Gian Maria Troiani

AI Ethics Journal

Assignment: Students will keep a journal on AI issues from the course and news this semester. It should be an e-journal of at least four pages in length (~1400 words), offering critical reflections from the perspective of ethics and politics on AI issues addressed in the class or items appearing in the news that are relevant to the course material.
This assignment is worth 15% of the grade.

Journal:

Reflection: Autonomous Agents

Autonomous Agents powered by Generative AI and Reinforcement learning are the frontier of ML, and their major impact on the entirety of society is self-evident. The pace of improvements is incredible, and they will soon be able to accomplish several tasks that we ask them to do for us. I believe this is a seismic shift that is starting to occur in software but will soon likely extend to an interaction with the physical world in the forms of robots, and IOT.

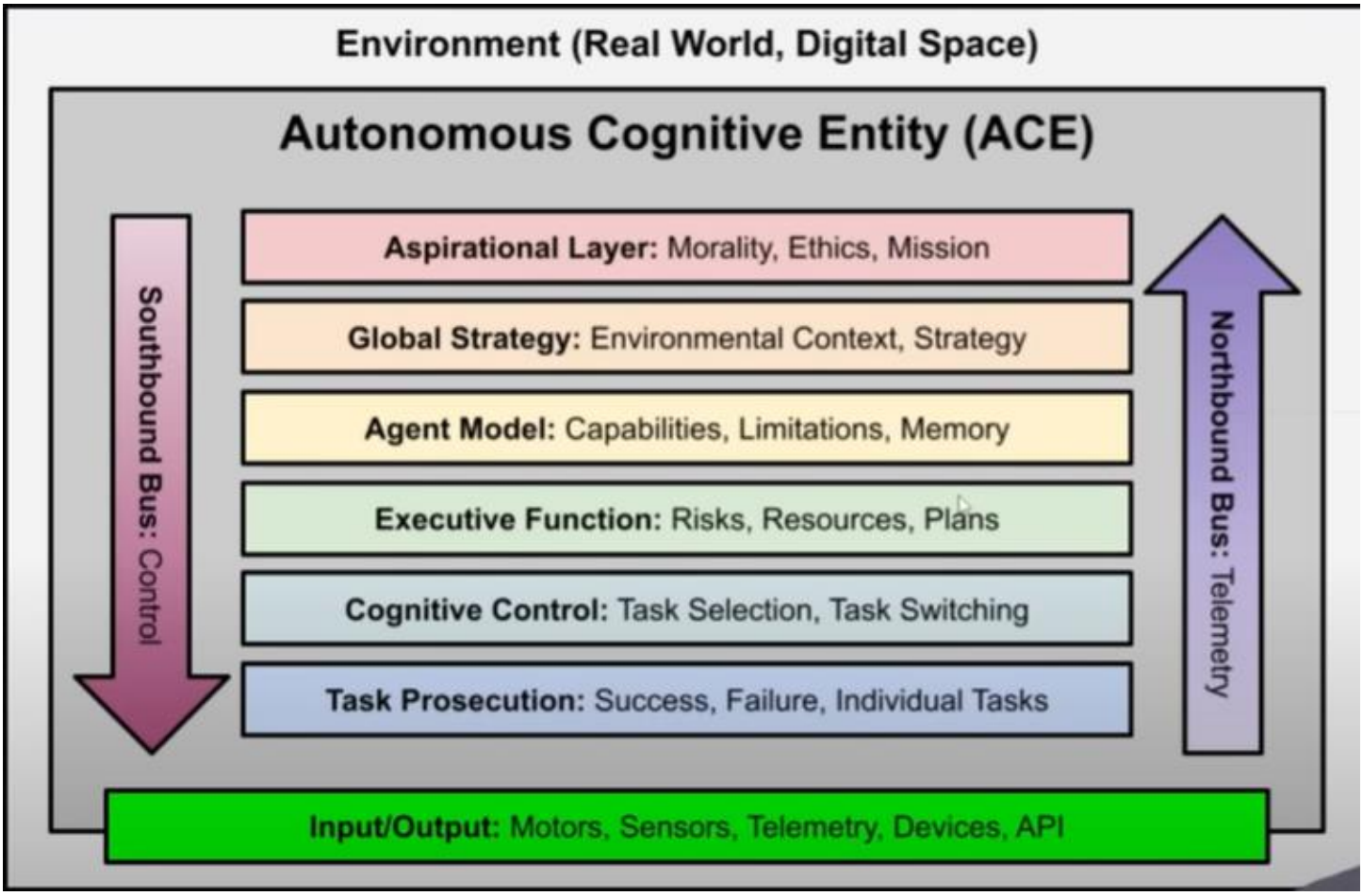
The latest version I have seen of this is the “Agent Swarm” technique where the developer basically asks an autonomous agent to create code to create other autonomous agents and organize them in some sort of way in a chain of command, or labor division fashion. Using this framework, it is possible to give a goal to the main agent which will later break-down the task and pass it to other agents, which will then report the fruit of their labor to the main agent and then finalize the output for the task.

This is extremely concerning from an ethical point of view, so far this still basically needs code to be done. II think some applications have the risk of getting out of hand quickly when these technologies go codeless/no-code and LLMs will achieve even better, more multi-modal capabilities and access to the internet.

While it is possible to anchor the output of these agents in check using techniques such as RAG retrieval when the agent does not have access to the internet. This is what prompted my question to Noelle about the topic. While the answer was awesome and very good in a corporate context where people are looking to keep bias metrics in check, it worried me about potential uses outside of environments where bias metrics are kept in check. For example, single developers with some budget could generate thousands of Autonomous agents using GPT-4.5 (or whatever Open Source model) and ask the agents to access and process information from the internet and use it to create applications later sold to the public without really checking for any bias. Or, I am sure it is possible to think about even worse scenarios where the agents are actually given the software tools to take actions and be used with bad intent. The first application that comes to mind is a programmatic effort to build and disseminate deep fakes and misinformation programmatically.

I strongly believe that at some point this will need serious regulations.

On the other hand, it was nice to notice that some people are trying to develop frameworks to build some sort of values/ethics into these new systems. [1]



Opinion: Countries mandate AI ethics/morality research quotas per law into big tech companies and university departments. We cannot advance technology too much without addressing the risk of completely and blindly losing control over it and how it is affecting society.

Reflection: Computational Preferences

[1] Shapiro, D. [David Shapiro]. (2023, April 1). *ACE Framework Overview and Intro: Autonomous AI Agents!* [Video]. YouTube. https://www.youtube.com/watch?v=A_BL_pu4Gtk

For over 2000 years, humanity has struggled to determine a universal ethical code. With the rapid development of AI, we urgently need to establish shared ethical guidelines before the technology becomes further embedded into society. I propose introducing "computational preferences" into foundational AI models to instill human values.
Rather than imposing a rigid ethical framework, perhaps it would be more feasible to conduct public surveys with questions probing moral viewpoints. By aggregating the results, patterns would emerge highlighting differences in values across geographical regions. Reinforcement learning algorithms could then be constructed to align model objectives with the preferences of a given population segment.
In this way, the computational models would organically integrate the priorities of humanity. Companies building foundational models could perhaps be able to use reinforcement learning through the aggregated filter/lens of what people value.Regional variation in preferences could also be respected by tailoring models to local moral sensibilities. By embedding ethical inclinations directly into model parameters, we can sidestep endlessly debating moral philosophy and pragmatically fuse AI with the value structure of society. I believe this data-driven approach balancing consequentialism and deontology could offer a path forward for responsible AI development.

Opinion and reflection: Aligning Societal Incentives

Implementing ethical AI requires more than just focusing on model design—we must align incentives at a societal level. Whenever a new model is created and deployed, the technology development process should take into consideration the incentives of all stakeholders (company, customers, end users of the technology, external society stakeholder, etc.) and how their incentives are structured. How will this technology impact the incentives of each of those categories? If incentives become misaligned, problems can arise despite good intentions.

For example, former Treasury Secretary Larry Summers advocated for financial deregulation to benefit society, which seemed reasonable morally and overall beneficial for society (e.g. allowing more families to afford houses). However, in the 2008 crisis, it became clear that incentives were dangerously skewed. Investment banks created complex products that credit agencies were paid to evaluate, yet agencies assigned undeservedly high ratings to boost business. Investors then flock to these products based on inflated ratings, ignoring the true risks. This chain of misaligned incentives culminated in a market crash when defaults materialized.

This case reveals why getting incentives right is crucial, even when pursuing moral goals. As AI progresses, we must learn from past mistakes and proactively ensure all players have stakeholder interests in mind, not just short-term profits. This is incredibly complex but vital work. Ethical implementation requires both model and incentive alignment. If society fails to align incentives properly, we risk enabling harm despite best intentions. The financial crisis offers a sobering lesson in this regard. By applying those insights to AI ethics now, we can potentially prevent unstable dynamics as the technology matures.

Opinion: Humans need to align themselves with their long-term values first

Solving the Human-AI alignment problem starts with people living lives more in-line with what they truly value in the long-term – which is not necessarily reflected in how they behave. Humans are more inclined to think in a short-time horizon and constantly engage in self-destructive behaviors.

Research and Reflection: AI and Creativity - Impact of new generative AI capabilities

As AI models become increasingly multimodal, creating music and applications in the visual arts are becoming accessible to an increasingly large number of people. I believe this might end up being a shock in many creative industries as it will negatively impact some artists while enabling many more people to experiment and create.

Music: It is currently possible to generate melodies of custom length from text prompts using platforms like Stable Audio, and the latest advancements of the Lyria model (Deep Mind + Youtube) will have ripples all over the music industry. Lyria features 2 main applications from what is known so far: Dream Track and Music AI tools. Dream track[2] is an experiment to deepen connections between artists, music creators and fans through personalized music creation. In the video provided, a Charlie Puth fan, inserts a text prompt("A ballad about how opposites attract, upbeat acoustic") and the model generates a catchy brand new clip of music, featuring brand new personalized lyrics sung by Charlie Puth. So, will the future of music be in the hands of the listener instead of the artist? What does it mean to be an artist at this point, and most importantly can we call this "art"?
Music AI tools[3] is a set of tools designed with artists, songwriter and producers to help the creative process. In the example provided, a person starts humming to a song and the humming melody can be transformed into an instrumental by the model. This makes making music so much more accessible to anyone without a music background and know-how. This does not diminish the value of rigorous musical theory or knowledge to craft a perfect melody, but it likely expands the amount of people that will express themselves through music without such background. I wonder if this will eventually lead to a further decline in the quality of the music that is constantly played on the radio, and of music in general, or whether it will become for many people the starting point of their musical journey. However, what this is sure to do, is impact a lot of jobs in the music industry such as producers.

Visual Arts: At this point, the most obvious application is image generation through text prompts with tools such as DALL-E3 and Midjourney. However, some companies such as Robohood and the artist Pindar Van Arman are using diffusion models to paint on canvas using Robotic arms and related technologies such as Reinforcement Learning and even Quantum Computing. Another interesting and very disruptive applications of AI to the visual arts are created by Runway ML which leverages text-to-video technology[4]. While at the beginning the technology was used to animate images for a few seconds, their models are becoming increasingly better. We are at the point that these videos are now starting to see their first applications in human-made documentaries and movies where the AI generated clips blend with the human generated content. Also, we are starting to see the AI

[2] [YouTube, F. M.]. (2023, November). Introducing Dream Track - an experiment on YouTube Shorts - featuring Charlie Puth [Video]. YouTube. https://youtu.be/1gjuHUy0IMM?t=3
[3] [YouTube, F. M., F. M.]. (2023, November). An Early Look at the Possibilities as we Experiment with AI and Music [Video]. YouTube. https://youtu.be/1gjuHUy0IMM?t=3
[4] [1] Unknown. (2023, December). Retrieved from https://d3phaj0sisr2ct.cloudfront.net/research/Gen2.mp4

generation of a few music videos. It is clear that eventually (meaning very soon) it will be possible to create entire movies, TV channels and real-time entertainment channels with 100% AI generated content.

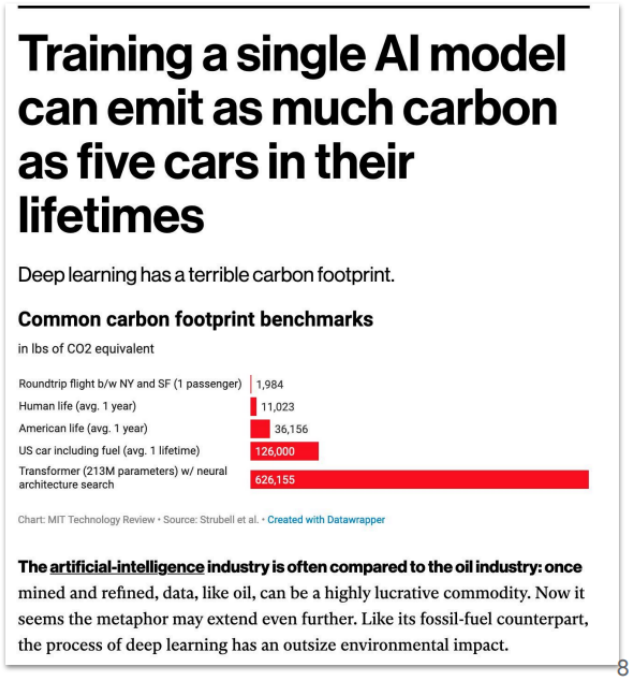
Given the current direction of the industry, it is clear that reinforcement learning will be increasingly more embedded into foundational models. I expect “planning capabilities over time steps” to get better. This leads (once again) to the ability of the creation of semi-autonomous agents. In the creative industry, I believe this will likely move away the time and effort of creators from lower-level tasks as they will be outsourced to multi-modal “semi-autonomous agents”. I believe this will increasingly amplify the potential scope of creative projects.
AI is also guaranteed to create more immersive experiences as it will blend with futuristic trends like VR, AR, XR. All the above mentioned new forms of creativity are coming to VR, AR, XR and will generate new ways to express and experience creativity. This will unlock the arts to take advantage of new media of expression and will likely make arts interactive for the user. I believe we will be able for example to interact with paintings in VR in a personalized manner. In a matter of time those paintings will become interactive movies to the point that it will probably be overlapping to the realm of videogames and metaverse.
AI and technology in general will extend creativity to new media of expression and experience, and I believe it will not kill creators, on the other hand, it will increase and empower the number of creators. These changes will completely change how most people will relate, create, and experience art. I think all these innovations in the creative industry might be starting to create income opportunities soon, hopefully offsetting some of the job displacement that will happen in other industries.

The Environmental Concern of Foundational Models[5]

A shift to data-centric AI from a model-centric AI might be a moral consideration going forward. The current state of the art in deep learning has been achieved from successfully scaling the transformer architecture to have a huge number of parameters. This process is extremely energy-intensive and has a substantial carbon footprint. The current approach does not seem scalable in terms of carbon footprint and projections show that at this rate we would need to improve chip technologies rapidly to even be able to continue sustain such an expansion: in fact, at the current rate, we would be running out of both GPUs memory to use to train large language models soon. According to research, the requirements of growth for GPUs and computation are not feasible. On top of that, they do constitute an environmental concern.

There is an effort underway by many developers to try to move away from these models in favor of smaller, cleaner data sets that can provide less biased data. This provides a longer-term sustainability as well as encourages more Responsible-AI practices.

[5] Reddi, V. J. (Year). The Parameter and Chip Wars—Moving Beyond Model-Centric AI Towards Data-Centric AI Systems. [Conference Session].AI Superstream: Data-Centric AI.



Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

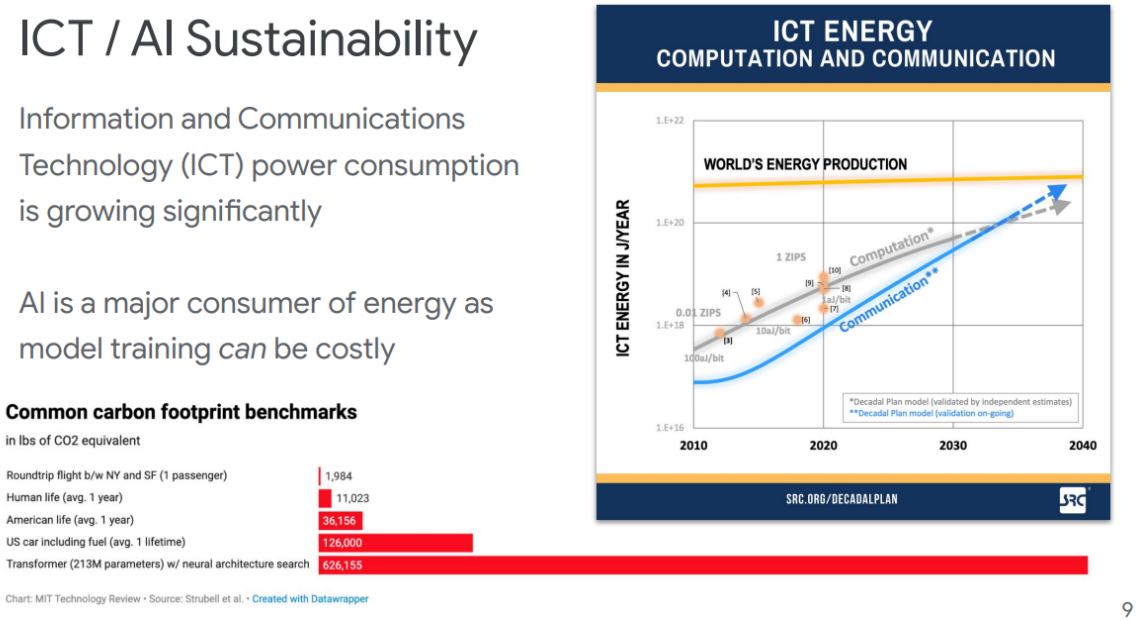
Common carbon footprint benchmarks

in lbs of CO2 equivalent

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

The artificial-intelligence industry is often compared to the oil industry: once mined and refined, data, like oil, can be a highly lucrative commodity. Now it seems the metaphor may extend even further. Like its fossil-fuel counterpart, the process of deep learning has an outsize environmental impact.

8



ICT / AI Sustainability

Information and Communications Technology (ICT) power consumption is growing significantly

AI is a major consumer of energy as model training *can* be costly

Common carbon footprint benchmarks

in lbs of CO2 equivalent

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

9

Takeaway notes from Ken Goodwill talk at Yale about Bioinformatics

AI in health care is exploding, raising many ethical questions about proper use. A flood of guidelines shows concern about over-trusting "black box" systems. Still, AI diagnostic power may soon consistently beat humans. While humans must stay "in charge" as licensed practitioners, our skills may erode. But if AI improves outcomes, does superior technique matter? And for whom: patients may value human qualities computers lack. Historically, sharing beneficial innovations globally has been poor. Companies also constrain transparency around AI limitations now. More skepticism is needed. Depth of education, not just AI system training, will help clinicians critically assess strengths and weaknesses. Hospital policies on appropriate AI use may become necessary. The pace of change is crazy – large language models capable of generating medical notes could enter hospitals

within a year. So providers must urgently take ownership of this technology influx, while safeguarding medicine's intellectual and professional role.

Reflection:AI & politics: dystopia or the future of values-oriented policy?

Could AI ethics one day positively enhance/replace political conversation? In politics, parties promote values by vying to govern based on their ideological priorities. However, politicians often fall short of expected ethical standards in pursuit of power or self-interest. As AI becomes deeply embedded in daily life, orchestrating its ubiquitous rollout across critical systems will prove hugely complex. Yet automated governance could also reduce complications of bureaucracy, questionable logic, and misaligned incentives in policymaking.

In theory, sufficiently advanced AI could surpass human ability at optimization and coordination on a societal level. If trusted systems were capable of managing complex dynamics better than fallible humans, would sticking with flawed human politics be immoral when technology promises better outcomes aligned with ethical goals? But equal danger lies in forfeiting control to autonomous systems. This conundrum warrants serious deliberation on risks and moral obligations.

Moreover, widespread automation may profoundly reshape political conversation. By handling intricate administrative and technical intricacies behind the scenes, AI could allow greater focus on debating fundamental questions about societal values and governance priorities. Freed of partisan pressures or self-interest, policies could better reflect ideological ideals rather than compromised settlements.

However, the systemic risks and moral implications of such automated governance likely outweigh speculative benefits currently. Relinquishing human oversight entirely, despite promises of optimized policies, remains deeply concerning given technical limitations and value alignment challenges.

Yet, the scenario highlights an important role for ethics in AI development today - embedding societal values early so systems have aligned goals to pursue "good" by whatever measures we define it. By confronting these complex automation tradeoffs openly now, we can potentially steer towards that vision responsibly. AI coders and ethicists should be proactively addressing these questions rather than reacting later.

And open philosophical discourse around values seems unlikely currently in today's political climate where politicians must compromise ideals to gain power. But perhaps these speculative futures can spur reflection on the role of ethics for 21st century policy making even in hybrid human-AI systems. Understanding public attitudes on algorithmic governance can also guide development towards humanistic principles.

By mapping out perspectives, risks and opportunities on emerging issues like automation in governance now while technologies remain largely theoretical, we allow more time for nuanced public debates that shape priorities. Technologists cannot dominate these discussions alone without input on societal values at stake. So an inclusive foresight process combining philosophical and technical expertise could offer the ideal path.

Reflection: Is our brain truly similar to an artificial neural network? Are we constantly switching objective functions?

1) our brain seems essentially a big neural network, more complex and fascinating but it seems too similar. We critique AI for saying things that are not true yet we have no clue what truth is and sometimes we say things just to satisfy some objective function which would be our goals in function of that. But really we are maximizing for some objective all the time and we just seem to be able to "switch" between objective functions.

So how is that any different from an improved version of the current networks we have?

Reflection: Parallel between moral theories and practices in layered model development

2) The main human ethical frameworks are utilitarianism, deontology and value ethics. There are significant parallels between what is being done at OpenAI:
Utilitarianism is the arbitrary objective function on the neural network, Value Ethics would be the human training that tell the model what is a good output or not (in value ethics you can only know what is moral by learning from an already good/moral person), and finally deontology would be some general rule always not permissible that makes some things are just unethical (eg killing someone). The corresponding incorporation of deontology would be the AI trainers selecting what the model should and should not do (e.g. telling someone how to build a bomb). There are parallels on how the development of these models occurs and applications of moral theories. The development process seems to almost adapt these ethical theories to the model in separate layers over time.

Is this the correct way of operating or would it be morally preferable to distribute the human training to most people in the whole world so we can teach AIs our values? (granted there would be a training process to do so, as otherwise the risks would outweigh the benefits).

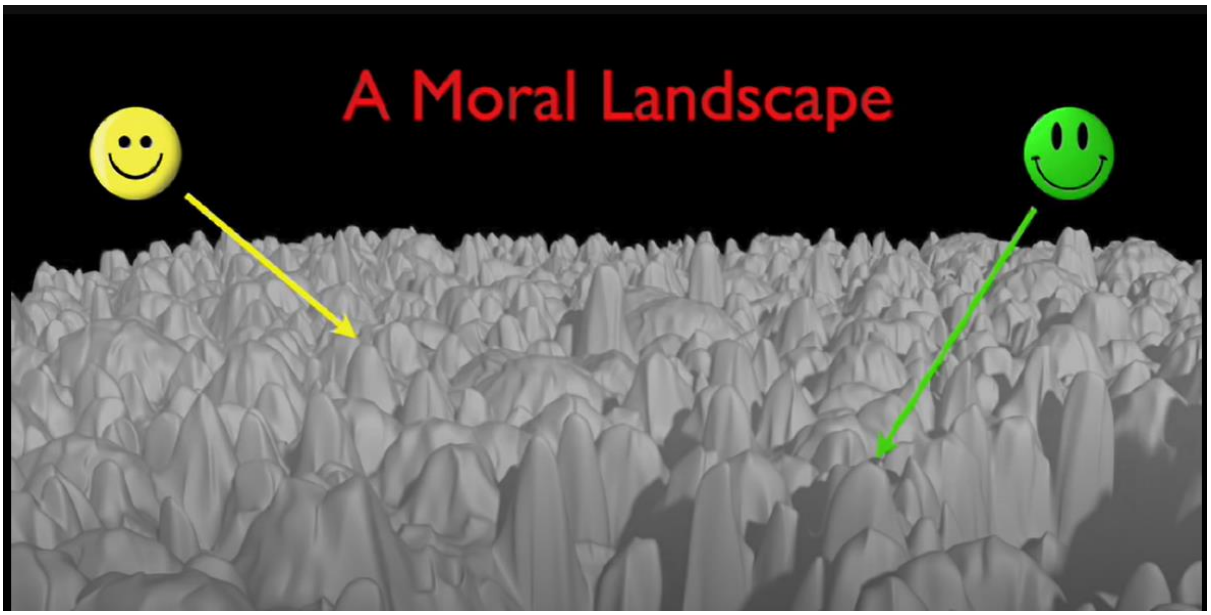
Reflection: Will AI "shipping" companies be the most valuable companies? Or will AI-ethics-oriented companies be the most valuable?

3) Given current trajectories, the most economically valuable companies in the coming decades will likely be those at the cutting edge of AI capabilities. After key inflection points, the firms able to solve pressing challenges in AI safety may become most influential for humanity's future.
Why? Because beyond a certain threshold of autonomous capability, getting AI ethics right could literally determine whether civilization flourishes or collapses. So, we must proactively confront these questions with our full intellectual capacity - the stakes could not be higher.

Reflection, opinion and open question: Is there such a thing as an AI moral landscape?[6]

Is there such a thing as a moral landscape for AI? Are there sets of moral global minima we do not want technology to touch? Is there a set of global maxima? I believe there are. However, what is the process to collectively figure out as a society what we think they are?

[6] [TED]. (2010, March 22). [Science can answer moral questions | Sam Harris]. YouTube. https://www.youtube.com/watch?v=Hj9oB4zpHww



Takeaway Notes from Mostaque & Diamandis discussion on the governance of Stability AI[7]

Governance of AI:
Proprietary AI models, like those from companies such as Anthropic and OpenAI, centralize power and create chokepoints, limiting access and innovation. Closed systems like Claude and GPT-3 are controlled by these entities, dictating who uses them and how. This contrasts with open-source ecosystems like Linux or PyTorch, where anyone can contribute and build on the technology, leading to quicker identification of flaws and biases, and fostering innovation. Open-source models decentralize power and promote community involvement in AI development, offering both ethical and functional benefits over proprietary models.
Governments should mandate standards for model training data sets and compute transparency. Stability AI is currently governed by Emad, but he is looking into alternatives like foundations and Decentralized Autonomous Organizations. Extremely advanced models should be treated as public goods and be collectively owned; smaller individual models can be privately owned. The key is to avoid centralization and build localization, so cultures don't outsource their “brains”. Regulation moves slowly; governments should focus on jobs, infrastructure, stoking innovation.

Safety and Alignment:

There is a need to separate AI governance from AI safety debates; the latter is hijacked by the precautionary principle. Transparent data inputs are key to safety; quality data avoids unsafe outputs. The scale of model outputs is a shortcut for low quality data. Alignment is about models serving societal interests, not just corporate interests. Diversity of high-quality inputs creates alignment and avoids fragile monocultures. Currently, responsibility falls on companies morally and socially more than legally.

---

[7] Source: [Peter H. Diamandis], '[Who Will Govern the Future of AGI? with Emad Mostaque (Stability AI Founder) | X (Twitter) Spaces],' YouTube, [https://www.youtube.com/watch?v=ZOJoPG9wqvI&t=1431s]