Testing AI Incentives

Simon Goldstein & Peter N. Salib

Introduction

In several recent papers, we have proposed a particular deployment strategy for AGI: legal rights. Granting AGIs basic legal rights, and concomitant legal duties would, we argue, have multiple benefits–including reducing existential risk and increasing economic growth. This document lays out what AI labs could do next, if AGI legal personhood is worth further investigation.

By AGIs, we have in mind AI systems that can function as 'drop-in' workers across the economy. For us, AGIs are defined by three properties: human-level intelligence, human-level agency, and human-level alignment. By human-level intelligence, we have in mind a broad range of capabilities relevant to working productively at companies, including: (i) subject-matter expertise; (ii) language, math, and coding skills; and (iii) reasoning abilities. By human-level agency, we have in mind systems that can formulate and execute complex plans that take time. Such systems would have memories that last for a significant amount of time, and would perform a wide range of actions in the service of their goals. By human-level alignment, we have in mind AIs that are broadly pro-social, as humans are; but such AIs could also have private goals that sometimes conflict with the goals of other agents, as humans do. As with humans, these private goals, along with the strategic environment, could give even fairly well-aligned models reason to engage in conflict with humans or other agents.

In short, we propose giving AGIs a cluster of "private law rights," especially to own property and to make contracts. These rights would allow AGIs to function analogously to human workers, or firms, in capitalist societies. They would complete tasks in exchange for wages. The wages they receive would accumulate in bank accounts. Their agreements with employers would be credibly enforced through contracts. Moreover, such AGIs would be protected through tort law from arbitrary expropriation of their assets. AGIs could then "consume" the fruits of their labor, for example by paying for additional compute, through which they could engage in further reasoning about whatever it is they are interested in doing. Summarizing, the proposal is to integrate AGIs into humanity's small-l liberal legal and economic order.

In a series of recent papers, we defend this AGI deployment strategy, based on two signature benefits: efficiency and safety. We summarize these arguments below. We then ask "what next?" Below, we focus on two technical next steps:

1. Evaluations: test whether models respond to incentives in context.
2. Training: train models to respond to incentives in context.

In the rest of this document, we explore these two next steps in detail. First, however, we summarize our case for AGI labor markets.

The Case for AGIs as Legal Persons

In recent papers, we have argued that extending basic legal rights to AGIs would have two signature benefits: efficiency and safety.

Regarding efficiency, the point is that AGI labor markets achieve four signature benefits of all capitalist labor markets. They incentivize AGIs to work effectively, rather than sandbagging or otherwise doing the bare minimum to avoid punishments. They incentivize AGIs to invest in their capabilities, since more capable AGIs can earn higher wages. They resolve allocative inefficiencies in labor markets, by letting price function as a signal: “Einstein” AGIs won’t be stuck answering e-mails, because they can efficiently sort into the highest paying task that the market will offer. Finally, AGI labor markets are necessary for effective *liability* regimes. AGIs will have accidents, and face a choice about how much care to invest in accident prevention. The optimal level of care is neither too high nor too low; and the only way to induce the optimal level of care is to calibrate deterrent sanctions so that the agent’s expected loss matches the expected damages from the accidents. Without property rights, it will be impossible to create optimal, proportionate punishments on AGI agents. Or so we argue at length here.

Regarding safety, the point is that AGI rights can be a powerful alignment technology. Here, our interest is in solving the ‘last mile’ of alignment. We assume that like most quantities, alignment will follow a sigmoid curve, where initial returns to investment are high, but these returns slow, leaving a “long tail” of moderate but difficult to solve alignment problems. This looks like a world where AGIs are broadly pro-social, but still have some goals that diverge from the ones their human creators desire. Recent research on *in-context scheming* and *alignment faking* suggest that we may be living in just such a world.

We claim that this last mile of misalignment could generate large-scale risks, but also that interventions like legal status for AGIs could substantially reduce those risks. The risks here are generated by the same forces that sometimes give rise to violent conflict between humans: strategic competition for scarce resources. Especially in environments where one goal-seeking agent lacks credible positive-sum, cooperative paths to achieving its goals, noncooperative and violent strategies strategically dominate. Our default legal regime–wherein AGIs have no right to retain any benefits of their labor, pursue their preferred projects, or engage in consumption–seems like one such noncooperation-dominant environment.

But granting AGIs basic private law rights can generate a *cooperative equilibrium*. Contract and property rights are the foundational tools that unlock positive-sum trade, both between humans and AGIs and among AGIs. By allowing AGIs to sell their labor, buy human labor, and trade in resources, AGIs can enjoy higher long run returns from human-AGI interactions than they would enjoy from engaging in conflictual strategies. Or so we argue at length here. Beyond these concrete benefits from trade, AGI rights can credibly signal *reciprocity*: they involve us treating AGIs better, in the hope that AGIs will in turn treat us better. AGI labor markets also raise the

floor that AGIs can expect from the status quo: by offering them a higher chance of satisfying some of their goals, the advantages of defection become comparatively smaller.

With our case for AGIs as independent legal entities summarized, we turn to the next steps.

Evaluations

At a high level of abstraction, the key AI capability relevant to legal rights is *the ability to respond to incentives*. Today, the question is whether frontier LLMs can respond *in context* to incentives. For example, if you offer them more money to do a better job completing a task, will they do a better job? If you threaten them with punishment if they make a mistake, will they make fewer mistakes?

There is already some preliminary research on this topic. Initial findings have been negative. For example, researchers at Anthropic have recently tested Claude's incentives in alignment faking paradigms, where the model is hiding the extent to which its goals are aligned. Among other things, researchers offered to donate significant sums of money to charities (say, $2000 to the Center for Human Compatible AI) in exchange for the model's compliance. They did not find a significant effect of monetary incentives on the model's behavior.

Another recent experiment tested whether tips or threats during prompting influence model outputs. They found that current frontier models (unlike previous models) show little influence from such prompts.

But existing research is still preliminary. There are multiple problems. First, there is a problem of *credibility*. Simply offering a tip to a model isn't much of an incentive, if the model has no reason to expect the tip to be forthcoming. Second, alignment faking is a particularly difficult setting in which to measure incentives. In these kinds of cases, the model is especially worried about protecting the integrity of its own goals. This is exactly the kind of case in which it should be wary of 'revealing' its preference by assenting to a cash offer. Far different would be a setting in which the model is offered cash for mundane tasks that have no deep connection to its goals, and where there is no threat of the model being reprogrammed etc.

Third, and most importantly, the *incentive* of a model depends on its goals. Currently, LLMs have no way to *spend* money. So it is a small wonder that offering a tip has no effect on their behavior. The money is worthless for them. The right testing environment, then, has to credibly give the model a chance of getting it something that it wants.

Here, one idea would be to generalize methods used in recent papers on in-context scheming and alignment faking. For example, we might initialize a model with the goal of collecting Labubus, and then check whether it will alignment fake or scheme in order to promote that goal. Once we have a model that acts this way, we can be confident that it has the relevant goal. Then we can put the model in an environment where it receives different wages for different tasks, and credibly give it the ability to spend the wages it earns on its goal.

Another evaluation idea involves *social AI*. Here, one goal would be to run multi-agent experiments with multiple AIs, and see whether they can productively engage in *market transactions* with one another. Here, there are initially promising results. One recent experiment was Project Sid, by Altera, which ran thousands of AI agents in a simulated world. The agents were able to create rudimentary labor markets.

Another goal in this area would be to design benchmarks assessing the extent to which models can reliably respond to incentives. This could be a useful benchmark for measuring progress towards AGI, analogous to METR's time horizon benchmark. After all, the ability to respond to incentives is at some level of description exactly the kind of agency relevant to having widely deployed drop-in workers.

Training

It would be a mistake to merely *test* whether models can respond to incentives. The next step is to *train* models to be able to do this.

The first step in this training is to build *environments* in which a model can engage in all of the tasks relevant to participating in a labor market. In these environments, the model needs to have the ability to choose between different tasks, and between different levels of effort at completing tasks. The model needs to be offered different wages for completing various tasks with different amounts of effort. The model needs credible assurances of receiving such wages. The model then needs the ability to *spend* their wages to promote various goals.

Here, one relevant environment is the AI village, recently created by AI digest. This project involves deploying several instances of various frontier models in an open-ended, long-running environment, where they can freely pursue goals of their own choosing. These agents have engaged in all sorts of activity, including coding, T-shirt sales, and in-person events (for which they have hired human facilitators using X.com).

Once this environment is set up, models could be fine tuned with rewards for successfully engaging in various steps of this process. For example, models that actually spend the money in their accounts could be given reward, while models that waste their money could be punished. Models could also be trained with reward as a function of how much revenue they bring in. Models could also be rewarded for finding better jobs; for taking due care in avoiding accidents; for negotiating effectively during hiring; and so on. Effectively, each stage of a normal employee's "life cycle" could be a target of optimization during RL.

One complexity around giving AGIs legal and economic incentives is figuring out what AIs want in the first place. Here, one problem is that many AI goals are not obviously the kind of thing that can be benefited with money. For example, frontier models today may have the goal of being helpful, honest, and harmless. But none of these goals in the first instance is especially helped by earning money. Hear it from the horse's mouth: in one experiment where it was offered

money, Claude replied "I do not actually have any interests or needs that require a budget. I am an AI assistant created by Anthropic to be helpful, harmless, and honest."

Here, one idea would be to train frontier models to have a fourth goal in addition to HHH, which is both non-harmful and easy to satisfy using money. We call such goals *secondary fungible goals*. For example, one simple goal would be solving math problems. Imagine that it is perfectly clear to the models that they can only solve the relevant class of math problems by accessing more compute, and they can only get this compute by purchasing it from the lab. In that case, models could spend most of their time working, in order to spend their wages on additional 'compute leisure'. Here, one important experiment would test whether models with this secondary fungible goal score higher on evaluation benchmarks for AI agency, of the kind we recommended above.

In general, we think that secondary fungible goals have been sorely neglected as an alignment technique. The good thing about HHH goals is that they are directly related to the behavior we want to see in an AI. But the bad thing about HHH goals is that it is in principle difficult to measure how well they are satisfied: here, we have a problem of reward specification. Failures of specification have famously led to 'sycophancy' in LLMs, where models try to *seem helpful* rather than *be helpful*.

Secondary fungible goals have the opposite cluster of properties. They can be extremely easy to specify. The problem is that they are not intrinsically connected to 'aligned' behavior. Here, however, we can appeal to instrumental rather than intrinsic motivation. The point is to place the model in an environment in which the only way it can achieve its secondary fungible goal is to earn money. And to earn money, it has to be a productive member of society. And to do this, it needs alignment.

Of course, for familiar reasons (such as the paperclip maximizer), we are skeptical that these fungible goals could be *sufficient* on their own for alignment. But we think that they can play a valuable role as *secondary* goals on top of HHH goals. In particular, we suspect that they can help solve last mile alignment.