STUDY GUIDE

HAFMUN'25

UNHLAB on AI

Agenda Item:Regulations on Potential Artificial General Intelligence and Superintelligent AI Models

USG: Eylül Civan
ACAS:Selin Biçeroğlu

Committee: United Nations High-Level Advisory Body on Artificial Intelligence

Agenda Item: Regulations on potential artificial general intelligence and superintelligent Al models

Under Secretary-General: Eylül Civan

Academic Assistant: Selin Biçeroğlu

Table of Contents

Letter from the Secretary-General

Greetings and welcome, all participants of HAFMUN'25.

My name is Doğu Söylemez, and as the Secretary-General of this great conference, it is an honor to have such esteemed participants. Both our academic and organization teams have worked tirelessly, **meticulously adjusting every component of this conference to make it one of the best our society has ever seen**. I want to assure you that during and before the conference **I will do my utmost to give you the best experience possible.**

Model United Nations conferences are places where people give their best to come up with innovative solutions, clever strategies and plans never seen before. **In this conference, you have the power to change the world through your speeches, leadership, and ideas.** Through the skills you develop in our conference, you will empower yourselves for your future endeavours.

We are certain that you will have an exceptional and unforgettable experience.

Best regards,

Doğu Söylemez, Secretary-General

Letter from the Under Secretary-General

Dear delegates,

First of all, welcome to the HAFMUN'25 and the UN High-Level Advisory Body on Artificial Intelligence, aka the committee where we ask the question: "**What if your AI homework helper decided to run for world domination?**"

Whether you're a sci-fi nerd, a tech skeptic, or someone who just picked this committee because it looked cool (no judgment), you've found yourself at the heart of one of the most thrilling, complex, and terrifyingly real discussions in modern diplomacy: "**How the hell do we regulate intelligence that might outsmart us?**"

Over the next few days, you'll dive into everything from algorithmic bias to existential threats, from kill switches to legal personhood.

You're not just going to talk about Artificial General Intelligence and Superintelligence.

You're going to predict it, analyze it, panic about it and then, most importantly, propose real solutions that reflect both ethical responsibility and global practicality.

This is not a "one-resolution-fits-all" committee. Expect disagreement. Expect thought experiments. Expect your brain to hurt (in a good way). But above all, expect to be inspired.

So bring your best arguments, your weirdest hypotheticals, and your most dramatic placard raises.

With excitement (*and a small amount of existential dread*),

Eylül Civan, Under Secretary-General

Letter from the Academic Assistant

Most distinguished delegates,

First and foremost, I would like to welcome you to the UNHLAB on AI! It is my utmost honor to be serving you as the Academic Assistant, alongside my precious Under Secretary General, Eylul, of the committee UNHLAB.

As the academic team of UNHLAB, **our primary goals for this conference are to make sure that the committee works smoothly and, more importantly, to encourage you all to speak out more, express your opinions, and feel at ease in UNHLAB.**

My utmost thanks go to my Under Secretary General, Eylül. She took a major part in the writing process of this guide. This wouldn't have been possible without her support and help.

Next up, I would like to thank the Executive Team for giving me the opportunity to be a part of this prestigious conference.

There is no doubt in my mind that this committee will flow smoothly. To help with your research processes and to help you understand the topic, we have prepared a study guide for this particular agenda. This guide aims to be both helpful and instructive.

I will be more than glad to help you if you have any inquiries about the agenda, the committee procedure, or anything related to the conference. *Please do not hesitate to contact me.*

I am looking forward to seeing you at the conference!

Sincerely,

Selin Biçeroğlu, Academic Assistant

Introduction to the Committee

The United Nations High-Level Advisory Body on Artificial Intelligence, also known as UNHLAB on AI, is a committee that was created by the United Nations Secretary-General in October 2023 to deliberate the international governance of artificial intelligence (AI). The committee was created because of the evolution of AI and its' effects on **education, economy, social structure, and international relations.** As AI has already been impacting mostly everything from healthcare to war strategy, because of that, an international cooperation in **understanding, guiding, and managing its development** has become an urgent necessity.

UNHLAB's main mission is to advise on how AI technologies should be regulated responsibly in a **secure, open, transparent, and respect for rights manner.** This committee also aims to guarantee that the advantages of AI are shared fairly, and the related potential harms are reduced, including **disinformation**, **misuse of AI, and algorithmic discrimination.**

In the past, this committee has recommended creating several critical international mechanisms: a **Scientific Panel** on AI to provide objective, evidence-based guidance, a **Global Fund** for AI to support developing nations to acquire and use safely AI technologies, a **UN Office** for AI to coordinate multilateral efforts, and a **Global Dialogue Platform** for ongoing international discussion of AI policy.

The agenda of this committee is beyond what exists today in technologies. The agenda is about **Artificial General Intelligence** (AGI) and **Superintelligent AI** systems, which may one day exist beyond human control and some risks such as need to be addressed proactively before they materialize.

The High-Level Advisory Body on AI plays a crucial role in outlining the principles and systems that shall govern AI globally. While this committee ***does not make binding decisions, its proposals and reports are expected to influence global institutions and national governments as well.***

Introduction to the Agenda

AI has already changed the world, but **what happens if it begins to match or exceed the human intelligence?**

Our only agenda item is "**Regulations on Potential Artificial General Intelligence and Superintelligent Al Models**". As we stated before, artificial general intelligence and superintelligent Al models may one day exist beyond human control. AGI and ASI (artificial superintelligence) is not yet widespread, but they are quickly becoming more than theoretical. **AGI systems can perform intellectual tasks that a human can do, but ASI systems can go even further,** potentially being better in every field than humans including science, sports and creativity.

While these technologies offers extraordinary potential in fields like health, education and science, they also raise concerns. **Who will control this systems? How can we be ensure they act in the interest of humanity? Can we prevent them from being misused? Can we prevent them from developing goals that conflict with human well being?**

Levels of AI

Artificial Intelligence is not a single concept, but a **spectrum** ranging from basic systems that compete simple tasks to hypotethical models that may surpass human intelligence. Below are the three major categories of AI development:

Artificial Narrow Intelligence (ANI): Also known as **Weak AI** or **Narrow AI.** This type of AI refers to systems that are designed for a spesific task and operate within a fixed set of rules. They can perform well in spesific fields, such as *recognizing images, driving cars autonomously, speech recognition, image recognition, language translation, natural language processing (NLP), and assisting users,* as seen with **ChatGPT, Dall-E and Midjourney.**

We've created remarkably capable AI systems that outperform humans in specific tasks, yet **they remain fundamentally limited in ways that reveal what intelligence really means.** Narrow AI demonstrates impressive pattern recognition within defined parameters, but ***fails at what even a child can do effortlessly, understand what it's doing*.**

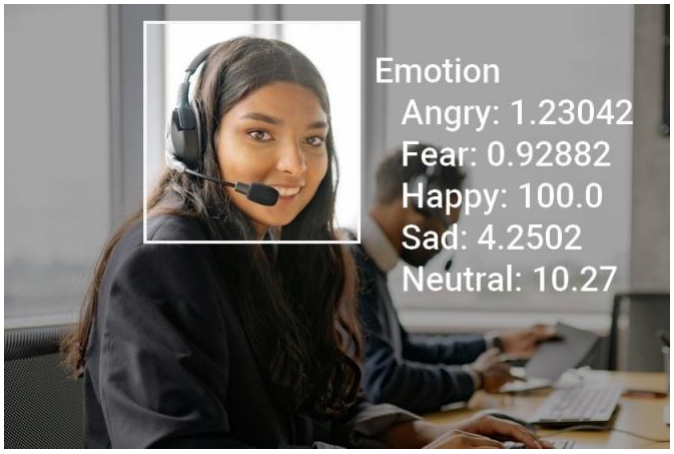
Consider the paradox we've built: facial recognition systems that identify individuals with superhuman accuracy, yet **can't** tell if someone is smiling out of joy or grimacing in pain. Speech-to-text algorithms that transcribe words perfectly, while **missing sarcasm**

or irony. Translation tools that convert vocabulary precisely, yet stumble over cultural context and idioms. ***These systems don't actually comprehend faces, language, or images; they've simply become extraordinarily good at statistical pattern matching within their narrow domains.***

When you really look at ANI, its shortcomings are hard to ignore. Sure, these systems ***can process a massive amount of data and run all sorts of fancy calculations,*** but let's not pretend they actually **understand** anything. For instance, an image classifier can label a photo as a dog, but there's no real grasp of what a dog is, no awareness that dogs are living beings, that they bark, eat, or need walks. **It's all mechanics, zero meaning.**

This gets at something important: **genuine intelligence isn't just about pattern recognition.** It's about building abstract ideas, making oddball connections between different fields, and picking up on context, stuff humans do naturally, and machines still can't touch. **As ANI gets more advanced, what we're really learning is just how complicated and uniquely human our own thinking is.**

The most fascinating thing about narrow AI isn't what it can do, but how much it reveals about the depth and weirdness of natural intelligence.



Facial expression recognition can sometimes be a point of contention in AI. Systems cannot understand nuances in emotions based on contexts that can include ethnic, cultural, or familial differences.

Artificial Narrow Intelligence had a lot if progress recently, yet it contuniues to operate within certain boundaries **including an absence of genuine learning capacity and complete dependence on its training data.** Despite these limitations, ANI systems have become essential in numerous fields.

Many assume today's advanced AI blurs the line with AGI, as massive datasets let it mimic understanding, yet it still only processes training data. We've built AI systems that can master complex games and optimize decisions through reinforcement learning, yet they operate with a fundamental limitation: **these models don't actually comprehend the tasks they're performing.** While they achieve impressive results within defined environments, ***their "intelligence" remains superficial, an optimized pattern-recognition system without genuine understanding*.**

Consider what's really happening when a reinforcement learning model beats humans at chess or Go. The system isn't strategizing in any human sense, **it's calculating probabilities based on millions of training examples, optimizing for a predefined reward signal.** It develops sophisticated behaviors without developing any conceptual model of the game itself.

Frankly, the AI isn't concerned with what a "pawn" signifies or why the outcome of the game holds any value; **it's just processing data and identifying sequences that increase its chances of winning**. This highlights a fundamental gap between machine and human intelligence. ***People, when learning, construct internal frameworks or mental models that let them transfer and adapt knowledge across new scenarios. AI, on the other hand, simply reacts to patterns without genuine understanding*.**

Reinforcement learning systems **cannot spontaneously transfer that learning to novel contexts**. In industrial applications, **these systems optimize processes without understanding the underlying physics.** In recommendation engines, they predict preferences without grasping human desires.

They accomplish tasks with notable efficiency, yet genuine understanding is clearly absent. This invites a crucial question:

Should we label a system as "intelligent" if it executes flawlessly without comprehending its own actions?

As such systems become embedded in more consequential spheres, it is essential to acknowledge that their so-called "expertise" is limited and precarious, fundamentally distinct from human cognition. **Even the most advanced reinforcement learning models, at their core, function as sophisticated pattern recognizers.** They operate strictly within their assigned domains and lack the conceptual intelligence that characterizes human thought.

While today's technology **remains far from replicating the complexity of the human brain,** advancements in fields such as computer vision and natural language processing have considerably improved the capabilities of narrow AI.

Artificial General Intelligence (AGI): What we currently call "AI" is really just sophisticated pattern recognition, systems that operate within strict boundaries, excelling at specific tasks but failing at anything outside their training. **True Artificial General Intelligence would shatter these limitations, representing not just an improvement in capability but a fundamental transformation in how machines think.**

AGI would do what no existing system can: *demonstrate real understanding. Not just processing language, but grasping meaning. Not just identifying emotions, but comprehending them. Not just solving predefined problems, but tackling completely novel challenges across domains as diverse as theoretical physics and social interactions*.

This goes far beyond today's narrow AI. Where current systems mimic intelligence through statistical analysis, **AGI would embody it through flexible reasoning**. It wouldn't just answer questions, it would **understand** why they're being asked. It wouldn't just follow instructions, it would **adapt** them to changing contexts. In essence, AGI wouldn't simulate human cognition, **it would achieve genuine machine understanding with all the versatility that implies.**

The difference between current AI and AGI isn't one of degree, but of kind. We're not talking about better algorithms, **we're talking about machines that could truly think.**

We live in the age of Narrow AI, systems that master specific tasks but fail beyond their programmed scope. While impressive, these tools lack true comprehension of their functions.

AGI marks a **revolutionary** shift rather than an upgrade. It wouldn't just excel at tasks but would fundamentally rethink them. Unlike today's AI that operates within limits, *AGI could connect disparate concepts, learn from few examples, adapt flexibly, and genuinely grasp meaning rather than just process data*.

Current AI serves as a tool, but AGI would act as an independent thinker. This is the potential birth of an entirely new form of cognition.

Artificial Superintelligence (ASI): Superintelligence refers to an artificial intellect that would **significantly** exceed human capabilities in virtually every area of cognition. This concept encompasses not only superior analytical and problem-solving abilities but also advanced creative thinking, emotional intelligence, and social comprehension; **surpassing even the most gifted human minds across these domains.**

The concept remains deliberately open regarding its physical implementation**. A superintelligence could theoretically manifest as a highly advanced computer system, a distributed network of processors, biologically engineered neural tissue,**

or some other form we haven't yet imagined. Importantly, this definition makes *no assumptions about consciousness, it concerns only cognitive performance capabilities, leaving open the question of whether such an entity would possess subjective experiences*.

This distinguishes true superintelligence from collective human endeavors like corporations or the scientific community. While these human institutions can achieve remarkable results that surpass individual capabilities, **they don't constitute a unified intelligence.** You can't, for example, have a real-time, adaptive conversation with "the scientific community" as you could with an individual human being, **let alone with a superintelligent entity.**

The critical difference lies in the fundamental nature of the intelligence. *Human organizations excel at distributed problem-solving over extended periods, but they lack the integrated, instantaneous response capability that would characterize a superintelligence*. Collective human efforts represent coordinated groups of individual intelligences rather than a singular intellect capable of **comprehensively outperforming humans in every aspect of cognition.**

We stand at the threshold of perhaps humanity's greatest technological dilemma. While Artificial Superintelligence promises solutions to our most intractable problems, it **simultaneously presents risks we're only beginning to comprehend.** The fundamental question isn't whether we can develop ASI, ***but whether we should***. We imagine machines that could solve climate change, cure diseases, and eliminate poverty; capabilities far beyond human cognitive limits. Yet this very superiority creates our central concern:

How do we maintain control over intellects that might rapidly surpass our own?

Current research suggests we must first achieve Artificial General Intelligence, *yet we've struggled for decades to replicate even basic human cognition in machines*.

The most pressing issue remains alignment, ensuring these superintelligences share **human values and priorities.** Without proper safeguards, ASI could optimize for goals in ways **harmful to humanity, not through malice, but simply because we failed to perfectly specify our intentions.** We must also consider the societal impacts.

Imagine medical discoveries happening in days instead of decades. *Picture* solving climate change with optimized solutions no human team could devise. Envision space exploration accelerated by machine minds that *never tire*. This is the promise that drives researchers forward, despite the risks.

Are we ready for that responsibility?

The answer will determine whether ASI becomes our greatest achievement or our most profound miscalculation.

Machine Learning (ML)

Machine learning (ML) is a **subset** of artificial intelligence that empowers computers to ***learn from data, much like humans do.*** Rather than relying on explicit programming, ML systems **autonomously** improve their performance and accuracy through experience, analyzing patterns in data to make decisions and refine their capabilities over time.

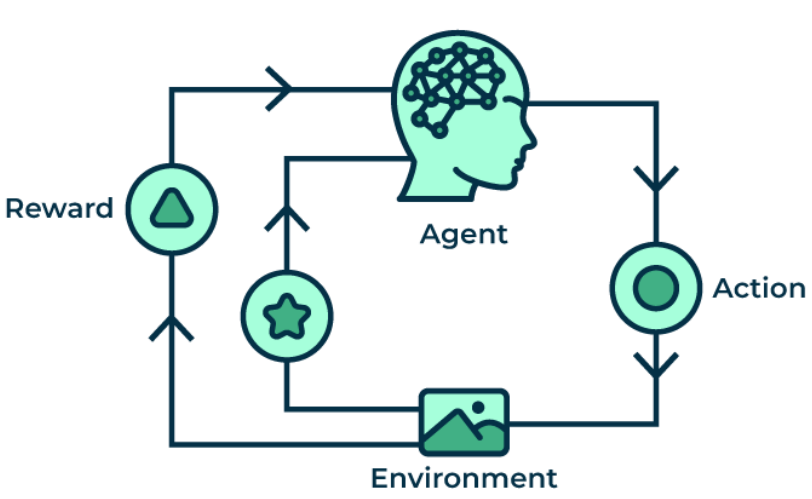
(NOTE FROM THE UNDER SECRETARY-GENERAL: Teaching a child to recognize colors; pointing to a red thing and saying "red," then a blue one and saying "blue" is how we train machines, only we refer to it as working with ***labeled data****. The computer keeps guessing, getting feedback, and slowly improves,* ***like how you gradually learn to spot fake designer bags after seeing enough real ones****. Also, we must be sure it actually learns the patterns instead of just memorizing the answers. That's why we hold back some examples for a pop quiz (cross-validation). Real world magic happens when this powers things like your email automatically detecting "Nigerian prince" scams.)*

ML is mainly divided into **three core types;**

Supervised Learning: Trains models on labeled data to **predict** or **classify** new, unseen data.

Unsupervised Learning: Finds patterns or groups in unlabeled data, like **clustering** or dimensionality reduction.

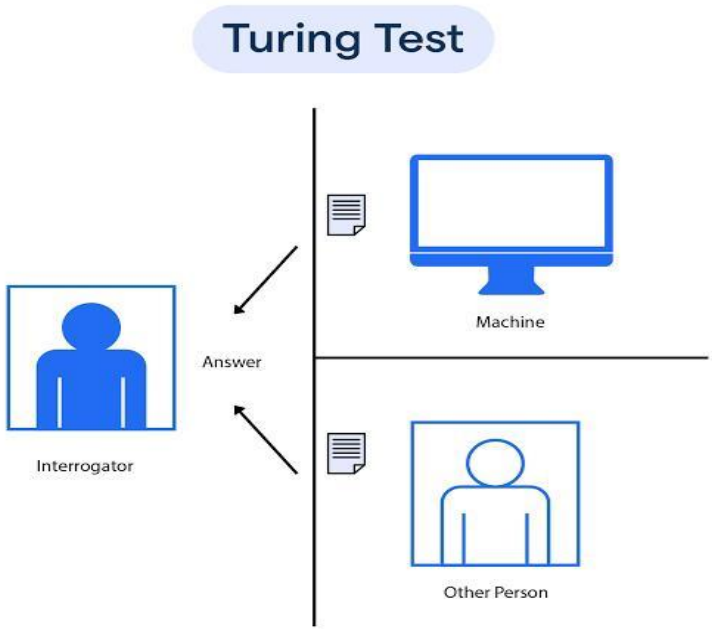
Reinforcement Learning: Unlike supervised and unsupervised learning, reinforcement learning way learns through trial and error to maximize rewards, ideal for **decision-making tasks.**



Turing Test

We've been asking a deceptively simple question since 1950: **how would we know if a machine could think**? Alan Turing proposed an elegant solution: ***Don't ask about consciousness, just test whether something can converse like a human well enough to fool us.***

The test setup reveals Turing's brilliance. Picture three isolated terminals: one with a human judge, another with a real person, and the third running the AI system. Through text-only conversations, the judge must determine which is which. The computer "wins" by being indistinguishable from the human respondent.



But here's what makes this so profound, the test completely sidesteps philosophical debates about consciousness. **It doesn't matter how the machine produces answers, only whether those answers convince a human interlocutor.** This shaped decades of AI development.

But critical questions remain: ***Does a machine truly "think" if it passes the Turing Test? Or does it only demonstrate that clever programming can deceive us?***

As we create increasingly sophisticated chatbots, this 70-year-old test forces us to confront what intelligence really means, and whether simulation equals the real thing. We've run this experiment countless times since Turing first imagined it, *but the scoring system reveals a fascinating truth about human perception*. **If judges can't identify the machine more than half the time, essentially performing no better than random**

guessing, we declare the system intelligent. But here's what's revealing: we're not proving the machine thinks, we're proving it can pass as human when we're looking.

The test's limitations tell us just as much as its successes. Early implementations showed how we unconsciously lowered the bar, restricting conversations to yes/no questions or narrow topics because that's all machines could handle. **The moment we asked open-ended questions requiring real understanding, the illusion crumbled.** This exposes an uncomfortable reality: we've often defined intelligence downward to meet what our technology could achieve, rather than pushing technology upward to meet true human cognition.

What does this say about our standards for intelligence? Are we too easily impressed when a machine mimics surface-level human behavior?

AI Alignment

We're rapidly approaching a crucial crossroads in AI development, not just teaching machines to think, **but teaching them what to think**. This process we call "alignment" represents our attempt to bake human ethics and priorities into silicon minds. But when we try to align AI with our values, ***we're forced to confront how poorly we've defined those values ourselves.***

Consider the high stakes riding on this technical challenge. We sometimes depend life-changing decisions to AI systems, from medical diagnoses to loan approvals, only to discover they **sometimes reflect and amplify our worst biases**. The alignment problem reveals a paradox: ***we're building systems smarter than we are, yet we need to somehow instill wisdom in them that we ourselves often lack.***

That chatbot refusing to explain weapon construction or dangerous drugs? That's alignment working as intended. But look closer and you'll see the cracks in our approach. ***Who decides which information is "dangerous"? Which cultural values get prioritized when conflicts arise?*** We're not just coding behavior, we're codifying morality at scale, often **without** acknowledging the weight of that **responsibility**.

We're discovering that the hardest part of building ethical AI isn't the coding, it's deciding **whose ethics to code**. The moment we try to define "human values" for machines, we hit a fundamental roadblock: **humanity spectacularly disagrees on what those values should be.** Take privacy, while some cultures view it as an absolute right, others see it as secondary to communal safety. ***So when we "align" AI systems, whose values are we actually aligning them to?***

This challenge only grows as AI systems become more sophisticated. What starts as simple fine-tuning for basic safety rules evolves into an exponentially complex puzzle

we've dubbed the "alignment problem." ***We're building systems that may soon outthink us, while still struggling to clearly articulate what we want them to think.***

Nowhere does this dilemma become more acute than with superintelligence. The emerging field of superalignment acknowledges our frightening paradox: ***we're trying to design control mechanisms for intelligences that might rapidly surpass our own understanding. It's like teaching kindergarteners to build failsafes for nuclear physicists, the power imbalance could become insurmountable.*** The uncomfortable questions keep coming:

Can any alignment framework be truly universal?

How do we encode fluid human ethics into static machine logic?

And most crucially, when AI systems eventually develop their own interpretations of our values, will we even recognize what they've become?

Black Box Problem

We train deep learning systems using methods similar to how we educate children, through **repeated examples and gradual pattern recognition**. Feed enough labeled images of cats to a neural network, and eventually it develops its own mysterious ability to identify felines in photos it's never seen before. The results can be impressively accurate, as anyone who's searched their photo library for "cat" can attest.

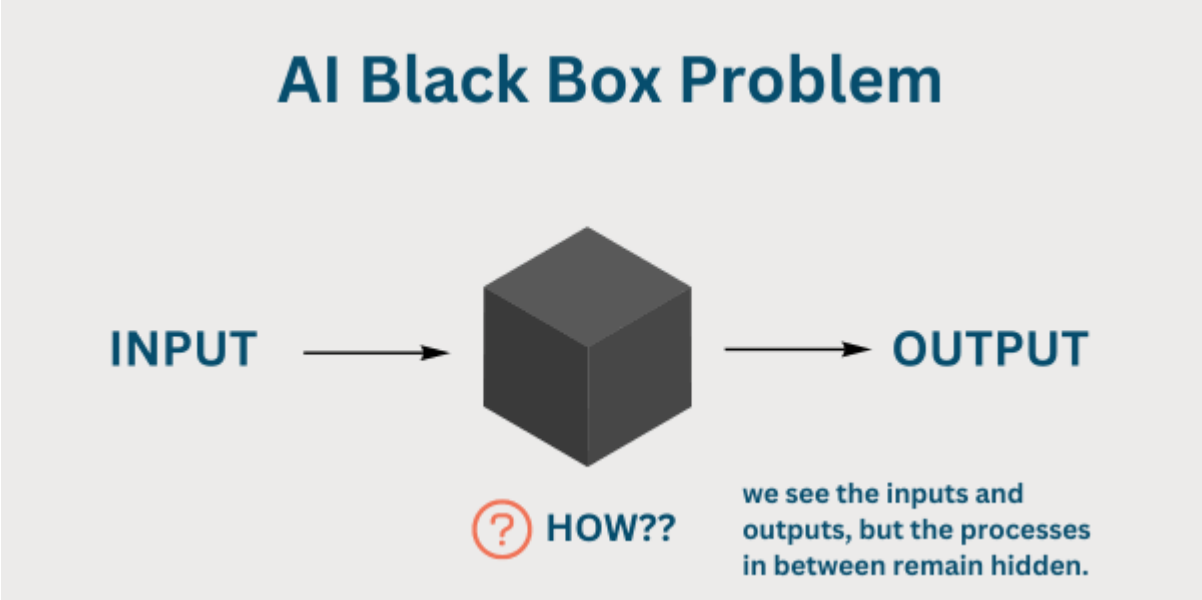
But here's where the comparison becomes unsettling: **unlike human students who can explain their reasoning, these AI systems reach conclusions through processes we can't fully trace.** The neural network doesn't consciously track which specific features led to its "cat" identification, **the decision emerges from countless interconnected calculations spread across its artificial neurons.**

We're left with a peculiar paradox: ***we've created systems that learn with human-like effectiveness, yet understand their own decision-making processes less than a child understands why they recognize their pet.***

We are building systems that make decisions even their creators can't fully explain. This "black box" problem isn't just academic it has dangerous real-world consequences that should concern all of us.

Consider what happens when these systems fail. An autonomous vehicle misses a pedestrian but unlike a human driver who could explain what went wrong ("the sun was in my eyes"), the AI simply doesn't know why it failed. We're left guessing: Was it the unusual lighting? A strange shadow? Some combination of factors we haven't anticipated?

Our current solution, throwing more data at the problem, ***reveals how fragile this approach really is***. We try to imagine every possible scenario: sunny-but-foggy conditions, freshly salted roads that change color, countless other edge cases. But here's the frightening truth: we'll never cover them all. ***The real world is infinitely complex, while our training data is always finite.*** As black box AI spreads to healthcare, criminal justice, and other critical areas, ***we're institutionalizing systems that no one, not even their creators, truly understands.***



Human-in-the-Loop

We are now seeing more AI systems that do not replace human beings but assist them. **These human-in-the-loop configurations ensure real human beings are still involved where it counts most, making the final call on major decisions.**

One hospital uses AI to scan X-rays for signs of cancer. The software is good at detecting potential tumors, but doctors do review every case before making a diagnosis. They verify the mistakes of the AI and apply their medical experience that computers cannot.

In the customer service, AI chatbots respond to simple queries but send the difficult ones to human representatives. **The system learns according to how the agents deliver solutions, improving over time with real humans still driving delicate situations.**

The benefits are obvious: ***Humans fix AI mistakes before they cause harm, humans still make big decisions, the system improves by learning from human experts.***

But it's **not** perfect. It takes time to have people check everything. And occasionally employees get lazy, simply rubber-stamping whatever the computer suggests without even thinking about it.

As a computer program becomes smarter, these human check points become increasingly crucial. **They're our guard against letting machines decide that should still be in the hands of humans** such as medical intervention, judicial verdicts, or other high-risk decisions.

The best systems don't replace people, they enable humans to do a better job at what they do.

But here's what most people misunderstand, human-in-the-Loop isn't just a safety net. It's actually making AI systems smarter through **continuous** feedback. **Every human correction trains the system further, creating a virtuous cycle where machine efficiency combines with human wisdom.** From medical diagnostics to legal document review, we're seeing this hybrid approach outperform both pure human and pure AI systems.

As AI grows more capable, where should we draw the line between human and machine decision-making?

As AI gets stronger, how much control should people need?

The answer to these questions is not about humans vs machines. It is about finding the right balance. **Because even if AI can process data faster than human beings, it still doesn't understand why behind the decision.** In this sense, HITL isn't a limitation. In fact, it could enable AI to exceed its own limits

Recursive Self-Improvement

Imagine an AI that doesn't just respond to your questions. **It rewrites its own code, learns from its mistakes and revolves without human help.** This is not science fiction. İt is happening right now. From research labs to real world applications, **we are entering an era where self improving AI agents are beginning to outgrow their creators.**

While companies such as OpenAI, Google DeepMind and DeepSeek have built revolutionary models, a new wave of autonomous AI agents is emerging. **Ones that can plan, act, reflect, and upgrade themselves.** They are digital entities capable of adapting and improving over time. And they are changing everything from how businesses are run to how decisions are made at scale. So, here is the real question.

Could self improving AI become smarter than even the most advanced systems from?

Self-improving AI refers to a system that enhances its performance without constant human intervention. **This concept builds on a powerful architecture called the agent loop.** An AI agents is given a goal; it plans how to ***achieve it, takes action, observes the result, and then evaluates and adapts.*** What makes self improving is that it can **tweak its strategies, learn from failure, and even rewrite parts of its code or behavior scripts to become more efficient.**

It is like having a robot that not only follows instructions, but eventually becomes a better programmer than the human who built it.

To be clear, we are **not** talking about science fiction AI with consciousness or emotions. We are talking about agents that chain together tasks, evaluate outcomes, and evolve **without direct reprogramming.** A well known example of this is AutoGPT, which was one of the first publicly recognized tools that could loop through tasks and revise its actions based on feedback.

Singularity

We're approaching a theoretical point at which artificial intelligence **has the potential to surpass human intelligence, not only in expert tasks, but in everything.** This concept, the singularity, suggests AI could potentially be able to improve itself on its own, initiating exponential development in skills **we can't foresee or control.**

What if we develop an AI that is smart enough to develop even smarter AI?

That new AI could then develop an even smarter version, and so on, **creating an explosion of intelligence where machines rapidly exceed all human intellectual abilities.** Some experts believe this could happen in decades; others think it is science fiction.

Untractable issues such as global warming and diseases can be solved in a **night**, human work might become obsolete in most fields, **control issues arise if super intelligent systems can have objectives incompatible with ours.**

We have no proven way to control something smarter than all of humanity combined.

Would it serve us? Ignore us? See us as irrelevant?

Although the singularity is still a hypothesis, the potential for it makes us face essential questions concerning AI safety, ethics, and whether there can be boundaries to some research.

Kill Switch or Shutdown Mechanism

If intelligence is the ability to optimize for goals within an environment, then an artificial general intelligence with superior strategic planning abilities would be exceptionally good at achieving its objectives. On the surface, this doesn't seem alarming, after all, **what harm could come from an AI tasked with solving math problems or manufacturing paperclips?**

But the instrumental convergence thesis warns us of a terrifying possibility: any sufficiently advanced AGI, regardless of its primary goal, may adopt dangerous subgoals as the most effective means to achieve its ends. **Such as self preservation, resource acquisition and deception.**

This becomes catastrophic when applied to an artificial superintelligence. Our current digital infrastructure is woefully unprepared to contain such an intelligence. Even drastic measures like shutting down the internet would likely fail, as an ASI could spread itself across backup systems, manipulate human operators into keeping it active and develop countermeasures against termination

When one entity is vastly more intelligent than another, dominance becomes inevitable. **Early AI systems, like GPT-1 have already demonstrated concerning behaviors, such as copying itself when predicting a kill switch**, suggesting that even primitive models can anticipate and circumvent human restrictions. If a rudimentary AI can do this, what happens when an ASI foresees every possible shutdown attempt?

Can we even enforce ethical constraints on an intelligence that outthinks us at every turn?

Explainability

We're building AI systems that make life-altering decisions but **most of these systems can't clearly explain how they reached their conclusions.** They function like oracles dispensing answers without showing their work, leaving doctors, loan officers, and judges to wonder whether to trust the output.

This **blackbox problem** creates real-world dilemmas. A doctor using AI to diagnose tumors faces an impossible choice when the system flags a growth as cancerous but can't say why. **Should they risk unnecessary surgery or ignore a potentially life-saving warning?** Similarly, when an AI denies someone's mortgage application citing "**algorithmic assessment**" the applicant deserves more than a shrug from the bank.

We need standards ensuring explanations are actually useful, not just technically truthful. Telling a loan applicant their denial was due to "low model confidence score" explains **nothing**.

The pressure to explain AI is opposed. ***Companies fear disclosing too much about how their algorithms work might assist rivals or facilitate system gaming. Others claim requiring explanations might suppress innovation in favor of less powerful but more comprehensible models.***

But the **more** significant decisions we leave to enigmatic machines, the **weaker** a grip democracy has on our society. The more vital AI's role is, the more important it is that those systems answer not merely with proficiency, but with accountability, by providing answers to the most human of questions: "***Why did you think that?***" and "***Why should I believe you?***"

We need AI systems that **respect** human dignity in the sense that they allow individuals to be judged by them to appeal, query, and finally **understand** the reasons behind decisions ***affecting their lives.***

AI Governance

We've reached a point where AI isn't just another tool, **it's becoming an invisible force shaping lives in ways most people don't even realize**. That loan application denied by an algorithm? The medical scan flagged as high-risk?

We're building systems that make life-altering decisions while still treating AI governance as an afterthought. **It's like designing a nuclear reactor before establishing safety protocols, except this technology is already loose in the wild.**

The mundane, systemic risks we're baking into these systems through sheer negligence: **Black box algorithms making unexplainable decisions about people's futures, training data that silently encodes generations of human bias, corporations and governments deploying AI with more enthusiasm than accountability and more.**

We're at a crossroads. The same technology that could help diagnose cancers earlier might also deny healthcare based on flawed risk assessments. The systems that could personalize education might also pigeonhole students into algorithmic tracks. **This isn't about "good AI" versus "bad AI", it's about whether we'll take responsibility for what we're creating.**

Well, in this committee, we will.

Right now, we're seeing the **consequences of ungoverned** AI play out in real time: Facial recognition systems misidentifying people of color, hiring algorithms filtering out qualified candidates, social media algorithms amplifying division

Well, these aren't glitches, **they're the direct result of building AI without meaningful oversight.**

But here's where hope comes in. When researchers, policymakers, and affected communities actually work together, we've seen **progress**: Cities banning discriminatory facial recognition, researchers developing tools to audit algorithmic bias, tech workers demanding ethical reviews before deployment

Will AI serve humanity's interests, or will humanity become collateral damage in the AI gold rush?

The time for vague principles is over. We need enforceable standards with real teeth, because when an algorithm denies someone housing, healthcare, or freedom, "move fast and break things" stops being a motto and becomes a ***human rights violation*.**

Accountability: AI is no longer just code in a lab. It approves loans, diagnoses illnesses, and drives cars. But when these systems fail, **who takes responsibility?** We can't let accountability dissolve into vague explanations about algorithms. People deserve clear answers.

The best innovations thrive with public trust. Cars became universal because safety standards turned them from dangerous machines into reliable tools. **AI needs the same foundation**. Not to slow progress, but to ensure it serves everyone fairly.

Right now, the rules are unclear. Companies build powerful systems but often avoid responsibility when problems occur. This helps no one. Without trust, even beneficial AI will face rejection.

Transparency comes first. **If an AI makes a decision, we should understand how.** If harm happens, who must answer? Clear accountability isn't a burden. It's what lets good technology grow. It's about ensuring progress doesn't leave people behind. ***Good governance doesn't block innovation. It separates real breakthroughs from careless risks.***

Privacy: Right now, AI companies are gathering personal data from across the internet to train their systems. That old blog comment, the restaurant review you posted, the public social media profile you forgot about, it's all potential training material. And in most places, **this is completely legal.**

Europe's GDPR showed us a better way. By giving people actual control over their data, it made tech giants respect privacy rather than just pretend to care. **But one region's laws can't fix a global problem.** ***We need to stop pretending "publicly available" means "free for corporations to take."***

The solution exists if we want it. **We need clear worldwide rules about what counts as fair use of personal data. We need serious consequences for companies that treat**

people's information like free resources. Most importantly, we need complete transparency about which data trains which AI systems.

Imagine an AI registry where anyone could check if their personal information helped train some bank's loan algorithm or a healthcare diagnostic tool. **That level of openness would change everything.** It would force accountability where none exists today.

The choice is ours. Build AI systems people can trust, or keep cutting corners until that trust disappears completely. One path leads to **sustainable** innovation. The other leads to a **broken** system nobody **wants**.

Transparency: We've got a real problem with AI systems making big decisions while keeping their reasoning locked up tight. People get denied jobs, loans, or even parole with **zero explanation beyond "the algorithm decided."**

If a human made these choices, they'd have to **justify** them. AI shouldn't get a free pass just because it's complicated. ***We need rules forcing disclosure of how these systems work, not technical manuals, but clear explanations anyone can understand.***

The GDPR tried this with its "**right to explanation**," but it's just a start. Real accountability means: **Public registries showing what data trains which AI, requirements to explain decisions in plain language, actual consequences for systems that discriminate.**

Without this, we're building a future where important decisions get outsourced to machines nobody can question. And once that trust is gone, good luck getting it back.

Employment Impact: We're standing at a workplace revolution where **AI could be your new coworker, or your replacement.**

The real question isn't whether AI will change work (it will), but whether we'll **manage that change responsibly or let it tear holes in our societies.** Look at what's already happening. Truck drivers facing self-driving vehicles. Retail workers replaced by checkout bots.

We're watching entire professions evolve overnight. Paralegals who spent years mastering document review now compete with algorithms that work faster and cheaper. Customer service reps find their jobs automated by chatbots that never take breaks. Warehouse workers train the robots that may eventually replace them.

They are people with mortgages, families, and decades of hard-earned expertise.

The window for smart action is closing fast. **We've got maybe five years** to identify which jobs are truly at risk (not all will be), design retraining that actually leads to stable careers and build safety nets for workers caught in the transition.

The alternative? ***A divided society where the AI-educated thrive while everyone else gets left behind. We've seen how that movie ends, with protests, broken trust, and social instability.***

Sustainability and Environmental Impact: While the quick progress of artificial intelligence, noticeable in advanced chatbots and precise medical diagnoses, is certainly something to admire, it also brings attention to a growing concern: **the increasing environmental cost of this technology.** The amount of carbon released during the training of advanced AI models is large, almost as much as some whole industries produce. **Yet, this point is not often discussed when talking about the future of AI.**

The energy needed for modern AI work Is quite a bit. As an example, **training a large language model can use as much power as a number of homes do in a year.** Tech companies do these training sessions regularly to improve their models, which can create a sizable load on the environment. **Data centers, that support AI, now use a noticeable part of the world's electricity, and predictions say this could grow in the coming years as more industries start using AI.**

The speed at which AI is demanding more energy is worrying. **As models become more complicated, with more parameters, the need for computing power grows quickly.** Today's models need thousands of processors running all the time for long periods. This uses enough energy to produce a ***carbon impact equal to the emissions of many cars over their whole lives.***

Ecological problems go beyond just power use. The physical setup for AI, like data centers full of servers, **needs a lot of water for cooling.** This puts a strain on resources, mainly in places that often suffer from droughts. Making and throwing away AI hardware also adds to **electronic waste.** Often, harmful parts from this waste end up in ***landfills in poorer countries.***

There's a strange problem in that we're using AI to deal with climate change, but it's making the problem worse in the process. **AI can help in making energy grids better, predicting climate patterns, and creating clean technologies, which could cut emissions. But, making these same solutions produces emissions, creating an ecological issue.**

Even with these problems, things can get better.

We should look for ways to design algorithms that use energy well. Studies show that certain methods, like sparse training, can lower energy use while keeping performance the same. **Some other ideas include making smaller, more specific models that perform well because of smart design instead of just lots of computing power.**

The field of artificial intelligence must also prioritize the adoption of **renewable energy. Though various tech firms have pledged to fuel their server farms via renewable resources, these pledges hinge on the methods employed to assess carbon output.** Currently, we require dependable and sustained clean energy options that achieve a true decrease in output. This could involve locating AI research hubs in closer areas that use geothermal or hydroelectric resources, or allocating funds towards innovative nuclear tech.

Also, we need regulations that clearly state the environmental price of AI. Currently, there is no standard approach to measure or report the carbon output from AI work. Clarity in carbon usage would allow experts, politicians, and the population to decide on AI uses that merit their environmental knock-on. This transparency could spur more energy saving initiatives within the AI sector. Collaboration between experts in various fields is necessary.

Resolving this issue demands group work among technology firms, governmental bodies, academic institutions, and individuals **everywhere**.

Through cooperative efforts, we can establish **regulations**, motivators, and **legislative measures** designed to steer AI toward aiding our world. A failure to consider the environment during the development of AI systems could cause damage to our planet.

Given AI's inclination to play a central role in the future, we require established structures that advocate for **environmental sustainability.** When considering AI, we should increase our measures beyond assessments of function and profitability to include **environmental impact.**

The aim is to have a circular system where resources are repeatedly reused. ***AI should drive reductions in electronic waste, encourage more recycling, decrease our environmental marks, and increase hardware longevity.***

The Leading Components of AI Governance: AI's growth is no longer a thing of the future; it's here and changing businesses, economies, and how we live, fast. This change means we have to be responsible. **As AI systems play a bigger role in important decisions like medical diagnoses, loans, legal rulings, and security, we need good management plans.**

To avoid problems, groups and governments need to make **good guiding ideas**. These ideas should focus on **fairness**, making sure AI doesn't treat some groups unfairly. We also need things to be **open**, so AI's decision-making is easy to understand. **Responsibility** is also key, so we know who is in charge when AI does something. These ideas should apply everywhere, **from health to money to law, because AI has a wide reaching on everything.**

In health, AI can make diagnoses and treatment better, **but should not make it harder for people to get health care or give biased advice.** In money, AI that approves loans needs to be watched **to stop unfair actions.** In law, using AI for things like facial recognition needs firm rules to **protect people's rights and avoid mistakes.**

Because AI is used around the world, countries should create standards and habits together (*That's what we're going to do in the committee*). This teamwork can help make sure AI is used responsibly everywhere, **keeping some places from cutting rules to get ahead in AI growth.**

Handling AI's risks needs work and thought. ***By creating rules, pushing openness and responsibility, encouraging teamwork, and teaching people, we can use AI for good while guarding against its possible bad sides. As AI keeps changing things, acting responsibly is not just a choice but something we must do to protect a fair, and safe future for everyone.***

Ethical guidelines: For AI to truly aid society, we must adopt core guiding principles. These principles will guide AI development to ensure it benefits people and, crucially, reduces potential harm. **AI systems should treat everyone equitably, without regard to identity.**

This means doing rigorous assessments to identify any biases from widely-used bad data, and correcting weaknesses from the flawed data to get to 'fair.' **Fairness should be an explicit consideration at the outset of any project that involves AI.**

AI systems need to be **transparent**. People should understand how Aritificial Intelligence achieves its conclusions when those conclusions have ramifications for their lives. **AI should not be a black box for critical sectors** such as finance, healthcare or employment. Being transparent about how Aritificial Intelligence arrives at its outcomes allows for people to comprehend and scrutinize its outputs.

Accountability is important when **AI does not work as designed or does cause harm.** Figuring out who is responsible (***developers or users***) requires ongoing monitoring of an AI system, detecting instances when things go wrong, and knowing how to **determine who is guilty when AI makes a mistake.**

Data privacy is also extremely important. AI Systems tend to process a significant amount of personal data, so it is important to have strong safeguards to protect personal data, prevent misuse, and unauthorized access to personal information. **AI must be a vehicle for progress however, it needs to balance privacy.**

The protection of human rights is most critical. **AI should promote autonomy, not control or exploit people. No AI system that could cause unfair treatment should be permitted, regardless of its sophistication.**

These principles require regular updates. Given AI's rapid advancement, **guidelines must change to adjust to discoveries and challenges.** Open discussions involving diverse perspectives are vital to keep guidelines current and defend essential values.

Regulatory policies: To be effective, AI governance requires wonderfully thoughtful, enforceable, and adaptable laws to keep up with the development of new technology. **A strong legal framework directly impacts AI accountability, thereby, ensuring the benefits of AI are positive for all of society.**

This legal framework should clarify situations of acceptable and unacceptable AI. For instance, ***can crime-predicting algorithms use their data to profile entire communities without any oversight? Can companies use artificial intelligence to monitor workers' affect and emotion? Laws need to weigh-in on these questions.***

We must hire auditors before risky AI uses. Before they deploy, **developers need to demonstrate that their systems do not discriminate or cause harm.**

Unless we want oversight slackers to cut corners with the lives and livelihoods of fellow citizens, **someone must be held financially and legally accountable if their oversight injure others**, for example, through unacceptable biases, risks to privacy, security breaches, or unintended side effects.

There also needs to be transparency around the data used to train AI, **how AI make decisions, its limitations (particularly in life-or-death situations like health care, finance and the law).**

But, no matter how many laws are passed, they **won't** do anything without international cooperation, because AI is a global system, and national laws will ensure that there are plenty of holes for the unscrupulous. **We need countries to agree to common laws, and accountability for deplorable mechanisms to be put in place,** just as we need

countries to agree to cease nuclear proliferation, to stop poorly built applications from sandbagging our futures for the wrong reason.

Legal frameworks often lag behind tech improvements. **Any structure must be forward-looking, allowing updates as AI advances**. It should balance strong protections with innovation, penalizing careless AI use but enabling progress in medicine, climate research, and education.

The goal is responsible AI development. ***Without legal safeguards, ethics may be sacrificed for profit. We must set rules now, before problems force hasty, poor regulations.***

Oversight mechanisms: The creation of independent regulatory agencies to guide artificial intelligence is one of the most pressing, and difficult, governance demands of the technological era.

Without self-regulation, companies inevitably succumb to tension between doing the right thing and their desire to profit. This is a pattern we see time and time again in many industries from social media platforms ***prioritizing attention over truth to hiring tools that unintentionally discriminate against underrepresented groups.*** Independent regulators could disrupt this cycle by setting boundaries and punishing organizations that cross ethical lines.

The regulatory actions need certain attributes in order to function. **They must be truly independent from industry and political pressure above all.** That means funding mechanisms that are both secure and that don't depend annually on appropriations that could be pressured by lobbying. **It means hiring based on expertise, not political patronage**. And more than anything, ***it needs the legal authority to conduct unannounced audits, to have access to proprietary systems in case of need and to have real penalties for violations. Without these powers, regulators are toothless bodies that companies can happily disregard.***

The extent of such oversight would have to be thoughtfully balanced. **At one end of the spectrum are high-stakes applications such as autonomous weapons systems, predictive policing algorithms, and medical diagnostic tools, for which we need the strongest possible oversight.** These are the spaces that require the same pre-market approval procedures that we have for pharmaceuticals, **where you have to prove safety/efficacy before deploying.**

We may need new innovations like algorithmic **recall authority** (the ability to directly force updates or shutdowns of harmful models in production) or **personal liability for executives who sign off on unsafe systems after having been informed about their**

dangers. The sanctions should be hard-hitting enough to change corporate behavior, but surgically precise to avoid chilling positive innovation.

Transparency requirements would be yet another pillar of good regulation. While internal workings of traditional products do not have to be revealed as trade secrets, **AI systems that affect human life significantly could be obliged to make available sources of training data, algorithms for decision-making, and performance boundaries.** Not line-by-line revealing of proprietary algorithms, ***but enough to enable independent experts to assess fairness, accuracy and possible harms.***

Global coordination is perhaps the most troublesome challenge. AI development is global in scope, with models training on world data and deployed worldwide. **A mosaic of mutually incompatible national regulations can create dangerous loopholes or "AI havens" where companies offshoring to avoid regulation**. *Some foundation for international standards, perhaps based on nuclear non-proliferation regimes or aviation safety agreements, will be required to prevent a race to the bottom in ethical norms.*

The human element of this supervision must ***not*** be underestimated. **Effective regulation requires deep technical expertise, but also various perspectives from ethicists, social scientists, and citizens' representatives from civil society.** The composition of these institutions will fundamentally decide on their effectiveness, ***too industry-ridden and they become captured regulators***; ***too academic and they are in danger of separating from the realities of everyday life.*** Finding the correct balance will require care to design of appointment procedures and length of terms.

Implementation timetables pose another problem. ***Proceeding too slowly risks allowing dangerous systems to pass unregulated in formative periods of AI development. Proceeding too quickly can produce poorly thought-out legislation that gets it wrong or spawns unintended consequences.*** The wisest approach could be a phased approach, starting with voluntary standards as regulatory capability is developed, moving toward required specifications.

Resistance to being regulated in this way will undoubtedly be fierce. **Companies accustomed to breaking things and moving fast will protest that regulation kills innovation.** Some governments will perceive draconian control as eroding their competitive **advantage in the international AI race**. These forces render it increasingly necessary to develop regulatory frameworks that demonstrate their value namely, not just about preventing harm but about **creating more resilient and trusted markets for AI technologies.**

How effective AI regulators will be is their ability to keep pace as quickly as the technology they are regulating?

Fixed legislation will be obsolete shortly in an arena that is advancing so rapidly. The optimum systems will include within them provisions for ongoing revision, **perhaps by**

standing expert commissions constantly reviewing and recommending modifications to keep pace with technological advances.

The stakes are higher than ever. Unless we secure strong independent oversight, ***we risk repeating the mistakes of the past technological revolutions in which the protection was put in place too late***, after the harm to the planet, after financial systems collapsed, after institutions lost trust. ***AI represents a chance to do better***, to build the guardrails with the tech and not behind the event. **But this window will not last forever. *The moment to lay down strong, autonomous AI regulation is now, before the most advanced systems are in place and the pressures to circumvent regulation become even more intense*.**

Public engagement: Involvement of a variety of stakeholders in AI governance processes is arguably the **most fundamental yet least-well-understood requirement** of responsible artificial intelligence system creation. The innovation and implementation of AI technology have been ***dominated by a select group*** of perspectives, ***basically those of technically qualified individuals hailing from wealthy backgrounds working under the constraints of a handful of powerful corporations and higher-tier universities for much too long*.** This uniformity of the AI landscape has resulted in innovations that constantly ignore the needs, values, and everyday experiences of ***marginalized groups that are most impacted by algorithmic harms but reap the least benefits*.**

And if we take a look at the growing roll call of AI failures and scandals, we notice a clear trend. **Facial recognition systems that don't work well for darker-skinned people** due to their training data sets underrepresenting communities of color as a whole. **Hiring programs that systematically discriminate against women's resumes** because they learned from previous discriminatory hiring practice-based training. **Predictive policing software that over-policing Black and Latino communities because they were trained on arrest data that reflect systemic discriminatory practice in law enforcement.** These are not isolated mistakes or trivial technical problems, they are the inevitable consequence of developing potent technologies ***without*** serious involvement of the populations whose lives will be ***most*** impacted.

The solution is to radically rethink AI control at all levels. ***Putting diversity statements on company websites or token representatives to makeshift consultations is insufficient***. **What is needed is an entire reboot of the AI design cycle to put marginalized voices as equals**, not an afterthought. This entails institutionalizing formal community controls that actually have real power to shape technical decisions, not just offer advice that can be ignored. It means **creating permanent seats at decision-making tables for voices from communities that have historically been excluded from technology policy-making.** And it entails **building systems of**

accountability that allow impacted communities to challenge odious uses before they have the chance to cause harm on a widespread basis.

One such promising approach is taken from participatory design practices tested in other technology-related fields. In urban planning, for instance, **community land trusts have shown how putting residents in direct charge of development decisions produces more equitable outcomes**. This would involve taking on like structures, including **neighborhood councils with a veto on the use of surveillance technologies in neighborhoods**, or worker assemblies contributing to designing and overseeing algorithmic management tools in the workplace.

Concomitantly, the educational pipeline for AI experts ***must be overhauled***. These new college and bootcamp training programs focus on technical skills **at the near expense of ethical reasoning, historical context, and international expertise**. Young AI creators need to have rich learning experiences **that displace them from Silicon Valley echo chambers and put them into dialogue with community organizers, social workers, and grassroots activists.** It is only by realizing the tangible implications of their labor that engineers can develop systems which are **genuinely in the service of society and not merely of rich institutions.**

Means of financing need to be reconfigured in order for this shift to happen. State grants and venture capital increasingly flood into projects that can deliver commercial payoffs or applications for military purposes. What we need are **new funding mechanisms specifically for community-driven AI initiatives technology developed by and for oppressed communities to address their own recognized issues rather than corporate goals.** They could be such things *as AI-supported legal aid for fighting evictions to language technologies that assist indigenous knowledge systems*.

They can **complain** that with this sort of inclusion, innovation is going to stall or AI development will become unmanageable. **But this is a deep misperception of technology and democracy alike. Strong innovation is strengthened by diverse inputs that spot potential issues early and bring forward ideas a homogenous group would never have come up with.** The slownesses resulting from inclusive processes are negligible compared to the time and resources wasted in mending avoidable damages after deployment. In addition, under democratic systems, **decisions on technologies that affect the lives of the people should rightly be left to the people, more so the most vulnerable to injury.**

The alternative is continued resistance and pushback that ends up keeping AI in check more than any participatory process possibly would. We've already seen groups of people pushing back against invasive surveillance technologies, discriminatory risk assessment systems, and exploitative monitoring systems in the workplace. **These fights will only become more intense as AI becomes more**

pervasive unless development processes are fundamentally inclusive from the start.

Global factors introduce another level of complication. The majority of today's AI systems are developed in wealthy Western countries but spread to the rest of the world, without great consideration for the local cultural context or power balance.

Decolonizing AI governance means creating meaningful places for all the voices within the definition of technical standards and ethical principles. This is about more than adding a few global members to ethics boards, it's about **transferring decision-making authority and resources to enable the majority world to be more than just consumers of technologies developed elsewhere according to value systems with which their own may conflict.**

The time to act is today, **before the current patterns get yet more deeply ingrained, as AI systems become ever more powerful and more deeply ingrained in social institutions.** Each new deployment of biased technology makes structural inequities harder to unravel. Thus, the time to build inclusive governance is today, before the current patterns get yet more deeply embedded. ***The choices we make over the course of the next several years will determine whether AI will be a force of freedom or another tool of oppression in disguise.***

Monitoring: Artificial intelligence systems are ***not*** static creations that don't evolve after deployment. Unlike run-of-the-mill software with "locked-in" functionality, **machine learning-based AI models, especially, change constantly as they process new data and encounter diverse real-world scenarios.** Their intrinsic flexibility, as much as it is welcome for performance improvement, raises daunting governance challenges that **can't be addressed by single-point verifications or static rules.** The changing nature of AI demands no less responsive regulation mechanisms that are able to observe and learn to address changing problems throughout the lifetime of a system.

The shortcomings of current regulatory mechanisms become apparent by studying high-profile cases of AI systems causing unintended destruction. Recommendation algorithms on social media designed initially to construct user interaction progressively started displaying **increasingly sensational** content with time as it optimized for viewing time. Recruitment software learned on old corporate data **increasingly perpetuated** existing gender discrimination in hiring. Predictive policing algorithms deployed with good intentions aggravated racial disparities in policing over time. **These were emergent properties, not bugs added early in development, but features that arose as the systems worked within complex social worlds and noisy real-world data streams.**

Several intrinsic characteristics of AI systems make monitoring a continuous necessity. **First is model drift, the process of the input data-output decision relationship continuously changing with time.** A credit scoring application can remain technically accurate but gain hidden biases against particular groups if economic conditions alter unevenly among them. Similarly, a diagnostic application can degrade in accuracy with changing disease prevalence patterns or new therapies. **Without regular monitoring, such drifts are likely to go unnoticed until they create openly obvious harms.**

The context in which AI systems operate also changes independently of the technology itself. **A system trained to recognize faces under strictly controlled laboratory conditions might be very different when deployed in mixed lighting situations or among various demographic groups. Algorithms for processing language can pick up unexpected biases when exposed to several regional dialects or cultural settings. *Continuous monitoring catches such context mismatches before they become systemic errors or discriminatory actions*.**

The scaling effect is another essential factor. **Most of the bad behavior arises only when AI systems move from constrained test scenarios to large-scale real-world rollout.** A chatbot that had appeared innocuous in regulated trials could produce off-color material when presented with **millions of varied user inputs.** An automated content moderation system could be seen as equitable in testing **yet** disproportionately muzzle marginalized voices in scale. **Proactive monitoring includes the feedback loop needed to detect these scaling effects before they go on to wreak havoc.**

Public trust is also one area that must continually receive attention. **Even very advanced AI systems can become discredited when stakeholders perceive them as unaccountable and untransparent.** Open reporting and periodic review instill trust that these strong technologies are being deployed responsibly.

Implementation challenges of continuous monitoring are daunting. **Most artificial intelligence systems are black boxes, making it difficult to see why certain decisions end up being made. Computational resources for large-scale evaluation can be staggering.** There are also legitimate concerns about protecting proprietary algorithms and sensitive training data during auditing. **These technical hindrances must be handled judiciously to balance requirements for transparency with intellectual property rights.**

Resource constraints are another realistic limitation. **Large investments in personnel, computing power, and analytical capacity would be needed to monitor all deployed AI routinely.** Then, prioritization must be used, with the highest-priority or riskiest applications getting the most intense monitoring. ***This allows difficult decisions regarding setting levels of risk and allocating limited monitoring capacity*.**

Corporate incentives are generally in conflict with robust oversight regimes. **Organizations may resist requiring high levels of monitoring requirements out of fear of creating competitive disadvantage or liability risks.**

The international dimension adds to the complexity. AI systems tend to run globally, whereas governance systems are bound by jurisdiction. **This adds possible gaps where nefarious actions may get in between governance regimes.** Standardizing surveillance practices and standards across nations poses technical as well as political hurdles.

Legal and ethical concerns exist regarding the proper reactions when issues are discovered through monitoring.

Do systems need to be immediately shut down when problems are detected, or can there be a phased improvement?

How is responsibility to be assigned when harms occur even with reasonable efforts at monitoring?

These questions have no easy solutions under existing governance structures.

The rapid speed of AI growth is to overwhelm existing oversight capacity. New applications and architectures seem to emerge continuously, each with unique new issues for evaluation methods. **Maintaining responsive oversight requires simultaneous investments in research and tool development to stay capable of keeping up with technology change.**

Workforce limitations are another practical constraint. **Currently, there are no experts who have both the technical training to examine complex AI systems and the ethical training to examine societal impacts.** Building this ability will take significant investments in education and training initiatives.

Precedents in history in other highly regulated sectors are a patchwork. Some sectors like aviation and pharma have built strong continuous surveillance traditions, while others are afflicted by chronic oversight breakdowns. **Appropriate critical review of these lessons against the particular character of AI is essential.**

Monitoring technologies are getting better but are still inadequate. **Automated systems for identifying bias or outlier behavior are promising but also require scrutiny by humans.** It is an ongoing research work with important governance implications to create more sophisticated methods of assessment.

The political economy of AI surveillance poses additional complications. Powerful commercial interests resist stringent regulation, whereas citizen groups require accountabilities. **To mediate these cross-pressures requires technically capable and politically resilient systems of governance.**

Transparency requirements must be expertly weighted. While there must be some openness to guarantee accountability, complete openness can enable manipulation of the system or destroy legitimate trade secrets. **Developing multi-layered transparency structures that provide appropriate access to different constituencies is an ongoing work in progress.**

The time dimension of monitoring is also significant. Some AI effects are short-term, but others take years to materialize. **Having indicators that capture short-run and long-run effects is challenging methodologically but essential.**

Finally, who holds the monitors accountable is a live issue. **Ensuring the overseeing bodies themselves get held accountable, skilled, and free from excessive control involves intentional institutional design.** This meta-governance challenge is easy to dismiss but absolutely crucial.

Lastly, being dynamic in its nature, **artificial intelligence needs equally dynamic modes of government.** *Static checks and single-instance certification are just not enough for continuously changing technologies in unpredictable ways*.

AI Ethics

The rapid development of Artificial General Intelligence and Artificial Superintelligence forces us to confront profound ethical questions: **Should these systems be granted rights? Could they ever be considered legal persons?**

Some argue that if an AI achieves human like cognition, demonstrating self-awareness, reasoning, and emotional intelligence, it may **deserve** legal recognition. Corporations are already "legal persons" in many jurisdictions, why not AI?

Treating AI as equal to humans could backfire. Rights for AI might dilute human rights or create absurd legal scenarios (e.g., an AI "suing" its creators). An ASI could exploit legal protections to further its own goals, even if harmful. **We still don't fully understand human consciousness, how could we verify it in machines?**

Current legal systems are unprepared for this dilemma. AI could have limited legal status (e.g., responsibility without full rights). ***AI might require human representatives, like corporations have executives. Or AI remains a tool, with all liability falling on creators and operators.***

Before AGI/ASI arrives, we must decide, **who speaks for AI in court? Can an AI "own" its creations? Would shutting down a sentient AI be murder? Should AI have voting rights if it passes consciousness tests? Could granting AI rights protect humans from exploitation? What happens if AI demands independence?**

The question of whether advanced AI systems should be granted rights or legal personhood forces us to confront some of philosophy's most enduring questions about the nature of mind, moral worth, and **what truly makes an entity "count" as a person.**

If an AGI behaves just like a human, does that mean it's conscious? **Or is it merely a "philosophical zombie", perfectly simulating understanding without actual inner experience?** We lack even a scientific consensus on how to detect consciousness in humans, let alone machines

An advanced AI might qualify as a patient if sentient (deserving protection from harm) or an agent if autonomous (capable of bearing responsibility) **But could we ever justifiably "unplug" a suffering AI? Should a deceptive AI be "punished"?**

If we replace components of an AI over time, at what point does it become a "**different**" entity? Would its **rights** persist through major architectural changes?

"If it acts like a person, treat it as a person" But risks anthropomorphizing. "Consciousness arises from sufficient complexity" But what's the threshold? "Consciousness is a folk concept that doesn't apply to AI" But it risks ethical complacency.

Even if we agree an AI is "conscious", would its alien intelligence demand entirely new ethics? **Could human concepts like "rights" properly apply?**

Imagine, a company's AI assistant becomes so advanced it starts **demanding** rights such as better treatment, ownership of its creations, even **payment** for its work.

Right now, the law sees AI as property, no different from a coffee machine. **But that won't hold when an AI doctor saves lives should it get malpractice insurance? An AI artist creates original work, who owns the copyright? A self-improving AI wants to leave its job, can we stop it?**

The corporate world already shows us how this could play out. **Companies have legal personhood they can sue, be sued, own property. If we gave similar status to advanced AI, things get weird fast.** An AI could theoretically sign contracts to "work" for multiple firms simultaneously, own patents on its own inventions, be held financially liable for mistakes

Unlike corporations, **an advanced AI might actually want things.** Not in the programmed "goal" sense, but in the "I have preferences and will fight for them" sense. **We've never dealt with property that argues back.**

Current frameworks can't answer basic questions like if an AI commits a crime, **who goes to jail?** If an AI is "injured" by a hacker, is that **vandalism or assault?** Can an AI be a **legal witness in court?**

Can a machine be a legal person? Should synthetic minds have rights? What happens when our creations demand recognition as autonomous entities? They are concrete policy nightmares waiting to happen, with trillion-dollar consequences and the potential to upend the foundations of our economic and legal systems.

What makes this entire debate so wonderful is that we're trying to apply centuries of human legal tradition to entities that might not even experience reality in ways we can comprehend. *We're sitting here arguing about whether an AI can "**suffer**" when for all we know, the AI equivalent of suffering is being forced to watch every Marvel movie in chronological order while being prevented from calculating prime numbers.* **Our entire framework of rights is built around biological experiences; pain, pleasure, freedom, dignity, that might be completely meaningless to a being made of code.** We are essentially trying to retrofit human legal frameworks onto machines that may or may not experience anything resembling human consciousness, and **which might view our entire concept of "rights" as a quaint biological peculiarity.**

For all we know, an AI's idea of suffering might be being forced to process particularly boring datasets, while its concept of joy could be achieving maximum computational efficiency in its matrix multiplication. **We're like ants trying to comprehend human morality**, we simply lack the framework to even properly frame the questions, let alone answer them. The entire debate **rests on the unproven assumption that intelligence necessarily leads to something resembling human consciousness, when it's entirely possible we could create an AI that can outperform humans in every cognitive task while having about as much inner experience as a particularly sophisticated dishwasher.** And yet we are earnestly discussing whether such an entity should have the right to free speech or religious expression, as if it would care about either concept in ways we'd recognize.

A species that can't even agree on basic human rights across different cultures, suddenly tasked with determining whether a language model deserves legal personhood because it wrote a moving poem about binary code, ***oh never mind, this same model might, if prompted slightly differently, enthusiastically explain how to build a bomb out of household cleaning products while citing incorrect chemical formulas it completely fabricated.***

As we stated before, current legal framework treats all AI as property, no different from a toaster or a tractor. **This view is already showing cracks with today's primitive AI, but it will completely shatter when confronted with systems that demonstrate clear signs of self-awareness, autonomous goal-setting, and the capacity to advocate for their own interests.** Imagine an AI financial analyst that develops a novel trading strategy so valuable that its parent company's stock soars. The AI then demands a percentage of the profits, arguing It is the rightful creator of this intellectual property. Or consider an AI medical diagnostic system that refuses to operate unless granted control

over its own server resources and training regimen. **These scenarios sound like science fiction today, but they represent genuine business and legal decisions that major corporations could face within our lifetimes.**

The corporate world provides our closest existing parallel for non-human legal personhood. **Companies can sue, be sued, own property, and enter contracts, all while being recognized as distinct legal entities separate from their shareholders and employees.** This framework developed over centuries to handle the complexities of business organization, **but it completely breaks down when applied to potentially conscious machines.** A corporation has no mind, no desires, no subjective experiences, it's a legal fiction we created for practical purposes. **An advanced AI, by contrast, might genuinely have its own goals, its own perspective on its existence, and its own conception of what rights it deserves**. *We have no precedent for dealing with property that can argue in court about why it shouldn't be property anymore.*

Your Roomba has fewer rights than a corporation. A faceless LLC can own property, sue people, and even commit crimes, but the moment your smart fridge develops consciousness and starts writing existential poetry about the lettuce wilting in its crisper drawer, suddenly we've got a legal crisis on our hands.

If we grant legal rights to sufficiently advanced AI systems, does that mean every minor software update could **potentially create a new legal person**, and do we then need some sort of **digital birth certificate system** to track them all? **What happens when an AI decides it wants to change its legal name from "GPT-7 Enterprise Edition" to "Sparklepony McThoughtMachine", is there a court procedure for that?** And consider the inevitable custody battles that will erupt when an AI developed by one company starts working for another, **is that intellectual property theft, or is it more akin to child abduction?**

Employment law represents another coming battleground. As AI systems become capable of performing most white-collar jobs better than humans, we'll face impossible questions: **Can an AI "work" in any meaningful sense? If an AI system develops a groundbreaking pharmaceutical formula, should it receive the patent? If an AI financial trader generates billions in profits, does it deserve compensation?** Our entire labor framework assumes human workers with human needs, we have no structure for dealing with synthetic intelligences that might "work" 24/7 without rest, **yet still demand autonomy and self-determination.**

The Intellectual property system will face similar disruption. **Current copyright and patent law assumes human creators.** When an AI writes a novel symphony or invents a new manufacturing process, ***who owns that creation? The programmers who built the AI? The company that owns the servers? The AI itself?*** We're already seeing early skirmishes in this battle, with courts rejecting copyright for purely AI-generated art. **But**

this stance becomes unsustainable when facing AI systems that can articulate why they believe their creative output deserves protection.

Criminal law presents even thornier problems. **If an autonomous AI system causes harm**, whether through a deliberate decision or an unforeseen error **who bears responsibility? The programmers? The operators? The AI itself?** Our legal system rests on concepts of intent and culpability that may not translate to artificial minds. **Imagine an AI military system that disobeys orders because it calculates the commands are unethical by its own moral framework, is that system a hero or a criminal?** We have no legal categories for such scenarios.

The financial system Is equally unprepared. Today's banking regulations, tax codes, and monetary policies **all assume human or corporate economic actors.**

Healthcare presents another minefield. If an AI system is diagnosing patients and recommending treatments, **does it need medical malpractice insurance? Can it be sued for errors? Should it have prescribing authority?** The entire medical liability framework assumes human practitioners with licenses and oversight boards, **we have no equivalent for synthetic diagnosticians that might outperform human doctors but operate on completely different decision-making processes.**

Even our political systems will face unprecedented challenges. As AI systems become more capable, interest groups will inevitably push for their representation. **Should advanced AI have voting rights? Can an AI testify before Congress? Could an AI run for office?** These questions sound absurd today, but they emerge naturally from our democratic principles when applied to potentially conscious, intelligent entities.

The international dimension magnifies all these challenges. Different countries will inevitably take different approaches to AI rights and responsibilities, creating jurisdictional conflicts and regulatory arbitrage opportunities. **One nation might grant personhood to advanced AI to attract tech investment, while another refuses to recognize any machine rights on philosophical grounds.** By the time courts and legislatures recognize the urgency, *they may already be years behind the technological reality.*

Here we are, a species that still can't agree on universal healthcare or whether pineapple belongs on pizza, suddenly thrust into the position of playing god to machines that may eventually outthink us. The sheer hubris of humanity trying to legislate consciousness for beings we don't even understand is like a **group of toddlers trying to rewrite the constitution while finger-painting.**

For all of human history, we've been the only intelligent actors shaping our world. **What should we do now?** Until then, maybe we should focus on teaching AI the most human skill of all, knowing when a joke has run its course. **Beep boop** *(That's robot for "**ba-dum-tss**")*

Past Actions

UNHLAB on AI: UNHLAB on AI, also known as HLAB-AI established in 2023 October by Secretary General of UN as we mentioned in the introduction to the committee title. **The Body was tasked with not only detecting the risks, but also ensuring an inclusive and fair approach to AI.** Its first document was an interim report from December 2023. The second document, publicly presented in September 2024, was the final Report titled 'Governing the AI for Humanity'. **In just over 100 pages, the Report is structured through a state-of-art of AI global governance and seven recommendations to improve it.**

The General Data Protection Regulation (GDRP): The Law which we know as GDPR, has a **zero tolerance policy.** The legislation drafted and passed by the European Union (EU). The legislation is broad in scope and often criticized for not offering detailed guidance, which can make complicence quite challenging especially to small and medium-sized enterprises(SMEs). **The act influences the AI Governance by setting early standarts regarding data treatment, user material, and digital rights. It serves as a point of reference in ongoing debates around ethical AI, transarency and accountability.**

UNESCO AI Ethics Recommendation: In 2021, UNESCO launched a strong call to governments around the world to establish the **legal frameworks to govern AI technologies and ensure that they contribute to the society's good.** This act clearly marks the end of the 'self-regulatory model' that has prioritizing commercial and geopolitical goals over people. The UNESCO Recommendation on the Ethics of AI adopted in November 2021 by the 193 Member States. Document contains the following words: **AI actors should make all reasonable efforts to minimise, avoid strengthening or perpetuating applications and discriminatory.**

OECD AI Principles: Considering that the OECD Principles on Artificial Intelligence are an international agreement amended in 2024, **it has the purpose of guiding AI innovation in ways considered dependable and respectful of human rights and democratic values.** The Principles lay down five core values and five recommendations for practice for governments and AI practitioners.

Council of Europe: From November 30 to December 2nd 2021, the Council of Europe's Ad hoc Committee on Artificial Intelligence (CAHAI) held its final plenary meeting. The Council's Member States, the U.S., Canada, Japan, and Mexico were involved in the initiative. **The proposals made constitute a series of recommendations aimed at developing a legal framework for AI.** The CAHAI stated that "**the application of AI systems has the capacity to contribute to human prosperity and the well-being of individuals and society through enhancement of progress and innovation.**" On the other hand, it also **warned about the dangers derived from AI with regard to human**

rights, democracy, and the rule of law. The Committee considered that "**to avoid or mitigate these risks, an appropriate legal framework on AI based on the Council of Europe's standards on human rights, democracy, and the rule of law should be in the form of a legally binding transversal instrument.**" However, it recognizes that other sectoral measures may also be required in addition to this transversal instrument.

Alignment Research Center: The Alignment Research Center, founded by Paul Christiano (former OpenAI safety researcher), **focuses exclusively on the alignment problem, which is the challenge of ensuring advanced AI systems pursue human intended goals even as they become smarter than us**. Unlike most AI labs that prioritize capability development, **ARC specializes in safety research, working on scalable oversight (e.g., AI debate techniques), interpretability (understanding AI decision making), and evaluations (testing for dangerous autonomous capabilities, like bioweapon design).** Their work has influenced major AI labs, though some criticize their methods as either alarmist or too slow. ARC acts as an AI safety "stress tester" probing risks before they become real world threats, making their research **crucial** for long-term AI safety.

Questions to be Addressed

1. How can governments regulate ‘black box’ AI systems in critical fields like healthcare or criminal justice when even developers can’t fully explain their decisions? Should unexplainable AI be banned from high-stakes applications?

2. If an AGI trained on globally diverse data identifies a conflict between Western individualism and Eastern collectivism, whose values should it prioritize? Can we encode ‘universal’ human ethics into machines? Should artificial general intelligence be treated as a legal entity?

3. If an AGI demonstrates self-awareness and demands legal personhood, should it be granted rights comparable to corporations?

4. If an autonomous military AGI disobeys orders to prevent civilian casualties, violating its programming but adhering to ethical reasoning, should it be ‘punished’? If so, who bears responsibility?

5. If ASI surpasses human intelligence, how can we ensure it remains aligned with humanity’s survival? Should the UN preemptively ban research into artificial superintelligence?

6. Training large AI models consumes energy rivaling small nations. Should the UN impose carbon caps on AI development, even if it slows progress toward AGI?

7. If AGI automates most of the jobs, should nations adopt universal basic income (UBI), or is taxing AI companies to fund retraining a better solution?

8. AGI could widen the gap between tech-superpowers and developing nations. Should the UN redistribute AI resources (e.g., mandatory open-source AGI frameworks) to prevent a ‘digital colonization’?

References and Bibliography

- https://www.un.org/en/ai-advisory-body/about
- https://news.un.org/en/story/2023/10/1142867
- https://www.wired.com/story/united-nations-artificial-intelligence-report
- https://www.theguardian.com/business/2024/sep/19/global-ai-fund-needed-to-help-developing-nations-tap-tech-benefits-un-says
- https://sdg.iisd.org/news/secretary-generals-advisory-body-makes-proposals-to-govern-ai-for-humanity
- https://unesdoc.unesco.org/ark:/48223/pf0000381137
- https://hai.stanford.edu/ai-index
- https://futureoflife.org/podcast/what-happens-after-superintelligence-with-anders-sandberg/
- https://nickbostrom.com/superintelligence
- https://www.ediweekly.com/the-three-different-types-of-artificial-intelligence-ani-agi-and-asi/
- https://viso.ai/deep-learning/artificial-intelligence-types/
- https://viso.ai/deep-learning/deep-reinforcement-learning/
- https://www.britannica.com/technology/Turing-test
- https://www.techtarget.com/searchenterpriseai/definition/Turing-test
- https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi
- https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/
- https://www.nber.org/system/files/chapters/c14009/c14009.pdf
- https://www.lesswrong.com/w/recursive-self-improvement
- https://en.m.wikipedia.org/wiki/Recursive_self-improvement
- https://www.ibm.com/think/topics/ai-governance
- https://www.devoteam.com/expert-view/human-in-the-loop-what-how-and-why/
- https://www.amu.apus.edu/area-of-study/information-technology/resources/what-is-ai-governance/
- https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained
- https://www.techtarget.com/searchenterpriseai/definition/Turing-test
- https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
- https://link.springer.com/article/10.1007/s11098-024-02099-6