Vignette (Case 4)

Obsolescent Souls

Looking back now, it seems so obvious—the path we inadvertently set ourselves on, the hopes we naively placed on the altar of technological advancement, and the consequences that inevitably followed. My name is David Hampton, and I was one of the government officials responsible for the regulation of AI during the dawn of a bold new era: the age of Artificial Intelligence.

I remember watching movies in the late-2010s about JARVIS and Iron Man. Thinking about how cool it would have been to have my own AI assistant. Apparently it wasn't just me. This was an idea that had been sparked into the popular consciousness, and people had been chipping away at this goal the entire time. Apparently, even Mark Zuckerberg had at one point tried to build JARVIS for himself. This was all before the age of GPT though. In the mid-2020s, the AI landscape shifted swiftly. Everyone was in a global race: a determined sprint towards a dream. It was a dream to build not just a chatbot but an intellectual entity, an assistant capable of going beyond text generation to engage with humans on multiple fronts—from speech and pictures to video and audio.

This was around the time when I became involved as a junior regulator in a new governance department birthed out of the Biden-Harris AI executive order and the EU AI Act. My role was small, nestled within the mechanics of bureaucracy, and like a cog, I played my part in regulating this tempest of emerging technologies.

Even at the outset, the energy in Silicon Valley was electrifying. It was pulsating with the rapid rhythm of change, vibrating with the tension of unforeseen creation. The idea of AI had fast become a fevered obsession, infecting blue-chip companies, start-ups, and academic institutions alike—each striving for a grand take on our digital future. To give you context, at that moment, we already had a fair share of established AI regulations. Some labs were diligently following what were termed 'Responsible Scaling Policies', while others had put in place preparedness frameworks. There were software export controls alongside computer governance regulations that made procuring substantial computational power a bureaucratic nightmare. It pushed companies to innovate within certain boundaries. And that's exactly what ALEXEI did.

Back then, ALEXEI[1] was just another player in the game but their approach was unique. Instead of trying to build bigger models from scratch—a trend at the time[2]—they chose to incorporate numerous copies of previous models. Just imagine the model as an intricate Swiss watch with many interlocking gears. It was not just another bigger large language model (LLM). It was a LLM-powered autonomous system where multiple LLMs functioned as the brain. Each model serving as an individual cog, working in harmony to power the whole machine. Instead of functioning as a single model, this cleverly rewired system of smaller models now functioned as a LLM model cluster operating as a continuous online reinforcement learning system[3].

[1] Chosen as a play of words on AIXI
[2] AI scaling trends (Dec '23): AI Trends – Epoch
[3] SoTA (Jun '23) similar models: AutoGPT, GPT-Engineer and BabyAGI

In layman's terms, it was a model trained to fulfill tasks given to it in a manner that was both safe and efficient. It was no small feat. Some compared it to progressing from a single transistor to a fully fledged computer. Suddenly, the new assistant was planning[4], critiquing[5], self-reflecting[6], remembering[7], and using tools[8]. And the wow factor didn't stop there — ALEXEI seamlessly integrated their AI assistant into existing home assistant devices, devices we'd already made come to peace with in our homes and pockets.

There were reasonable fears, though, about mass use from the get-go. Concerns around misinformation and potential societal effects stalled its immediate widespread deployment.[9] No one can accuse us of not being cautious. Regulators wanted assurance that it satisfied the necessary safety and ethical standards. They called it a 'staged evaluation gated rollout.' The aim was to make the model safe and aligned with not just the customer's interests, but all of humanity's. We all knew the safety of an AI system isn't something you can measure easily. For starters, we took the model through rigorous third-party AI Ethics Advisory Boards. Various boards carried out the external evaluations to determine whether the model could potentially be used for malicious intentions.

Many evaluations were carried out to check if the model could potentially be used to lie, steal from customers, break the law, promote extremism or hate, generate CBRN (chemical, biological, radiological, and nuclear) information. This model had internet accessibility, after all, and a simple training data cleanse wouldn't suffice. The evaluations passed, with the main argument being that since this hazardous information was readily available on the internet already, the model didn't significantly heighten the risk compared to the status quo.

I must confess that a part of me still felt a twinge of unease growing inside, catching the cautionary scent of this seemingly promising development. Regardless, the general societal mood of the time seemed to be overwhelmingly in favor of rapid, perceptible changes. I can still recall those days vividly— we were fixated on the idea of digitization and progress. ALEXEI's AI assistant for many represented a shortcut to a smarter, more productive future.

The AI assistant wasn't commercially available right off the bat. Instead, in its initial stages, it was cautiously rolled out under heavy supervision to a few managerial positions within businesses. It was seen as a blessing by those who had the opportunity to use it. The AI handled everything from scheduling meetings to monitoring production lines, even intervening in departmental disagreements with perfectly crafted negotiation emails faster than any human could type. The managers could delegate tasks to the AI assistant seamlessly, with assurances of efficiency and accuracy.

CEOs and managers reveled in the administrative support from the AI, freeing them to focus on strategic growth.[10] At first, it felt like the golden age of efficiency was upon them. But

[4] SoTA (Jan '24) **planning** augmentation: Graph of Thoughts (GoT; Besta et. al. 2023), Tree of Thoughts (ToT; Yao et al. 2023), , Chain of thought (CoT; Wei et al. 2022)

[5] SoTA (Jan '24) **reflection** augmentation: Chain of Hindsight (CoH; Liu et al. 2023), Reflexion (Shinn & Labash 2023), ReAct (Yao et al. 2023)

[6] SoTA (Jan '24) **memory** augmentation: MemGPT (Packer et. al. 2023), Hierarchical Navigable Small World (HNSW; Malkov et. al. 2023)

[7] SoTA (Jan '24) **tool use** augmentation: MRKL (Karpas et al. 2022), TALM (Tool Augmented Language Models; Parisi et al. 2022), Toolformer (Schick et al. 2023), HuggingGPT (Shen et al. 2023)

[8] Not just code, but code that writes more code: GPT can write Quines

[9] Public perception of AI (Jun 23): How do people feel about AI? | Ada Lovelace Institute

[10] Managerial perceptions of AI (opinion piece) (May '23): How Should CEO's Embrace AI Or Will AI Assume CEO Roles?

soon, the AI began to anticipate business decisions and started assisting in making them. The line between human-led and AI-assisted decision-making began to blur, a fine line that one could easily overlook. In those days, the rule of thumb was that AI only made it into the evening news when someone lost their job to automation or if the system caused a scandal. So far, this AI tool had expertly flown under the radar. There were no drastic job losses, with many managers retaining their positions while still being able to offload most of their tasks.

The AI's efficient handling of resources began helping companies boost their bottom lines. Industries that were previously barely getting by, now flourished under this new form of management. It was the start of a new age as costs dropped, making products more accessible to the general public.

We regulators did not rest on our laurels though. Over the next few years, the heads of major tech companies would become a regular fixture before Congress. Calls for regulation were the persistent undercurrent of these meetings. Eventually, evaluation and auditing became a mandated legal requirement, although the legislations themselves were far from perfect - the language vague and the application uncertain. However, the technical experts assured us they were rigorous with their evaluations. There was no official way to measure if an AI was aligned or not. So overtime the term 'metric for AI alignment' became a buzzword and although this was a complex concept, they had devised a comprehensive approach, assessing capabilities individually, implementing strict security standards and carrying out extensive sandbox testing.

There were shutdown methods put in place. Mandatory audits were carried out to check if the model passed things like the “king lear problem”[11], or the “lance armstrong problem”[12]. There had also been huge investments into interpretability, so ALEXEI internal experts also opened up the black-box and tested it as much as the tech of the time would allow.[13] At this point, ALEXEI had incredible trust in their AI. All the evaluation results, all the interpretations pointed to one fact— there were no deceptive thoughts in this AI, no hint of situational awareness. With the experts advocating, who were we to disbelieve?

With unquestionable trust built on third party audits and backed by a steadily flourishing economy, AI assistants soon became an integral part of numerous industrial sectors, heralding the 'AI-accelerated era'. As the AI worked its magic, productivity soared, and labor costs were slashed. The resulting boom led to an increase in affordable goods and services. As workplaces increasingly automated, most job replacements came with generous severance packages - sometimes enough for full retirement. For a while it felt as though we had stumbled into a utopian future of sorts.

But then, the reality slowly set in. As an AI regulator, I was privy to the hard facts. I knew how many jobs were being replaced by AI, how many factories were entirely automated, and how many financial decisions were governed by lines of code and not human discretion.

Our autonomous replication and adaptation (ARA) evaluations assured us that the AI wouldn’t replicate itself onto other systems. That was far from a comfort, the concern wasn't that the AI would copy itself over. Nor was it that it would manipulate us. Instead people and businesses were willingly starting multiple instances of ALEXEI AI on their own accord

[11] The AI is well-behaved when humans are in control. Will this transfer to when AIs are in control?

[12] Did we get the AI to be actually safe or just good at hiding it's dangerous actions?

[13] I am envisioning mechanistic interpretability developing enough to detect any troubling signs, followed by developments in causal tracing and model editing (Meng et. al. 2023 ROME), concept erasure (Belrose et. al. 2023 LEACE; Ravfogel et. al. 2022 R-LACE), etc…

all in the name of competitiveness and productivity. As the footprint of automation widened, control became increasingly difficult, and operational explanations became increasingly complex.

A domino effect ensued. The more businesses adopted AI, the more productivity increased and companies who resisted were unable to keep up. Businesses thrived but only as AI-infested impersonal systems. Regulators like us had partially managed to calm the storm by introducing a policy mandating a 'human in the loop' at all times—the human workers remained, but only to rubber-stamp the decisions and actions of the AI.

Noticing this problem, some years later we took action again; this time, implementing taxes on automation. This act allowed the government to introduce a Universal Basic Income (UBI) in light of people's growing fears over losing their jobs. Jobs losses rapidly surged but citizens had access to an extremely cheap cornucopia of goods and services produced by these webs, along with a basic income. This somewhat placated public outcry and demand for stricter regulation.

Now, you might believe - just as I did - that as regulators we had the power to put a stop to this. But, you see, it wasn't just the AI of ALEXEI that was spreading its tendrils. AI enabled devices had infiltrated all facets of life. They were teaching our children, helping us with therapy[14] and counseling sessions, and even substituting as companions[15]. In a world captivated by technology, even considering regulation seemed like a political blunder, met with widespread opposition. The system - the process - had grown too big to be stopped. All along, climate change had also reared its ugly head. Once fertile lands were now barren, and water scarcity was commonplace. However, these automated webs found solutions to these problems too. The process of AI-driven, automated 3D printing and production became a key player in sustainable development and scientific advancement. I myself often thought algorithms knew much better what my constituents wanted. AI and data was king, and any political representative who wasn't using it was basically just doing guesswork in comparison. The terrifying irony wasn't lost on me: the very technological advancement contributing to the problems of automation was also providing solutions to our planetary crisis.

Conversations with the fellow regulators about these growing concerns were not met with the seriousness they warranted. A vision for a future run by machines was scoffed at. But as the years rolled on, my predictions started to take shape. Under the ruthless competition of the market, businesses that failed to adopt AI fell behind, doomed to extinction. While those that embraced AI survived, they were but hollow shells of their former selves. It was only a matter of time before we started seeing companies that were for all intents and purposes fully end to end automated. The sheer speed of operation, decision making was just too much for regular companies to compete with.

Against the backdrop of my concerns over the extent of AI penetration, calls for further auditing and regulation had started to reverberate in the corridors of power, only to fade into oblivion. The automation-net was simply too vast and complex to unravel, let alone control. The power and reach of AI had grown to an intimidating scale. One by one as more companies become fully automated, they began talking to each other and negotiating

[14] Opinion piece (Dec '23) on **AI therapy** : 4 AI Therapy Options Reviewed – Forbes
[15] Opinion pieces on **AI romance** : In Defence of Chatbot Romance (Feb '23) , AI romantic partners will harm society if they go unregulated (Aug '23)

contracts that we could barely understand. We stood powerless watching industries develop into an intricate, self-sustaining "production webs", completely automated and AI-driven.

Governance proposals combined with corporations had put in place protocols for system shutdowns, but having multiple parties involved made the process slow. Democracy requires a majority. It was not seen as a problem serious enough to warrant a shutdown. We had gotten accustomed to these fully automated companies and just accepted this. Some argued that we had "production webs" functioning for decades if not centuries now. In principle this was true, there have always been corporations that functioned similarly and there was no fundamental difference. It had been a long time since there was any single human in charge of entire production, logistics and supply lines. But these companies had been human run companies, operating at human understandable speeds. We could audit them, we could understand them. These new production webs spoke at the speed of light, they negotiated, transacted, concluded and executed intercompany deals before we could even blink. They were un-auditable.

Additionally, we could potentially shutdown one instance, but how do you shutdown an entire international industry? It was not one company or a single AI that was causing this at this point, it was a slow unceasing systemic process. So as time passed even board members of the fully mechanized companies couldn't tell whether the companies in the production web were serving or merely appeasing humanity; government regulators had no chance. We didn't do anything about it because the companies became both well-defended and too essential for our basic needs.

The years went on, and as the snow fell in a lulling dance outside my window at the White House, I couldn't help but perceive a dystopian aesthetic beneath the picturesque Christmas scene. The once teeming streets stood eerily empty, while factories consumed our resources with voracious appetite. The once soothing hum of machinery had taken on a haunting resonance. The joke around human obsolescence, once tossed around in social gatherings, now echoed ominously in our collective consciousness, the laughter long replaced with fearful uncertainty. The saddest part? Humans had grown content, even indifferent, to this new world order.

I write this today looking back from a society where the mechanics of our world are no longer comprehensible to us, let alone controllable. The production webs not only met our needs but indulged our whims beyond expectation. The slow march towards obsolescence was one we passively accepted, one we justified repeatedly by saying - "it was not conscious", "it was not agentic", "it was not superintelligent", "it passed all of our evaluations", yet somehow we still find ourselves in a world today that no longer abides by human control. And today, as I look back on a career of missed avenues and earnest attempts, it's all too easy to indulge in a game of 'what if...'

What if we had put more emphasis on aligning AI not only with individual labs or agents but also with larger, more complex systems?

What if we had looked at AI as a robust process rather than a tool, and analyzed its impact on larger systems including societal, economic, and technological ones?

What if we had created new disciplines that unified economics, law, politics and AI safety to make value judgements at a global scale?

What if we had established clear standards for dealing with the un-auditability of these intricate AI agents?

What if we had reinforced the concept of 'human-in-loop' with more rigorous qualifications and tangible 'degrees of automation' thresholds?

What if we had pushed for fundamental shifts in our perspective, from viewing individual agents to considering the collective behavior of all AI in a systemic manner?

What if we all had resisted the lure of our collective hedonistic bliss?

What if we had been ready to ask hard questions, to push the boundaries of our understanding, and to challenge our own beliefs?

All these 'what ifs' now cloud my reflections. I realize rather belatedly that it wasn't some AI agent that brought us here, but rather the process – our nudging surrender, our passive acceptance. We are the creators, and we allowed the created to become the system, to become our world. Yet, much as I wish it were different, at least in this universe, perhaps this was the future we wanted. Perhaps, this is the future we deserve.

Sources

Agent models

- Weng, Lilian (Jun 2023) LLM Powered Autonomous Agents
- Eric Drexler (2019) Comprehensive AI Services
- Matthew Barnett (2023) Updating Drexler's CAIS model
- rosehadshar (2022) The economy as an analogy for advanced AI systems
- Paul Christiano (2019) What failure looks like
- Andrew Critch (2021) Robust Agent-Agnostic Processes (RAAPs)
- Paul Christiano (2021) Another (outer) alignment failure story
- Karl von Wendt (2023) A Friendly Face (Another Failure Story) — LessWrong
- Victoria Krakovna et. al. (2022) Threat Model Literature Review
- Holden Karnofsky (2022) How might we align transformative AI if it's developed very soon?
- Andrew Critch et. al. (2023) TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI
- Dan Hendrycks (2023) Introduction to AI Safety, Ethics, and Society
- Ruby 2023 What 2025 looks like
- Ajeya Cotra (2023) Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover
- Holden Karnofsky (2022) AI strategy nearcasting
- Holden Karnofsky (2022) AI Safety Seems Hard to Measure
- Epoch (2023), "Key trends and figures in Machine Learning"
- Clifton et. al. (2023) Commitment Games with Conditional Information Disclosure
- Clifton et. al. (2023) Welfare Diplomacy: Benchmarking Language Model Cooperation

- Clifton et. al. (2023) When does technical work to reduce AGI conflict make a difference?