THE AI ENGINEERS' GUILD

Proposal for an AI risk mitigation strategy

by Morris Stuttard and Anastasia Slabukho

CONTENTS:

ABSTRACT

Taking into account the incentives for self gain in a global market economy, any consultation-dependent effort to mitigate the risks of the AI race runs the risk of effecting very little actual change. This paper proposes a more proactive form of intervention modelled on writers' guilds, which have managed to successfully regulate development activities in a $324 billion Film and TV industry. By means of combining extrinsic and intrinsic incentives, it might be possible to establish an equivalent AI Engineers' Guild and use it as a platform for ensuring safety in global AI development and even application. A call for discussion is issued on what form an AI Engineers' Guild might take as a risk mitigation tool and how a Guild could be created where no such entity currently exists.

LONG ABSTRACT

This paper proposes that efforts to intrinsically promote safety and openness in AI development face a long and hard struggle ahead due to the incentives of capital gain and national competition at the heart of the AI race. Jared Diamond's *Collapse* offers historical warning of how immovable past societies have proven to be even when the direst of consequences were already unfolding. It is posited that central to our social tendency to ignore self destructive activity is the desire for 'exclusive advantage' - where preservation of the 'self' takes priority over preservation of the wider world. Companies and governments in our current economic and political situation mostly follow the same drive. Indeed, the very fact 'incentives' have become central to the discussion on AI Risk mitigation strategies is a *de facto* acceptance that rational thinking on this question can not be depended upon - we need to be motivated to save ourselves too.

Rather than viewing this situation as negative, it is proposed that efforts towards AI safety should adopt a competitive rather than consultative attitude. Once we accept that AI developing firms are not motivated to invest heavily enough in risk mitigation, we will be in a much better position to address the issue by taking on more direct approaches.

But neither does this paper recommend the extrinsic incentives of government regulation. Since governments have their own stake in the AI race, any effort to lobby them for greater caution is also likely to be met with disappointing results, as already seen in efforts to address the global climate change problem. Resistance to change will be redoubled because, even more so than climate change, the threats posed by unsafe AI development or monopolisation of AGI are difficult to define, both in timescale and nature.

Instead, the solution proposed here is that AI engineers and researchers harness their own potential to create a positive and safe culture of AI design and implementation. Parallels are drawn with the screenwriting vocation of the Film & TV industry, where scriptwriters have successfully exploited their creative importance to their employers, uniting under powerful guilds who to this day hold great sway on the industry's activities. Their influence is such that major studios are themselves signatories of the Writers' Guild of America and are compelled to abide by the rules laid down by the guild when working with writers.

The recent boom in salaries of AI engineers suggests that these senior individuals have a similar value in the AI race. This paper recommends an immediate and ambitious effort to form an AI Engineers' Guild around these luminaries - one which would nurture safe practise through employment contracts and compel employers to become guild signatories to the same end.

Incentivisation would be key to the success of a new AI Engineers' Guild - both for members as well as for signatory companies and governments. Initial incentives should not rely upon the intrinsic desire to promote safe AI but incentivise through the promise of capital gain. On top of regular guild initiatives such as promoting fair pay and equal opportunity, an AI Engineer's Guild would guarantee valuable legal indemnification to both members and signatory companies in the event of AI system failure; hold a monopoly on all top level AI researcher talent adhering to safe AI development policies; and offer dynamically-priced Guild contracts to allow non-profits, SMEs and low GDP governments access to advanced AI systems.

The paper ends with a call for discussion on what form such a Guild would take in the long term to prevent internal abuse of power and achieve its 'safe AI' goals even at the level of AGI discovery. An additional suggestion is given to create a Guild 'spec' market of AI systems and, following that, an in-house AGI research program in order to draw, fund and keep the best AI research talent, while ensuring the safe development, sale and application of AI systems to the global tech industry.

OBJECTIVES

Primary objective: find a solution or set of solutions for mitigating the risks associated with the AI race.

Secondary objectives: create discussion around the topic in order to gain a better understanding of the nature of the AI race, raise awareness of the race, and to get as diverse an idea pool as possible.[1]

THE PROBLEM

EXCLUSIVE ADVANTAGE

Any effort to influence the course of technological advancement must begin with an awareness of the primary incentives that drive it. Without this understanding, solutions which appear sound in theory are unlikely to be embraced on the scale required to effect any real change.

When seeking to understand these incentives, there is every reason to follow the money. In a global market economy, certain dynamics can be counted upon. Economic self interest is one of them.

Popular imagination would have us believe that the driving force behind technological invention and innovation is the human instinct for discovery. A slightly more grounded but still not entirely useful view is the cliché that ‘necessity is the mother of invention’ - and presumably of its younger sibling, innovation, too. But the view most substantiated by research[2] is that, within a market economy, technological discovery is driven primarily by the desire for capital self gain.

It is useful to remember that interests of capital self gain are not limited to the ‘individual self’. It is true that many independent specialists contribute to the technological march of our species, but even such luminaries rely upon financially incentivised support from wider groups of investors to bring their discoveries to the market and the world. This leads us to corporate entities, which use their financial muscle to both absorb pioneering startups and drive major innovations in-house (more so innovation than invention due to the higher risk of failure that comes with moving development away from existing technologies). Corporations are primary drivers of technological innovation today, not only because of their R&D budgets but because they are so well placed to introduce new products to society through highly-evolved marketing and logistics infrastructures. The ‘self’ in ‘capital self gain’ must therefore be extended to the conglomerate.

---

[1] https://www.general-ai-challenge.org/ai-race

[2] Huesemann, Michael H., and Joyce A. Huesemann (2011). *Technofix: Why Technology Won’t Save Us or the Environment*, Chapter 11, "Profit Motive: The Main Driver of Technological Development", New Society Publishers, Gabriola Island, Canada, ISBN 0865717044

But also to nations. Governments are more complex beasts, since their agenda is more dispersed - and not at all limited to fiscal return. Nonetheless, even when the return on government technological advancement is *not* financial in nature, there still exists a 'self' with regard to incentives. Governments are motivated, indeed mostly required, to invest public funds into the service of their own nation's people over the world's. The Space Race is an obvious example of this.

For our purposes, it would be extremely useful to define a single primary incentive driving technological development at the personal, corporate and governmental level, if such a single incentive exists. In the event the desire for capital self gain does not paint the whole picture, it might be argued that 'advantage' does. Would it be so erroneous to claim that most technological advancement of any kind, including the AI race, is overwhelmingly driven by the desire for exclusive advantage - more explicitly, the advantage of one person, company or country over every other?

INCLUSIVE VOICES

If one accepts that technological advancement is driven primarily by exclusive advantage, it could be said that the passenger seats are occupied by the highly vocal but oft-ignored voices of collective advantage.

What defines the voices from the back seat, and makes them extremely important, is that they are *in*clusive in nature. Ethics, equality, environmental protection - whatever these voices are calling for, they are singular in their service of the world over the 'self'. They are also characterised by long term over short term perspectives. These voices are ours when we call for protection from the dangers of unchecked AGI/ALI research.

However, if past warnings on the subject of climate change are any indicator, our voices will go largely unheard. There are no voting systems for technological development, so, much like children's calls from the back seat of "Are we there yet?", it might be optimistic to believe that simply by raising awareness of the risks of the AI race we will have any significant impact on its course.

THE LONG TERM LONGSHOT

The anthropocene era owes much of its wonders, waste and weirdnesses to the commercialisation of technological inventions and innovations. Outside military tech, one would have a very difficult time finding rational strategies behind the introduction of many of the technologies ubiquitous in our society today.

This only supports the view that those wielding the tools of technological change in our global economy are much more incentivised to create and release new products and services for their own gain than they are to guide the wider world to a better place.

It would be reassuring if we were at least more conscious of this dynamic, but the drivers of tech development appear not only to be ignoring the voices from the back, but also in possession of no real roadmap. In a market economy, drivers of technological advancement are not even incentivised to take the prudence and contemplation that comes with time. Thanks to the institution of patenting, a simple dynamic lies at the heart of all commercially valuable tech development - the first new product or service on the market wins.

Compounding this incentive for haste is the fact that profit only holds value for as long as it confers wealth upon those still alive to enjoy it, which means capitalist driven technological innovation has no reward mechanism in place for long-term pioneering. A cynical view would be that safety concerns are a factor in technological development only for as long as there is a danger to profit.

Such short-sighted market incentives can only lead to the neglect of dangers and drawbacks of new technologies in the long term. While governments do recognise their responsibility to think beyond the immediate and mitigate the harm that modern day tech development does alongside the good - be it to public health or the environment - actual high-impact action towards *safe* and *sustainable* global development of any kind is rare. The US planned withdrawal from the Paris Agreement is a prime example of how exclusive advantage so easily trumps sustainable development at the government level.

Even domestic government safety mechanisms suffer from the same incentivisation towards acceleration rather than careful advancement. EU legislation tends to view the tech sector as benign in relation to others such as the chemical sector, and economists advocate the removal rather than tightening of tech sector regulation in order to better compete with the US and other markets[3]. This makes adequate legislative protection unlikely where long term dangers of innovations are concerned, nevermind those threats which are difficult to predict, such as the invention of AGI.

HISTORICAL WRECKS ON THE SIDE OF THE ROAD

The solution one would think, lies in raising awareness of the risks of the AI Race and accompany this effort with the provision of actionable solutions to mitigate these risks, as with the Asilomar AI Principles[4]. There is no doubt that a consultative approach will have some positive impact but serious consideration needs to be given as to whether it will be enough.

[3] Jacques Pelkmans and Andrea Renda; Nov. 2014: *Does EU regulation hinder or stimulate innovation?* CEPS Special Report.
[4] https://futureoflife.org/ai-principles/

Broader history offers its own warnings. Jared Diamond argues in his book, *Collapse,*[5] that our current global society is driving blindly towards multiple ecological cliffs. Civilizations throughout history have demonstrated a chilling habit for ignoring long term dangers to the point of complete societal collapse. Multiple case studies are presented - the Norse and Inuit of Greenland, the Maya, the Anasazi, the indigenous people of Rapa Nui (Easter Island), Japan, Haiti, the Dominican Republic, and modern Montana. Even when the cliff is clearly seen up ahead, our species has a tendency to just keep on driving. Our failure to manage global climate change is a worrying example of this dynamic unfolding globally today.

Like the agricultural and industrial revolutions that brought us to our current stage in technological evolution, it is widely thought that we are on the brink of a new threshold in the development of our species[6] - an era of high-impact ALI or even AGI-administered systems. Of all the dangers we are driving towards, this has only even been conceived as a threat in the last few decades and is still by far the cliff most shrouded in fog. It is unclear when we will hit this precipice, or how far we will fall. We only have the voices from the back warning us that a danger is there - be it corporate or national hegemony, acute global inequality, or, if science fiction writers prove prescient, enslavement or destruction at the hands of an AGI system.

We call for safeguards to be put in place but the question remains whether the drivers of AI development - the engineers, investors, companies and governments - will heed these voices in the back when not only a cliff but the light of exclusive advantage shines from up ahead.

[5] Jared Diamond 2005: *Collapse: How Societies Choose to Fail or Succeed* (ISBN 978-0241958681).
[6] C Last. Foundations of Science 22 (1), 39-124 *Big Historical Foundations for Deep Future Speculations: Cosmic Evolution, Atechnogenesis, and Technocultural Civilization*

A SOLUTION IN PRINCIPLE

THE INCENTIVE GAME

AI development is so far unfolding largely as one would expect. With the exception of companies like GoodAI and OpenAI, investment has drawn the best engineers to projects serving corporate or national agendas, meaning AI research and application is already being funneled down commercial and military corridors. The very fact 'necessary incentives' have become central to AI Risk mitigation debates is a *de facto* acceptance that rational thinking alone cannot be depended upon - we need to be motivated to save ourselves too.

EXCLUSIVE TO INCLUSIVE INCENTIVES

How might AI development actors be convinced to introduce risk-mitigation measures into their development work?

The strategy taken on by Musk's OpenAI is to issue a call to arms from AI engineers to create safe AGI available to everyone. OpenAI's Charter[7] is noble and well-thought out. It does, however, rely heavily on the firm winning the AGI race (its commitment to dropping tools and joining any other actor who comes close to discovering AGI depends heavily on that external actor even wanting researchers involved who do not share exclusive advantage as their goal).

Winning the AI race will not be easy. Open AI's horse in the race is burdened by its prioritization of safe AGI development. The firm must also continue to attract and retain the world's leading AGI researchers. Despite their lead researcher earning up to $1.9 million in 2016, the inability of the non-profit to offer stock options to employees - one of the primary lures used by private companies - means OpenAI staff are still being underpaid by industry standards. The result is that Open AI is dependent on talent being altruistic enough to put material wealth second to humanity's welfare. In 2017, five of their researchers left for the private sector.[8]

But is it possible that our social tendency towards exclusive advantage is where the solution lies? Instead of addressing the issue of AI development safety with a call for safety itself - and thereby running the risk of joining the other ignored voices in the back - might it be possible to create market-competing financial incentives that reward the enforcement of safe standards?

Whether or not the following proposal is deemed feasible itself, there is every reason to believe that some solution exists along this line of thinking.

[7] https://blog.openai.com/openai-charter/
[8] Cade Metz, 19 April 2018, New York Times, *A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit*

A SOLUTION IN PRACTICE

INSPIRATION - WRITERS' GUILDS

It should not be necessary to reinvent the wheel. The purpose of this paper is to draw attention to one particular institution founded in multiple countries around the world within the $324 billion Film & TV industry. That of writers' guilds.

For all the many differences, enough parallels exist between screenplay development in the Film & TV industry and AI development in the tech industry to make a guild system worthy of consideration as a mitigation solution to the risks of the AI race[9].

Before financing or production of a film or TV series, a painstaking process of script development must be successfully completed. A filmic story needs to be conceived, shaped and revised before a screenplay is drafted, revised and then revised some more. Similarly to AI engineers working *independently*, this work can be carried out by the writer/s alone - working without funding but retaining ownership of the intellectual property in what is known as a 'spec script' - which is then sold to the market. Similarly to AI engineers working as *employers* for companies (or governments), writers can also be hired by producers, studios or networks to develop projects that are owned by the hiring companies (such as Marvel Studios, for example). Only when the script development process, in either form, converts into what is perceived to be a fully-functioning screenplay - a blueprint for the filmmaking process to come - is the project produced. Once production and post production is complete, the resulting work is then distributed to the world. In this sense, the input of screenwriters represents the creative foundation of all output from the Film & TV industry.

How far can we take this comparison? There may be value in thinking in terms of 'talent' and what role senior talent means to the wider industry. If the work of highly-skilled writers represents the creative foundation of all products released by the media industry, is not the work of highly-skilled AI engineers the foundation of related products and systems released by the tech industry? Instead of films and TV shows, there are self-drive cars, data analytics services and emergency response systems - but whatever the finished product, in many senses it has been built around the work of highly-skilled AI researchers and engineers.

Screenwriting is an unusually top heavy profession - only 5,227 WGAW screenwriters reported earnings in the year ending March 2017.[10] A 2017 study by Element AI showed that, *globally*, only 22,000 PhD-educated researchers were working on AI systems.[11] Further supporting this parallel, the high salaries of leading AI researchers and engineers suggests the tech industry views the AI developing talent pool similarly to how the Film & TV industry perceives leading

[9] The following description of activities have been simplified and generalised for the purposes of this paper.
[10] http://www.wga.org/uploadedfiles/the-guild/annual-report/annualreport17.pdf
[11] https://www.bloomberg.com/news/articles/2018-02-07/just-how-shallow-is-the-artificial-intelligence-talent-pool

screenwriters. Big-name AI researchers attract finance and supporting talent[12] in exactly the same way big-name writers do, and both attract yearly incomes of $500,000 plus.

How might this comparison be useful? Because, as individuals, writers have very little influence on the industry they serve - they are simply employees incentivised to write what can feasibly be produced. They, like the Film & TV industry itself, are driven by the goal of exclusive advantage - self-marketing their way through the crowds of competition to make a living in a highly competitive industry. AI development is arguably much the same - only, thanks to the risks of AI development - with much higher stakes. With the potential rewards of AI breakthroughs in military applications, it is only a matter of time before state defence AI engineering jobs offer competing salaries (there are indications that they currently do not, at least in the US[13]). When that happens, if it has not already, top AI engineering talent will be contributing to the creative foundation of the defence industry's output too.

Among the many obvious industry-related differences between modern day screenwriters and AI engineers, one important institutional distinction stands out. Screenwriters are unified under powerful guilds. As of yet, AI engineers have no such institution representing them or their work. There exists a guild of AI Game Programmers but no guild where it is most needed - in general purpose AI R&D. The intention of this paper is to show how the establishment of a globally functioning guild of AI engineers offers a potential platform for the introduction of much-needed safety mechanisms in the field of ALI and, particularly, AGI development.

HOW GUILDS WORK AND WHY

When united under a guild, writers are able to collectively influence the industry they serve, rather than representing silent employees. The guild conveys commercial value both to the writers it represents and to the market it serves, using that muscle to challenge 'exclusive advantage' incentives in order to promote values of collective betterment. For example, guilds have made significant achievements in addressing social issues such as unequal representation/pay, racism, ageism, harassment, financial crisis and mental health. Were there a single cliff that writers were driving humanity towards through their work, as there may be with the work of AI engineers, there is little doubt that writers' guilds would exert both the will and influence to divert the industry away from that precipice.

The success of writers' guilds, especially the powerhouse guilds in the US, the WGA West and WGA East, owes much to the status of their membership roster. It is mandatory for a screenwriter to join the WGA upon winning a contract with a studio and, in this way, the guild oversees the activities of almost all top-earning screenwriters. This industry clout of the WGA

[12] Cade Metz, 19 April 2018, New York Times, *A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit*
[13] Cade Metz, 15 March 2018, New York Times *Pentagon wants Silicon Valley's Help on A.I.*

was built over many years through reciprocal incentives - on one hand for writers to become members and, on the other, for the Film & TV industry to employ guild writers and thereby bow to the rules the guild imposes on employment.

If an AI Engineers' Guild is ever to be established, it could borrow a good deal from the success of writers' guilds. The following are some incentives that reciprocally bind writers' guilds to the industry they serve, taken from the WGA and the British WGGB.

INCENTIVES FOR WRITERS TO JOIN A GUILD

- Status. Only the top professionals in the field possess guild membership, meaning it acts as a form of industrial accreditation and attracts more work.
- Fair pay. Guild writers are required only to accept work that is paid in line with minimums agreed with studios and producers' guilds - and that means paid well. This agreement sustains the commercial value of writers' contributions to the Film & TV industry while also shielding individuals from commercially-driven exploitation.
- Protection. Writers' guilds provide protection in the form of legal advice, contracts, and security of intellectual property.
- Fairness. Positive social values are upheld by guilds, such as ethnic diversity and gender equality.
- Professional development - most guilds promote and subsidize specialist training for their members to keep their skills up to date with contemporary industry requirements and ensure they are trained in non-industry-specific skills such as self-marketing, negotiation and financial planning.
- Awards. Writers' guilds recognise works of exemplary accomplishment through annual award ceremonies.
- Legislative influence. The guilds play a significant role in the formation of new legislation pertaining to the arts - be it through consultation or lobbying.
- Pension schemes and emergency financial support. Long term financial security is promoted by the Guild.

INCENTIVES FOR COMPANIES TO EMPLOY GUILD WRITERS

- Market dominance. America's WGA has grown so strong that the Hollywood studios themselves are signatories of the guild, meaning any studio involvement, be in production or distribution, requires the writer to join the WGA. The result is that all studio-produced screenwriters are current, post-current or emeritus guild members.
- Quality assurance. Even in guilds outside the US, the lofty entry level requirements for membership act as a guarantee that a writer has the experience, knowledge and skills to deliver a commercially viable product. For this reason, a guild writer brings significant value to the financing stage of new projects.
- Assurance of professionalism. A guild writer adheres to a code of conduct or 'code of working'. This acts as a behavioural guarantee to the employer.

- Services. The guild provides services that can be of use to production companies, networks and studios such as credit arbitration, contract litigation and intellectual property protection.

CHALLENGES OF USING A GUILD SYSTEM FOR RISK MITIGATION IN AI DEVELOPMENT

Would a guild of AI engineers offer a feasible solution to our goal of mitigating the risks associated with the AI race?

One thing can be certain - many challenges would lie ahead.

The Writers' Guild of America did not wake up one day with the influence it currently holds over the Film & TV industry. Its authority evolved over decades of hard work. And still, in 2018, the WGA struggles against the forces of commercial self-interest. Currently, the WGA is embroiled in one of the most bitter feuds[14] since the writers strikes of 2007/8, this time with the talent agencies, following the monopolization and business diversification of the two powerhouse agencies, WME and CAA. If the Hollywood agencies do not concede to the WGA's demands, it is conceivable that the WGA's influence over the industry could diminish considerably. The forces of exclusive advantage are not only strong but unabating.

In addition to general challenges of any guild in any industry, there are many that pertain specifically to AI R&D, especially when the goal of an AI Engineers' Guild is to secure safe and available AI development across the globe.

It is not the purpose of this paper to analyze these incongruences in detail, but it is important to be cognizant of some of the unique qualities of AI development which would make a guild system complicated as a risk mitigation tool. Among these qualities are:

- **The AI Arms Race.** Military application of AI is the sphere where risk mitigation is most needed[15] and yet the defence industry is heavily protected from external influence and regulation.
- **Diversity of research.** The diversity of AI development is enormous - both in its theoretical framework and in its application. Codes of conduct or safety measures that might usefully mitigate risks in one field of innovation might not be applicable to another.
- **Rapid change.** The face of AI development is changing with extraordinary speed. With every advancement, new opportunities and new dangers open up. Any institutional effort to mitigate risks can quickly become at best only partially effective, at worst wholly redundant.

[14] David McNary April 24, 2018, Variety. *Hollywood Agents Slam Writers Guild Over Proposed Rules Revamp*
[15] Brian Tomasik, 2013-2016, Foundational Research Institute, *International Cooperation vs AI Arms Race*

- **Employer loyalty.** For almost a century, screenwriters have been freelancers employed on a project basis - even ‘staff’ jobs are limited to the lifespan of a TV series. This makes screenwriters far better disposed to seek guild membership and abide by its rules than it would AI engineers. Permanent employee contracts, as is commonplace in AI development, undermine the influence a guild can have over a member’s work and the employer’s use of that work.
- **Dangerous use of safe work.** Even if an AI engineer completes a project for an employer with no transgression of the guild’s AI development rules, that work might still be applied in ways that are dangerous, or combined with the work from other engineers in a way that would be deemed unsafe.
- **The corrupting value of AI.** There is significant danger that any guild member or group of members on the verge of a world-beating discovery in their AI work would be tempted to turn their back on the guild and pursue exclusive advantage.

THE AI ENGINEER’S GUILD - PROPOSAL FOR AN AI RISK MITIGATION SOLUTION

Before we turn our attention to the viability of building a guild system for AI development where no such institution currently exists, let us first analyze whether the guild model even represents a solution at all for the mitigation of risks associated with the AI race.

For evaluation purposes, let us imagine for a moment that a functioning guild of AI engineers is already in place across the globe - what form would this guild most perfectly take and how might it act as a mitigative force against the risks of AI development and application?

Without a doubt, the following ideas will require refinement, change and even complete elimination. The purpose of offering a set of proposals at all is to begin dialogue among various experts, not to make largely uninformed prescriptions.

AI ENGINEER’S GUILD - INFRASTRUCTURE

- The Guild operates globally as a network of subsidiary guilds.
- Each ‘sub-guild’ is devoted to a specific branch of AI development, allowing for industry-specific variation.
- National guild offices take care of legal compliance across multiple sub guilds.
- Guild HQ acts as an umbrella institution working towards continuous unification of objectives across sub-guilds and general policy guidance where possible.
- The Guild operates its own Marketplace of AI systems and solutions (possibly unnecessary/unfeasible - see below).
- An elected Guild Council is responsible for policy-making.

AI ENGINEER'S GUILD - ACTIVITIES

- Work to ensure AI research and development is conducted safely.
- Work to ensure AI systems are deployed to the market within safe parameters.
- Prevent the development and application of potentially destructive AI systems.
- Work to prevent monopolization and centralization of power off the back of AI technologies.
- Lobby for non-proliferation of certain forms of AI systems in the military.
- Act as advisers to government policy makers.
- Lead research into positive new applications of AI systems.
- Lead research into negative futures as a consequence of AI-related systems.
- Give the general public a voice regarding the integration of AI systems into society.
- Develop a framework for the safe development of AGI.
- Lead AGI research with a primary focus on safe development and application.
- Actively foster AI engineers' skill development.
- Promote equal opportunities among AI engineers.
- Promote activities of the Guild, its signatories and its members.

AI ENGINEER'S GUILD - MEMBER OBLIGATIONS

- Guild members are encouraged to work under contract-basis and not as permanent employees. Benefits such as social security contributions, pension plans and emergency financial support are provided by the Guild for those who work on a limited contract basis.
- Guild members are required to work under contracts prepared/vetted by the Guild. Legal and technical consultation is provided by the Guild to ensure their work is limited to specific goals and then only used for agreed upon purposes.
- Any employment of a Guild engineer requires the employing company or government to become a signatory of the Guild and operate under its code of safe conduct where powerful AI systems are used.
- Guild members are encouraged to produce 'spec' work - systems which are then made available for sale to signatory companies and governments through the Guild's own Marketplace (possibly unnecessary/unfeasible - see below).
- Members are forbidden from taking contracts which in any way support or operate weapon systems.

AI ENGINEER'S GUILD - SIGNATORY ORGANISATION'S OBLIGATIONS

- Signatory membership of companies or governments are directly tied to the contract under which they hire Guild engineers.
- Additionally, or where no such contract is in place, the signatory agrees to basic safety mechanisms for the use of AI systems as determined by the Guild for their activities.
- Permission is granted for safety audits by Guild representatives.
- Governments and companies may apply for signatory status with the Guild.
- All AI engineers employed by a signatory company/government must take on Guild membership.

- Any use of contracted AI work for systems not explicitly agreed upon represents a breach of contract - the client may be liable to legal action.

AI ENGINEER'S GUILD - STRATEGIES & INCENTIVES FOR EXPANSION

- Affiliate membership to the Guild is awarded to graduating students of AI-focused university courses. This creates a culture of loyalty to safe AI development and discourages young professionals from applying their skills to 'dangerous' non-signatory organisations.
- The weight of skills from the collective members outweighs that of any individual company or government, thereby incentivising private and public organisations to become signatories to the Guild, whereupon they will be required to adhere to the Guild's code of conduct pertaining to safe AI. In return, signatories are granted employment access to Guild engineers and to the Guild Marketplace, where they may purchase 'spec' AI systems (if such an entity is formed - see below).
- Only signatories of the Guild gain access to the Guild's Marketplace.
- Guild engineer contracts, and sales from the Guild Marketplace, are dynamically priced to enable startups, non-profits as well as local and low-GDP governments to afford safe AI solutions.
- The Guild takes full legal responsibility for the work of their engineers so members and signatories are indemnified against legal action following the failure of AI systems commissioned or purchased from the Guild.

AI ENGINEER'S GUILD - CONSOLIDATION

- Copies of all AI engineering work done by members is stored on off-shore guild servers, using quantum encryption with engineer-identifying authentication signatures.
- Security safeguards are put in place to deter member temptation from quitting the Guild following a major discovery.
- Members of the Guild Council are democratically elected and serve a limited term. Privileges are automatically revoked at the end of that term to prevent abuse of power.

As you see, these suggestions go way and beyond the achievements of even the most powerful writers' guilds. But ambition befits a sector which is set to be so much more impactful to society and humankind than writing for the Film & TV industry.

The most highly ambitious deviation from the writers' Guild system is the proposed creation of an AI Engineers' Guild Marketplace. This represents an effort to guide the AI race to safe ground through direct market competition. A radical strategy for any guild, this strategy is nevertheless comparable to OpenAI's mission to prevent unchecked monopolization and misuse of AGI by winning the AGI race themselves. The main difference is that the proposed Guild Marketplace would monetize minor Guild-regulated ALI breakthroughs first, in order to build the Guild's influence over time, and steadily impose safety regulations into ALI-applications even before AGI is discovered.

Needless to say, forming a market-dominating Guild Marketplace might well be a step too far into idealism. But with so many other strong incentives for individuals and signatory companies to join the Guild, our goals of risk mitigation may well be achieved without the Guild needing to compete directly in the global market for AI systems.

BUILDING THE AI ENGINEER'S GUILD - A PROCEDURE

In the event this proposal for the formation of an AI Engineers' Guild is deemed meritable as a force of AI risk mitigation, it is important to also explore how such a Guild can be built from the ground up. Whatever challenges might await a successfully established AI Engineers' Guild, many more lie immediately ahead in its establishment. One can be sure, however, that these challenges will only get harder the longer we delay.

It should be noted here that one important consideration for the establishment of a Guild is the need for overwhelming success - with anything less likely to spell failure. Should only half of all highly-skilled AI developers across the globe become Guild members, for example, it is safe to say that the companies, corporations and governments financing AI R&D would be incentivised to work with the remaining non-members, who are free of commercially hindering commitments.

Again, the below procedural proposal for building a successful AI Engineers' Guild should be taken only as a framework for discussion. It is expected that professionals in various fields will apply their expertise to shape a more realistic procedure that maximises the chances of success.

YEAR 1 - GUILD STARTUP

- A new General AI Challenge is introduced to explore in-depth analyses on how an AI Engineers' Guild might best function, and how it might best be set up. Meanwhile, Good AI collaborates with organisations like Partnership on AI to mobilize discussion through conventions and conferences, building a body of knowledge and opinions towards a blueprint. An online think tank processes discussions and highlights the most lauded ideas.
- Intense surveying takes place into the current status quo and modus operandi of AI engineers and researchers across the globe. Surveys explore the current and prospective foci of AI engineering work, employment incentives, market forces, human resource questions pertaining to performance and breakthroughs, potential dangers of AI engineering both in gravity and likely timescales as wellas whichever other questions are identified as important. This information also feeds into the blueprint for the Guild.
- An initial foundational Council is built around the most actively engaged parties to create a skeletal long term plan based on a vision of the infrastructure and management of a successful Guild, along with a roadmap for its establishment. The shape of the early Guild will be distinctly different to its long term vision and much more focused on incentivisation for membership growth than risk mitigation. It is imperative at this stage to safeguard against conflicts of interests among those participating.
- Leading AI engineers are identified - especially those deemed to be high-profile. These luminaries are approached and invited to contribute to the formation of the Guild as an active organisation, both in design and implementation. Those who accept become the

first members and, thanks to their reputations, draw attention to the Guild project throughout the AI development sector.
- The Guild is officially established and more high-profile members are sought out. Preliminary member obligations are modest.

YEAR 2 - FIRST TIER EXPANSION
- Guild research departments are set up, focusing on risk mitigation. Their findings feed into the early contracts between Guild members and signatories.
- The founding members are asked to invite their employers to become Guild signatory members and abide by the (also initially modest) rules regarding safe AI use.
- The company membership roster of Partnership on AI is invited to become Guild signatory members.
- Discussions take place with OpenAI about signing their researchers as Guild members and aligning goals towards creating a competitive Guild Marketplace (if such a branch of the Guild is deemed feasible/necessary).
- Current contracts permitting, the founding members are encouraged to develop ALI systems for sale in the Guild Marketplace. These systems are projected to be of high commercial value and created with the specific goal of commercially attracting more Guild signatory company and governments, who may only purchase these systems if they agree to signatory membership.
- The engineers responsible for the creation of these spec AI systems will retain full ownership of their work and thereby receive market-value reward for any sale. The only restriction to their ownership is full adherence to the Guild's rules of safety and conduct. These systems will not be open source, in order to attract greater fees from private companies.
- Marketplace pricing is dynamic so as to allow SMEs, non-profits and small/local governments to access new AI systems and themselves become Guild signatories.
- High security systems and encryption protocols are put in place to protect Guild IP.
- Revision history code is incorporated into all AI work, tied to authentication for decryption, thereby creating a culture of accountability.
- Signatory companies/governments (or potential signatories) are approached for application-specific commission work by Guild members.

YEAR 3 - SECOND TIER EXPANSION
- The first Guild Council elections are held. New checks and balances are put in place to prevent corruption and abuse of power.
- The bar for Guild engineer membership is lowered to include the next tier down in terms of skillsets and talent, as well as diversifying into new areas of AI research and application. All new members are commercially incentivised to create more AI systems for sale in the Guild's Marketplace.
- As membership increases, plans are put in place for the establishment of sub guilds to represent different fields of AI development.

- The first AGI taskforce is set up with the goal of developing and delivering safe AGI under the protection of the Guild. Its work is financed by ALI sales from the Marketplace as opposed to financing from signatories in order to safeguard against conflict of interest.
- Research on the contemporary global state of AI research guides next steps in the Guild's evolution and activities.
- The Guild Marketplace is developed as a competitive source of AI solutions - offering cheaper rates for safety-assured products and services. For the first time, legal guarantees are now provided along with contracts.

YEAR 4 ONWARDS

- Affiliate membership is granted to junior AI engineers and university graduates from signatory universities.
- The Guild's activities move towards goals defined and evolved since the organisation's conception - promoting decentralized safe development and application of AI systems.
- As those goals are achieved, the Guild shifts its attention towards consolidation of safety-driven influence over the tech industry.
- Polling systems are established so the general public may democratically influence the development and release of AI applications on behalf of the Guild.

CLOSING STATEMENT

The basic model of modern day writers' guilds potentially offers a tantalizing, though ambitious, solution for the safe development and application of AI. A long list of obstacles would lie ahead in the creation of a successful, globally functioning guild for AI engineers, especially given the diversity of work in this sector, the enormous incentives towards exclusive advantage and the permanent nature of many AI engineers' contracts. But, given the potential dangers of unregulated AGI development, especially at the hands of the world's militaries, the stakes are high for intervention. At the very least, an AGI Engineers' Guild would be able to mitigate the risks of ALI research and application. At best, it might place the race for AGI superiority under its guardianship and protect the interests of the many over the few.