MID Working Thesis

Pitfalls and alternatives of foundation models

Ben Po-Sheng Cheng*

February 15, 2026

Abstract

Ever since OpenAI unveiled ChatGPT to the public, the idea of Artificial General Intelligence (AGI) has stormed the tech industry with its promise to be "the most impactful technology in human history". (Altman 2025) This is part of a larger paradigm shift of AI development towards general capability Foundation Models that consumes humongous resources. The conversational chatbot interface of ChatGPT has become a norm of how people expect to interact with AI. However, the paradigm of foundation model is dangerous. Enormous amount of energy, labor and capital have been poured into the development of foundation model despite its several overlooked issues, like reinforced bias and inequalities, unhealthy distortion of the science community and the environmental cost of datacenters.

The goal of this work is to challenge this AI paradigm by painting an alternative picture of what AI can be when it is on the opposite side of large, cloud-based, omni-capable and corporate-owned. The body of work is a designed electronics device and its tangible user interface. With them, I wish to provoke the conversation around new implications in terms of societal human-AI relationship.

*Rhode Island School of Design. Contact: pcheng01@risd.edu

Contents

1 Introduction

In recent years, AI has definitely become the technology in spotlight. AI startups are raising significant money every year, CEOs feel compelled to implement AI in their companies, while artists and news outlets are filing lawsuits against data infringement and the fear of being replaced by AI among workers are once again soaring. People seem to have mixed and strong feelings (perhaps too strong) about AI, but a proper conversation about what kind of AI we actually want is missing. The narrative around AI is usually extremely polarized – either a utopian picture or dystopian one. On the utopian side, prominent voice in the industry describe AI as the "the most impactful technology in human history", (Altman 2025) with the potential to solve cancer and climate change. On the other hand, an intelligence machine with super-huamn capability has always been the one of the greatest fear of humanity as seen in sci-fi movies and novels.

Neither of the narratives helps moving the technology forward to where we want it to be, because they concude flooring either the throttle or the break pedal. In reality, what AI actually can become deserves a much more granular answer, and this work seeks to paint a version of that. To begin with, I want to look at the current dominant paradigm of AI development and in particular, as an attempt to outline the potential improvement and also to establish the fact that AI has problems worth fixing, focusing on the issues and downfalls created by the paradigm. Subsequently, other alternative paradigm for AI development are explored, finally leading to my proposal as a designed tangible user interface.

The ultimate goal of this project is, however, to provoke conversations on the different societal implications of AI other than the two extreme utopian/dyspotian narratives. Through the designed device/interface, I wish people can start to see different ways AI can be built, used, and impact our life.

1.1 Issues of the current paradigm

Almost all of the modern large language models (including multi-modal ones and image generation models) are based on an architecture called transformer. (Bommasani et al. 2022) Published by researchers in Google in 2017, (Vaswani et al. 2017) way before ChatGPT became popular in around 2022, it is a type of machine learning models specilizes in understanding and generating natural language. It iteratively predicts the most likely next token (word) given a input paragraph.

The original transformer model does not posses the same level of conversational capabilities as GPT models. What OpenAI did to increase the capabilities of GPT was dramatically increasing the number of parameters inside the mathematical model. The original transformer model has up to 213 million parameters, while GPT-3 is almost 800 times bigger with its 175 billion parameter. (T. B. Brown et al. 2020) This paradigm of making models bigger in order to achieve better performance is famously characterized as OpenAI's Scaling Law. (Kaplan et al. 2020) Behind this paradigm, its seemingly impressive capabilities are made possible by an extremely large amount of computing power and taining data. The training of a state-of-the-art large language model require tens of thousands of computing units running non-stop for weeks. Each computing unit can consume up to 300 watts of power when running at full speed. They also generate a lot of heat, which takes additional energy and water to dissipate. Power consumption of the largest modern datacenters are designed to consume around 150 megawatts, equivalent to nearly 122,000 american households.

These humongous models are also commonly referred to as Foundation Models, because they are large (in terms of number of parameters/weights of neural networks), possess general capability, and act as a foundation or intermediate for downstream AI products. (Bommasani et al. 2022) Their large scale means they are extremely expensive to develop, and only a handful of resourceful companies are able to afford the computing power, energy, water, and labor required. Their general capability means their behavior is unpredictable, and there is a pressing lack of proper benchmarks to evaluate their performance and externalities. They have a broad impact and implications as they are adapted by all kinds of AI applications; however, their intrinsic biases are also inherited by every downstream product, reinforcing and amplifying existing stereotypes and inequalities.

1.2 Alternatives

1.3 Tangible interface

A tangible interface is a physical device that allows spectators to imagine how it can be used and also what context it can be situated in. It opens a broader conversation on not just what AI can be but also wider speculations of the societal impact of AI. If the alternative AI paradigm in my proposition is small, local, private and personal, naturally the machine learning model has to live in a personal electronics device, and people inevitably interact with the model in some ways through this device.

Ever since the popularity of ChatGPT skyrocketed, the designed interface for Large Language Models has been centered around various forms of chatbot. This make sense, as these models are built to understand and predit natural langauge. However, chatbot alone is a very inefficient way to interact with these models. For example, in task like coding and image generation, it is very rare that the models can produce the ideal result on the first try, usually it takes several iterations and additional prompts to modify the initial result to perfection. Under the hood, these modifying prompts are appended to the initial prompt, making the input for all subsequent requests longer and longer. Additionally, chatbot as a human computer interface also provides no clue to how the technology actually works, and thus deprive the user from learning best practices and trouble-shooting techniques. This is not the say chatbot is the wrong interface but to argue that more control and configurability should be implemented, either in hardware form or software form.

Through my designed tangible interface, I wish to illustrate a different type of AI models that is small, local, private and personal, and to explore the new possibilities of how AI can/should be used and the alternative scoietal implication in terms of human-AI relationship.

2 Unseen downfalls

2.1 Nature resources

As estimated by Strubell, Ganesh, and McCallum 2019, the training of a large language model can generate more than 5 times the carbon footprint of the lifetime of a car including fuel. Besides training, running these models is also a energy hungry task. According to Luccioni, Jernite, and Strubell 2024, the energy required to generate an image using large foundation models is equivalent to one fifth of the energy stored in an iPhone battery. Energy consumption of these datacenters is only predicted to continue rising, surpassing that of the entire of state of Japan (1000 TWh) in 2026. (IEA 2024)

Humongous datacenters not only take electricity to run, they also require water for cooling down the computing units. Li et al. 2025 estimated that the training of OpenAI's 175 billion parameter model GPT-3 consumed 5.4 million liters of water in Microsoft's datacenters. Depending on where the model is hosted, it also consumes 500ml of water for every 10-50 medium-length responses generated.

2.2 Labor

As big techs race to spend tens of billions building out the biggest datacenters to support the foundation model boom, land, communities and nature environments are being exploited. On the outskirts of Santiago, Chile, the municipality of Quilicura hosts various factories and Google's datacenter. Only two decades ago, Quilicura was a rural town home to the rich biodiversity of the Atacama Desert. Now, there is only industrial complexes and the "Quilicura Urban Forest", built by Google as a community giveback in 2019. However, there are not many residents still living there as it has become an industrial area. Like most datacenters around the world, Google's datacenter also does not provide any job for the locals except for the temporary construction work – in 2024, a job posting for mechanical technician for the datacenter requires applicant to submit resume only in English. (Hao 2025)

Generative foundation models like GPT are trained on enormous amounts of datasets, many of which are impossible to filter or censor. As a result, the original models produce sensitive content, and companies like OpenAI hire Silicon Valley middlemen like Sama or Scale to access cheap labor in developing countries to moderate the generated content. It is reported that Sama hired people in Kenya for an average of $1.46 to $3.74 to do this work. They are exposed to sexual, violent text content including child sexual abuse, slavery, and bestiality while their employer provides minimum mental health support, which often results in serious mental health conditions. (Hao 2025)

These examples highlight the multifaceted impact of western technological advancements. They exploit resources from the underprivileged and destroy their land, community, and lifestyle.

2.3 Bias and stereotypes

Foundation models act as a general intermediate for various AI products; they are often fine-tuned, augmented, or adapted for more specific use cases. Because they are trained with extremely large datasets like the Common Crawl, the bias of the datasets is inherited by the models. For example, there is more content on the internet in English than in any other language. The large amount of online content that is hateful, abusive, and toxic inevitably becomes part of the training material and also directs the model's behavior. The overrepresentation, underrepresentation, and misrepresentation across cultural, geographical, and racial groups are also carried through to foundation models.

These intrinsic biases are inherited by downstream applications that adapt foundation models, creating widespread performance disparities for various use cases. For example, Koenecke et al. 2020 looked into how African American English speakers who cannot access reliable speech recognition technologies are in a disadvantageous position when these technologies are used in job interviews or transcribing courtroom proceedings. According to Hofmann et al. 2024, models exhibit stereotypes about speakers of African American English. They may be more likely to suggest that speakers of African American English be assigned less-prestigious jobs, be convicted of crimes, and be sentenced to death. Wilson et al. 2021 found that models fail to detect faces of people with darker skin tones.

2.4 Data ripoff

On top of that, the training data are often obtained without clear consent. Several news outlets have filed lawsuits against foundation model developers like OpenAI and/or Anthropic for illegally using their publications as training materials. These materials are usually consumed by the model training to update its parameters. They only exist in the generic form of numbers and vectors, erasing the identity and intention of the original authors. Image generation models are trained on images created by artists and illustrators, only to eventually replace them in the workplace.

Actress Scarlett Johansson famously accused of OpenAI creating a ChatGPT voice assistant that sounded almost identical to her even after she declined to license her voice to OpenAI. "I was shocked, angered and in disbelief that Mr. Altman would pursue a voice that sounded so eerily similar to mine that my closest friends and news outlets could not tell the difference," she said. (Allyn 2024) In response, OpenAI's CEO Sam Altman stated that the model was trained on the voice of another actor.

2.5 Economic inequality

It's easy to describe AI as a groundbreaking technological advancement that benefits all of humanity, solving climate change, cancer, or other troubling issues that pose a significant danger to our species. However, none of these promises seem remotely close to being realized, and the actual economic impact of AI paints a much grimmer picture.

Acemoglu 2024 estimates the Total Factor Productivity[1] growth generated by AI to be 0.66% over a ten year period, which is much more modest than people might expect. Moreover, as estimated by Bursztyn et al. 2025, emerging technologies like social media may have negative impact on social welfare even if they create economy growth on paper. We've already seen various malicious and manipulative uses enabled or exacerbated by AI, like disinformation. (Buchanan et al. 2021)

According to Acemoglu and Restrepo 2019, when new technologies are able to drive substantial productivity growth, even if a large portion of routine tasks are automated and thus lower-skill workers are replaced, the reorganization of production processes creates new tasks and jobs for human labor. However, if the productivity growth is modest or if the new technology's sole focus is to replace human labor, new tasks may not emerge, and task displacement means low-skilled workers will suffer from job loss. This is what Acemoglu described as "so-so" automation, and AI, with its estimated 0.66% productivity growth, seems dangerously close to falling into this group.

Recent developments in economics suggest that technology, in particular those that enable automation in the production process of goods or services, creates an imbalance and unequal outcome for people in different income or demographic groups. (Acemoglu and Restrepo 2022; Moll, Rachel, and Restrepo 2022) Automation usually replaces low-skilled labor in the production process. In other words,

[1] In simpler terms, the combined productivity of production input (capital and labor). Traditionally its growth is attributed to technological advancement.

the people who lose their jobs to AI are theoretically those who are already in the lower income bracket. Additionally, more automation means a higher return on capital investment, delivering more earnings to the wealthy. Skare, Gavurova, and Blažević Burić 2024 shows that the capital stock of AI is unevenly distributed and positively correlated with exacerbated income inequality. Using online job posting data, Berger et al. 2024 finds that hiring for lower-level office jobs declined after ChatGPT's release, while demand for higher-skilled workers increased.

Economic inequality not only exists across income groups but also across demographics. Preliminary estimation in Acemoglu 2024 suggests that AI might cause the wages of low-education women to decline. Appel et al. 2025 finds that countries with higher AI usage are concentrated in North America, Western Europe, and Northeast Asia, while the global south almost entirely falls into the minimum usage group. In their theoretical framework, Ide and Talamàs 0000 found that AI is more likely to displace workers from complex problem-solving work to routine tasks in a developing economy, unlike in an advanced economy where people will be better positioned to use AI to help them tackle more complicated work. All of this seems to suggest that the prevalence of AI is empowering the already-empowered while underprivileged groups will benefit less from AI.

Rising economic inequality has significant implications. Long-lasting stagnant wages and loss of employment opportunities in certain groups create popular discontent that sometimes fuels disruptive populist movements. Uneven distribution of income also leads to the consolidation of wealth for certain individuals or business sectors, which in turn results in the consolidation of power. This is particularly true in the case of the current AI landscape, in which very few people possess immense power over the trajectory of its development and are able to steer the future of AI to benefit themselves.

2.6 Power Concentration

TODO: Add description on how the trajectory of AI is dictated by very few powerful entities/men. OpenAI's power struggle. Capital investment. Big techs are the only player. Effective Altruism

2.6.1 Death of Open Science

The paradigm of foundation models represents the homogenization in the field of machine learning research and development. Specifically, Bommasani et al. 2022 characterize homogenization as “the consolidation of methodologies for building machine learning systems across a wide range of applications.” The origin of foundation models stems from the field of Natural Language Processing (NLP), with key breakthroughs including self-supervised learning, word embedding, the transformer (Vaswani et al. 2017), and the bi-directional encoder-decoder (Devlin et al. 2019). These technical architectures represent the essential properties of foundation models, laying out their uncanny capabilities but also their intrinsic downfalls.

Today, almost all state-of-the-art NLP models are derived from a handful of foundation models like the bidirectional encoder model BERT. (Bommasani et al. 2022) The transformer architecture of large language models is also beginning to be widely applied to other areas like speech recognition, image, biology, and reinforcement learning. The phenomenon of foundation models as a paradigm taking over other forms of machine learning systems is the key motivation of this work, particularly because of the ongoing issues it’s creating and the huge amount of resources required to develop them. AI, as a frontier discipline of science, should be explored openly. However, we are observing foundation models crowding out the resources for research and development in other machine learning systems.

Even worse, the development of foundation models is so expensive that only a handful of “big-techs” can afford it now. It’s getting more and more difficult for independent researchers unaffiliated with big-tech companies to conduct research in cutting-edge foundation models. Because of the sheer size of these models, their performance claimed by the private developers is extremely hard to verify independently. Evaluating their real-world cost and externalities has become equally difficult for the same reason.

Founded as a non-profit to ensure the safety of AI development, OpenAI stopped open-sourcing its models after GPT-3 in 2020, citing security reasons. This has become the norm across many big players in the field, including Google and Anthropic. (Hao 2025) Subsequently, OpenAI has transformed itself into a private for-profit company. The field of cutting-edge foundation model development increasingly looks like another technology race in Silicon Valley, rather than an open community in which members share knowledge and breakthroughs with each other for the advancement of science that benefits humanity. If AI is what we believe to be the next leap forward in human civilization, the path we are taking to get there seems like an extremely dangerous one.

The thriving community of AI Ethics researchers set out to tackle these threats by collaborating on research evaluating the issues of foundation models and coming up with appropriate benchmarks. Nonetheless, big-techs still find their way to try to consolidate their power in this community. Mohamed Abdalla and Moustafa Abdalla 2021 finds that big-techs use the same strategy as big tobacco in the 1950s to undermine research questioning their products. They publicly claim their emphasis on and care for AI safety and even set up internal research divisions to conduct research. They also sponsor and provide funding for top conferences, events, and individuals in AI Ethics. This presents conflicts of interest and threats to academic integrity. According to the cited work, there are more researchers in AI Ethics at top institutions that have received funding from or been affiliated with big-techs than those who have not. In 2020, researcher Timnit Gebru was ousted by Google over the publication of her now-famous paper Stochastic Parrots, highlighting the issues of large language models sheerly repeating their own training data. (Hao 2025, Bender et al. 2021)

3 Forgotten paths

Known as the "father of AI", congnitive and computer scientist Marvin Minsky wrote in his book *The Society of Mind*, "The power of intelligence stems from our vast diversity, not from any single, perfect principle." He portraits the human mind as a society of simple but very different cognitive processes, also known as agents. (Minsky 1988) Each of these individual thinking entity has different internal logics, combined to form the vast intelligence of our human mind.

The current paradigm in Artificial Intelligence, however, is devoted to building systems that are the complete opposite. The paradigm of foundation model suggests that a single, humongous "master algorithm" will be able to reproduce the extremely complicated and inter-weaved cognitive capabilities of the human mind. In particular, the uprise of Large Language Models implies that language is the only necessary tool for intelligent systems to reason and make decisions.

Instead of incorporating a fundamentally different architecture for vision capabilities, the recent development of multi-modal Vision-Language Models uses the same word embedding techniques for understanding images, essentially treating each small “patch” of an image as a word. (Dosovitskiy et al. 2021) It seems like the world firmly believes that this one single way of developing AI systems can lead to something on par with the human mind.

The core proposition of this thesis is that instead of pursuing an omni-capable master algorithm, resources should be more equally devoted to developing various kinds of AI systems. This section outlines the other alternative paradigms of AI that I believe should have been given the same amount of attention. Some of them draw inspiration from the intrinsic downfalls of foundation models, others from the negative externalities we are observing around the world, and still others from the cognitive science that depicts how human intelligence works.

3.1 Application-Driven AI

The premise of foundation models, or more specifically transformer-based large language models, as the major breakthrough towards artificial general intelligence is that one giant model will be sufficient to encapsulate all sorts of human intelligence. Indeed, in recent years, we’ve seen substantial improvement in the performance of these models. However, The so-called “performance” are usually evaluated with hypothetical datasets and metrics that doesn’t translate to real-world capabilities.

This is what Kiri Wagstaff of the California Institute of Technology describes as “machine learning for machine learning’s sake”. (Wagstaff 2012) A lot of these improvements don’t necessarily translate to real-world usages. For example, a computer vision model might be able to successfully detect pedestrians with a 99.9% rate on a benchmark dataset; nevertheless, failing to identify one pedestrian in every 1000 is still fatal and doesn’t mean the model is nearly close to being capable enough to be applied to critical applications like autonomous vehicles. The long tradition of machine learning is based on a stochastic process instead of deterministic decision-making. Much of artificial intelligence’s vast improvement in performance is driven by raising the probability of success, not real-world usage. Moreover, the same metric for success rate is not comparable across different applications. 99.9% might not be enough for self-driving cars but might be sufficient for identifying food ingredients. This is again to say that the pursuit of performance in benchmarking metrics tells us nothing about how usable it is in real-world applications.

Additionally, Wagstaff also argues that the machine learning community needs to engage with real-world domain experts more closely. The prevailing paradigm of collaboration with other domain experts usually only happens for curating and annotating training datasets (Sambasivan et al. 2021), but little for creating the appropriate evaluations for real-world applications and for their potential impact.

Creating models that have real-world applications in mind from the initiation of their development also boasts the benefit of having smaller, more efficient models. Over the past couple of years, we've seen developments in machine vision, audio, and robotics all converge to the paradigm of a multi-modal transformer architecture. The one-model-to-rule-them-all concept does seem tempting, but the truth is that different tasks may require different architectures for more efficient computing. Even having more application-specific transformers may also be beneficial.

Smaller models also enable the possibility of being run locally on smaller hardware, which allows their users to have a very different relationship with the models. Models that can be run locally will also be easier to be trained and fine-tuned by their users, which means they can adapt more granularly to a user's personal needs and lifestyle.

3.2 Cognitive Algorithm

If the goal of Artificial Intelligence is to build computing systems that rival human intelligence, it makes sense to mimic how the human mind works. Unfortunately, the human brain and mind are such complicated systems that to this date, we still haven't solved their mystery. However, psychology, cognitive and neural science are still able to give us some hints.

New York University professors Gary Marcus and Ernest Davis wrote in their book *Rebooting AI* that drawing from Kant's *Critique of Pure Reason*, "time, space, and causality are the three philosophical grounds fundamental to the human mind." (Marcus and Davis 2020) When seeing a car coming while trying to cross the road, humans are able, in almost no time, to estimate how far the car is, how long it's going to take for it to be dangerously near, and how long it would take to cross the road, and then make a decision. Such simple everyday tasks are usually comprised of many complicated thoughts. We are able to "sense" the distance and time of the car with little to no brain effort, at the same time considering other factors, like the sound volume of the coming car or whether you are about to be late or not.

3.2.1 Purely Analog Senses

Given a well-defined context and sufficient knowledge, existing machine learning systems might be able to understand the time duration of a certain event, but they are only able to do so by inducing from mathematical formulas. They have the knowledge of time, but humans have a sense of time. Similarly, AI has an understanding of the shapes of objects, but it might not be able to immediately induce their functionality from the shapes. Imagine seeing a cheese grater for the first time; we are able to reason with its scale, shape, material, and texture to immediately guess what it might be for and how to use it. This kind of holistic reasoning combining our innate understanding of time, space, and causality is exactly the kind that intelligent machines should learn from the human mind. Of course, by showing a cheese grater and its text description to a model during training, current machine learning systems would have no problem understanding what a cheese grater is. But perhaps that's why foundation models have to be large models consuming almost all the available existing data to train. It is important to note that fundamentally, current state-of-the-art models are still language models that understand everything through language and language only. Computers fundamentally cannot "sense" anything but can understand its "abstraction" as language or other symbolic systems, like a mathematical formula.

A maybe more efficient way to build AI systems is to look into the other "senses" that do not yet exist in the digital world so that AI can reason and induce more holistically. Upon seeing the picture of a textile, we are immediately able to guess its texture, flexibility, or even whether it's waterproof or not, all without touching it. The existing "senses" computers have include light, sound, force, temperature, electrical potential, magnetic fields, and more. All of these are mathematically well-defined physical properties, and they all have corresponding sensors that can transform them into digital readouts. We humans, however, do not rely on the understanding of physical properties to perceive the world. We understand the world through smell, texture, haptics, and more. Machine learning systems, on the other hand, do not possess this kind of capability to understand these "mathematically ill-defined" and thus purely analog properties.

Some recent developments in the field of the so-called world model claim to be able to generate realistic simulations of the physical world. (Bruce et al. 2024; Gupta et al. 2022) What these models actually are is models capable of playing video games. It is unsure whether they are just reproducing common physical phenomena in video training data or if they indeed have an understanding of the rules of physics. Even if they do, their output is still limited to just video or visual images – far from all the complicated analog senses humans have. We should tap into these unknown territories of the digital world, trying to come up with more abstraction or symbolic systems to teach machines more nuances of this physical world. This can potentially eliminate the vast resources required to teach models everything with empirical training materials. And in the context of the current paradigm of large language foundation models, giving computing devices more "senses" to characterize the world also saves us the effort and data size of trying to describe everything in text.

3.2.2 Common Senses and Causality

When we say Novak Djokovic has been playing professional tennis for 20 years, we don't mean that he has been playing tennis non-stop for the continuous time duration of 20 years. The human mind is capable of understanding the implied meaning that tennis has been Novak Djokovic's profession for 20 years. However, a language model that is trained to understand the literal meaning of texts might not be able to decipher the latent implication.[2] This is an example of what we vaguely describe as "common sense". We, as humans, do not need to be taught that it is impossible to continuously play tennis for 20 years, even for a professional athlete, and are able to comprehend the idea of a profession because we are able to understand the world and human lives by employing different knowledge frameworks simultaneously. Instead of reading 20 years as 365 days and one day as 24 hours, our sense of 20 years as "a very long time" substantially outweighs

[2] I prompted Claude Sonnet 4.5, "suppose one year is 365 days, one day is 24 hours, one hour is 60 minutes. Novak Djokovic has been playing tennis for 20 years, how long has he been playing tennis?", and then the chatbot went on to actually convert 20 years to hours and minutes. Although it is also quite astonishing that the model was able to understand the implied meaning of the first part of the prompt, a human being would probably first ask for clarification.

the symbolic meaning of a year as a unit of time. The human mind is firstly a sensing entity and then an abstraction decoder. We built the various symbolic systems based on our innate senses; machine learning systems, however, can only induce from the abstraction and symbols. As a result, the lack of common sense has been a challenge for AI.

An alternative AI paradigm is trying to program common sense, or the general understanding of causality, into purely mathematical models. We understand the basic causal relationship of everyday events and objects, but even the largest foundation model today does not have any mechanism to employ this kind of knowledge. It only has the most surface-level understanding of our common sense even after being trained on the largest corpus of text material. Carnegie Mellon University computer scientist Douglas Lenat was an advocate for intelligent machine systems with programmed causality. (Lenat and J. S. Brown 1984) His long-term project Cyc aims to create a symbolic system and a knowledge base database that maps out the relationship of everyday common sense and implicit knowledge. It's a decision-making algorithm that is based on purely deterministic and pre-programmed relationships. After almost 40 years of employing philosophers and programmers to write down rules of the world for machines to understand, the project unfortunately never gained much traction. However, it is a sharp contrast to the strictly stochastic and training-based paradigm of foundation models.

Amidst the enormous resources being poured into developing large models, maybe it is a good time to revisit the idea that machines should be able to employ some simple pre-programmed rules that describe how the world works, instead of trying to make the algorithms learn every single relationship embedded in our common sense.

3.2.3 Learning versus Innateness

Large foundation models, as statistical pattern-matching machines, try to generalize everything into a widely applicable algorithm. Even if they were able to learn the entire common sense of the human mind from a large corpus of data, not everything follows rules. Our human intelligence makes decisions based on not only the abstraction and generalization of our various senses and experiences but also our innate values. We internalize our experiences into an individual's unique logic.

Psychology provides us with plenty of evidence that human decision-making does not only come from learnings from the outside world but also from internalized values. Psychologist John Watson of Johns Hopkins University, a major pioneer in *behaviorism* psychology in the early 1900s, famously claimed that a child's behavioral pattern can be dictated purely by controlling every factor of the environment in which the child was raised. (Watson 1928) Only a couple of decades later, Noam Chomsky, who is usually seen as the founder of *cognitive* psychology, published the groundbreaking work in linguistics *Syntactic Structures*. (Chomsky 1957). Along with his other works, he characterizes language understanding not as pattern-matching of past encounters of sentences but as a result of internalized grammar. (Chomsky 1959) The old perspective of behaviorism explains behavioral patterns entirely with an external reward system; in contrast, Noam Chomsky's cognitive psychology attributes the driving force of human intelligence to internal representations like desires, beliefs, and purpose. To this date, this view from cognitive psychology is still agreed upon by scholars, while behaviorism theories have vanished. (Marcus and Davis 2020)

The paradigm of foundation models, in a sense, is like dictating a child's behavior by raising them in a controlled environment. Every model starts from a bunch of random numbers and then learns everything from data by maximizing rewards or minimizing loss. Although so far we've seen this approach make big progress, it is an extremely resource-intensive process, and we must remember that this is not how the human mind works. We should look into ways to design models that are capable of internalizing values and generating internal representations of these values. Structuring experiences and learnings to create a unique internal logic is a crucial "human" aspect of human intelligence. Artificial intelligence should also be able to replicate its user's internal representation in the training/inferencing process. Most large language models might be able to adapt to a user's response, but they do so by appending the user's previous feedback to future prompts. Internally, it's just the same model processing longer inputs with more added context. Compared to the human mind's capability of generating internal representations, this is extremely inefficient and again adds to the already enormous resources required to run these models.

3.3 Trust

TODO: Talk about why people lose their trust on AI? How can we rebuild trust? Can we reimagine our relationship with AI? Perhaps more private, transparent and tactile experience.

3.4 Participatory AI

TODO: A more democratized process of AI development. Data-centric AI (Liang et al. 2022; Sambasivan et al. 2021)

4 Observations

4.1 AI is not good enough

By saying that AI would benefit the world because it can solve cancer, climate change and poverty, popular AI advocative sentiment usually implies that if we solve AI, then we essentially solve everything. However, the actual implication of a solved world is far beyond general prosperity and panacae. Philosopher Nick Bostrom describes the condition of a solved world as *Technology Maturity*, meaning that with AI we would have all the technologies we can possibly have and use them to solve whatever problem that remains. (Bostrom 2024) When we look into what exactly these technology are, you will see that we are still so far away from technology maturity.

TODO: Examples. Brain-computer interface, solving AI is almost contingent on solving the human brain, which we are still far.

4.1.1 Language alone is not enough

4.2 Is AI ever going to be good enough?

TODO: Can we have human-level intelligence machine without inducing moral-status? Sentience machine (Bostrom 2024)

4.2.1 Do we really want AI to be that good?

TODO: Purpose? Return of Malthusian economy. Extreme inequality – all earning goes to capital. Reference Sec. 2.5

4.3 AI is machine, not human

The discussion of the long-term utopian (or dystopian) vision and speculation of AI is beyond the scope of this design work. We are designing for the world with current technologies and socioeconomical conditions. However, extrapolating the extreme future informs us a lot about what the currrent vesion of AI we have actually is and how it sits in the imaginable roadmap of the technology. Only once we have those understandings can we design for the AI we have instead the AI we envision.

Following previous discussions, we know that the current version of AI we have is still far away from what's promised. That is, the technology itself is still not good enough. However, in the discipline of design, we are designing the derivative product or human interface of AI *as if* the technology was that good. The perimeter of the very limited capability of models is intentionally invisible. The intrinsic downfalls and externalities are hidden. In pursuing ease-of-use and "seamless user experience", user-configurablity is completely abandoned. It seems that designers want users to believe that the models can figure out what exactly what's needed where in fact more means of user input could potentially improve the performance dramatically. The prevalent chatbot interface gives up on informing the user how models work, essentially depriving user's oppurtunity to make models work better for them.

Chatbot is a human-human interface that lives on computing devices, not a human-computer interface. One of the implication of technology maturity is that whatever intelligence machine we come up with, we can interact with it in a manner extremely similar to human-human interaction. AI, as it is now, is a sort of intelligence machine that is still far from technology maturity. However, that interface that we came up with is the one that is supposed to make sense only once technology maturity is reached. In designing a interface like chatbot, the desginer forgot the fact that AI is still fundamentally a machine. Machines need to be operated in a certain way to yield maximum performance and efficiency. Operating a machine requires that the user has some technical knowledge. A well-planned learning curve is what makes a human interface usable. Typing on a keyboard is a learning curve. Scrolling, pinch-to-zoom, long-press on a touch screen all demand leanring. For a technology like AI that is still novel to the public, opting for a human-like interface that has close to no learning curve seems more like evading the hard work rather than being considerate.

In every other technology humanity has ever seen before, its human interface is always one that bridges the needs of users and the limitations of the technology. The purpose of human interface is to make the technology works for the human, hence the interface has to work for both the technology and the human. Chatbot, however, is one that only works for human but takes no consideration for the intrinsic properties and operational parameters of the models.

A good human interface informs the user the capabilities (and limitations) of the technology, best operational practices and potential means of trouble-shooting. Every usable machine in the history has a human interface that accomplishes these things. When a machinist is operating a laith, they can observe whether the power, gear ratio and cutting tool are making a clean cut. If they want to change the speed of the rotor, it's clear that the most preferable way is to change the gear ratio. If the cut is not clean even under optimal rotational speed, they would know maybe the cutting tool needs to be sharpened. Using a laith requires a somewhat steep learning curve, but at least there is one. On a desktop computer interface, the relative size of the menu bar vaguely indicates the capacity for multitasking. You can easily drag and resize each window to have the best combination of applications running. If the computer becomes slow and laggy, most people can easily induce that they should close some windows to keep everything running smoothly. These are all examples of how a human interface bridges the gap between what the user wants to do and what the technology can do.

What the chatbot interface assumes is that the user can do all of these by giving more instructions in natural language to correct or optimize the model's performance. If the AI technology we have is good enough in terms of its cognitive capabalities (see Section 3.2), this is a quite viable assumption. However, our AI is still not there and this chatbot interface only results in prompts augmented with more and longer prompts, which ultimately exceeds the context window of the model and makes it disoriented. Chatbot is an interface for the AI that can do everything (or, one that is very close to reach technology maturity), but it is an extremely ineffective and inefficient way to "operate" today's large language models.

TODO: Why augmenting chatbot instead of radical new design. Because it's langague model. Because language is the primary way of communication

References

Altman, Sam (2025). *Reflections*. URL: `https://web.archive.org/web/20250829145028/https://blog.samaltman.com/reflections`.

Bommasani, Rishi et al. (2022). *On the Opportunities and Risks of Foundation Models*. arXiv: `2108.07258 [cs.LG]`. URL: `https://arxiv.org/abs/2108.07258`.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. arXiv: `1706.03762 [cs.CL]`. URL: `https://arxiv.org/abs/1706.03762`.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: `2005.14165 [cs.CL]`. URL: `https://arxiv.org/abs/2005.14165`.

Kaplan, Jared et al. (2020). *Scaling Laws for Neural Language Models*. arXiv: `2001.08361 [cs.LG]`. URL: `https://arxiv.org/abs/2001.08361`.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (July 2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: `10.18653/v1/P19-1355`. URL: `https://aclanthology.org/P19-1355/`.

Luccioni, Sasha, Yacine Jernite, and Emma Strubell (2024). "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, pp. 85–99. ISBN: 9798400704505. DOI: `10.1145/3630106.3658542`. URL: `https://doi.org/10.1145/3630106.3658542`.

IEA (2024). *Electricity 2024*. Tech. rep. Paris: IEA. URL: `https://www.iea.org/reports/electricity-2024`.

Li, Pengfei et al. (June 2025). "Making AI Less 'Thirsty'". In: *Commun. ACM* 68.7, pp. 54–61. ISSN: 0001-0782. DOI: `10.1145/3724499`. URL: `https://doi.org/10.1145/3724499`.

Hao, Karen (2025). *Empire of AI*. Penguin Press.

Koenecke, Allison et al. (2020). “Racial disparities in automated speech recognition”. In: *Proceedings of the National Academy of Sciences* 117.14, pp. 7684–7689. DOI: 10.1073/pnas.1915768117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

Hofmann, Valentin et al. (Sept. 2024). “AI generates covertly racist decisions about people based on their dialect”. In: *Nature* 633.8028, pp. 147–154. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07856-5. URL: https://doi.org/10.1038/s41586-024-07856-5.

Wilson, Christo et al. (2021). “Building and Auditing Fair Algorithms: A Case Study in Candidate Screening”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 666–677. ISBN: 9781450383097. DOI: 10.1145/3442188.3445928. URL: https://doi.org/10.1145/3442188.3445928.

Allyn, Bobby (2024). “Scarlett Johansson says she is ’shocked, angered’ over new ChatGPT voice”. In: *NPR*. URL: https://archive.is/gtXbi.

Acemoglu, Daron (May 2024). *The Simple Macroeconomics of AI*. Working Paper 32487. National Bureau of Economic Research. DOI: 10.3386/w32487. URL: http://www.nber.org/papers/w32487.

Bursztyn, Leonardo et al. (2025). “When Product Markets Become Collective Traps: The Case of Social Media”. In: *American Economic Review* 115.12, pp. 4105–4136. ISSN: 0002-8282. DOI: 10.1257/aer.20231468. URL: https://libkey.io/10.1257/aer.20231468.

Buchanan, Ben et al. (May 2021). *Truth, Lies, and Automation: How Language Models Could Change Disinformation*. DOI: 10.51593/2021CA003.

Acemoglu, Daron and Pascual Restrepo (Dec. 2019). “The wrong kind of AI? Artificial intelligence and the future of labour demand”. In: *Cambridge Journal of Regions, Economy and Society* 13.1, pp. 25–35. ISSN: 1752-1378. DOI: 10.1093/cjres/rsz022. eprint: https://academic.oup.com/cjres/article-pdf/13/1/25/33213534/rsz022.pdf. URL: https://doi.org/10.1093/cjres/rsz022.

— (2022). “Tasks, Automation, and the Rise in U.S. Wage Inequality”. In: *Econometrica* 90.5, pp. 1973–2016. DOI: https://doi.org/10.3982/ECTA19815.

eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19815. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19815.

Moll, Benjamin, Lukasz Rachel, and Pascual Restrepo (2022). "Uneven Growth: Automation's Impact on Income and Wealth Inequality". In: *Econometrica* 90.6, pp. 2645–2683. DOI: https://doi.org/10.3982/ECTA19417. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19417. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19417.

Skare, Marinko, Beata Gavurova, and Sanja Blažević Burić (2024). "Artificial intelligence and wealth inequality: A comprehensive empirical exploration of socioeconomic implications". In: *Technology in Society* 79, p. 102719. ISSN: 0160-791X. DOI: https://doi.org/10.1016/j.techsoc.2024.102719. URL: https://www.sciencedirect.com/science/article/pii/S0160791X24002677.

Berger, Philip G. et al. (Feb. 2024). "Employer and Employee Responses to Generative AI: Early Evidence". In: DOI: http://dx.doi.org/10.2139/ssrn.4874061.

Appel, Ruth et al. (Sept. 15, 2025). *Anthropic Economic Index report: Uneven geographic and enterprise AI adoption*. URL: www.anthropic.com/research/anthropic-economic-index-september-2025-report.

Ide, Enrique and Eduard Talamàs (0). "Artificial Intelligence in the Knowledge Economy". In: *Journal of Political Economy* 0.0, pp. 000–000. DOI: 10.1086/737233. eprint: https://doi.org/10.1086/737233. URL: https://doi.org/10.1086/737233.

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

Abdalla, Mohamed and Moustafa Abdalla (2021). "The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. Virtual Event, USA: Association for Computing Machinery, pp. 287–297. ISBN: 9781450384735. DOI: 10.1145/3461702.3462563. URL: https://doi.org/10.1145/3461702.3462563.

Bender, Emily M. et al. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

Minsky, Marvin (1988). *The Society of Mind*. Simon & Schuster.

Dosovitskiy, Alexey et al. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

Wagstaff, Kiri L. (2012). “Machine learning that matters”. In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML’12. Edinburgh, Scotland: Omnipress, pp. 1851–1856. ISBN: 9781450312851.

Sambasivan, Nithya et al. (2021). ““Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI”. In: CHI ’21. Yokohama, Japan: Association for Computing Machinery. ISBN: 9781450380966. DOI: 10.1145/3411764.3445518. URL: https://doi.org/10.1145/3411764.3445518.

Marcus, Gary and Ernest Davis (2020). *Rebooting AI*. Vintage.

Bruce, Jake et al. (2024). *Genie: Generative Interactive Environments*. arXiv: 2402.15391 [cs.LG]. URL: https://arxiv.org/abs/2402.15391.

Gupta, Agrim et al. (2022). *MaskViT: Masked Visual Pre-Training for Video Prediction*. arXiv: 2206.11894 [cs.CV]. URL: https://arxiv.org/abs/2206.11894.

Lenat, Douglas B. and John Seely Brown (1984). “Why am and eurisko appear to work”. In: *Artificial Intelligence* 23.3, pp. 269–294. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(84)90016-X. URL: https://www.sciencedirect.com/science/article/pii/000437028490016X.

Watson, J.B. (1928). *Psychological Care of Infant and Child*. New York: W.W. Norton Company, Inc.

Chomsky, Noam (1957). *Syntactic Structures*. Mouton & Co.

— (1959). In: *Language* 35.1, pp. 26–58. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/411334 (visited on 11/26/2025).

Liang, Weixin et al. (Aug. 2022). “Advances, challenges and opportunities in creating data for trustworthy AI”. In: *Nature Machine Intelligence* 4.8, pp. 669–677.

ISSN: 2522-5839. DOI: 10.1038/s42256-022-00516-1. URL: https://doi.org/10.1038/s42256-022-00516-1.

Bostrom, Nick (2024). *Deep Utopia*. Washington, DC: Ideapress Publishing.