not as a research prototype but as a product - I can exchange money for fully automated transportation. In its present operating range, the industry has reached full autonomy. And yet, overall it’s almost like no one cares. Most people I talk to (even in tech!) don’t even know that this happened. When your Waymo is driving through the streets of SF, you’ll see many people look at it as an oddity. First they are surprised and stare. Then they seem to move on with their lives. When full autonomy gets introduced in other industries, maybe the world doesn’t just blow in a storm. The majority of people may not even realize it at first. When they do, they might stare and then shrug, in a way that ranges anywhere from denial to acceptance. Some people get really upset about it, and do the equivalent of putting cones on Waymos in protest, whatever the equivalent of that may be. Of course, we’ve come nowhere close to seeing this aspect fully play out just yet, but when it does I expect it to be broadly predictive.

economic impact

Let’s turn to jobs. Certainly, and visibly, Waymo has deleted the driver of the car. But it has also created a lot of other jobs that were not there before and are a lot less visible - the human labeler helping to collect training data for neural networks, the support agent who remotely connects to the vehicles that run into any trouble, the people building and maintaining the car fleet, the maps, etc. An entire new industry of various sensors and related infrastructure is created to assemble these highly-instrumented, high-tech cars in the first place. In the same way with work more generally, many jobs will change, some jobs will disappear, but many new jobs will appear, too. It is a lot more a refactoring of work instead of direct deletion, even if that deletion is the most prominent part. It’s hard to argue that the overall numbers won’t trend down at some point and over time, but this happens significantly slower than a person naively looking at the situation might think.

competitive landscape

The final aspect I’d like to consider is the competitive landscape. A few years ago there were many, many self-driving car companies. Today, in recognition of the difficulty of this problem (which I think is only *just barely* possible to automate given the current state of the art in AI and computing more generally), the ecosystem has significantly consolidated and Waymo has reached the first feature-complete demonstration

of the self-driving future. However, a number of companies are in pursuit, including e.g. Cruise, Zoox, and of course, my personal favorite :), Tesla. A brief note here given my specific history and involved with this space. As I see it, the ultimate goal of the self-driving industry is to achieve full autonomy globally. Waymo has taken the strategy of first going for autonomy and then scaling globally, while Tesla has taken the strategy of first going globally and then scaling autonomy. Today, I am a happy customer of the products of both companies and, personally, I cheer for the technology overall first. However, one company has a lot of primarily software work remaining while the other has a lot of primarily hardware work remaining. I have my bets for which one goes faster. All that said, in the same way, many other sectors of the economy may go through a time of rapid growth and expansion (think ~2015 era of self-driving), but if the analogy holds, only to later consolidate into a small few companies battling it out. And in the midst of it all, there will be a lot of actively used Tool AIs (think: today's Level 2 ADAS features), and even some open platforms (think: Comma).

AGI

So these are the broad strokes of what I think AGI will look like. Now just copy paste this across the economy in your mind, happening at different rates, and with all kinds of difficult to predict interactions and second order effects. I don't expect it to hold perfectly, but I expect it to be a useful model to have in mind and to draw on. On a kind of memetic spectrum, it looks a lot less like a recursively self-improving superintelligence that escapes our control into cyberspace to manufacture deadly pathogens or nanobots that turn the galaxy into gray goo. And it looks a lot more like self-driving, the part of our economy that is currently speed-running the development of a major, society-altering automation. It has a gradual progression, it has the society as an observer and a participant, and its expansion is rate-limited in a large variety of ways, including regulation and resources of an educated human workforce, information, material, and energy. The world doesn't explode, it adapts, changes and refactors. In self-driving specifically, the automation of transportation will make it a lot safer, cities will become a lot less smoggy and congested, and parking lots and parked cars will disappear from the sides of our roads to make more space for people. I personally very much look forward to what all the equivalents of that might be with AGI.