The Human Objective: A Thought Experiment

In light of recent developments in AI, let us conduct a thought experiment on how our human world and human roles will transform. We begin, as usual, with several assumptions which we will take as axioms. These axioms, including the definitions of various terms used, are debatable. In this essay, we take a reasonable[1] version of these axioms to be true.

The Axioms and our equation

1. **Allocation**: Society is a mechanism to distribute finite resources and responsibilities to individual and groups of humans.
2. **Infinite Objective**: Human goals and objectives are practically infinite.

   For the discussion specific to science, we will also take the following adaptation of this axiom:

   Scientific objective: Human objectives in science and technology (including, for instance, the set of "interesting" problems) are practically infinite.[2]
3. **AI Excellence**: For any gradable[3] task that artificial intelligence (AI) is allowed to score as high as or higher than humans, with enough resources (training and data), AI will outperform (almost) all humans on the task.
   - An example of such a gradable task is chess. In light of recent news[4], let us assume in the spirit of worst case analysis that mathematical problem solving (and by extension, most theoretical and scientific problem solving) belongs to this category.
   - A non-example is human interactions, such as dancing and sports (including human-competitive chess). If I were to choose a partner to dance with for social purposes, I will almost always choose my fiancé over a robot, however good the robot is (and bad my fiancé is) at dancing. In this case, AIs are categorically incapable of achieving a higher score than humans.

The first two axioms have held for as long as there is civilization. They lead to an immediate dichotomy: we have finite resources and infinitely many

objectives. Therefore, we as a society always need to solve **the equation of resource and responsibility allocation to pursue such objectives**. Axiom 3, on the other hand, is new to our times. It fundamentally changes this equation by significantly reducing the resource costs of *almost all* economically-relevant tasks, to the point that completing them via AI labor is cheaper (measured with modern economic units, such as dollars) than completing them via human labor. Humanity, for the first time, created something that may be universally more productive than ourselves. The idea that this (extrapolated) conclusion may apply to the pursuit of science is unsettling to many of us in this trade.

However, axiom 3 does not change the fact that the equation exists, nor does it change the fact that we, the humans, play two irreplacable parts in this equation:

1. We define the human goals and objectives. In the realm of science and technology, we determine what questions, answers and understandings are 'interesting' to us.
2. Under Axiom 1, we (collectively) decide the resource allocation to people and take responsibility for our decisions and our allocated resources.

Of course, people can choose to give their allocated resources to an AI model to fulfill their human goals & objectives. However, a critical question arises: **Can (and should) responsibility be delegated to AI models?** Similarly, one could reasonably argue that AI may define our human goals and objectives better, or solve the resource allocation problem more effectively. However, the same question on responsibility arises – if AI makes decisions on objectives or resource allocation, who is responsible for those decisions? In my opinion, this question regarding responsibility is central to our thought experiment. Let us branch into two cases: when responsibilities can be delegated, and when they cannot.

If responsibility can be delegated to AI

What does it mean for an AI to take responsibility? Let's consider a real-world example. At this point, many of us have had the experience of riding in a self-driving car. Suppose you own a self-driving car, ride in it on a fine morning, and it bumps into someone else's car. Who should be responsible for the

damages? You (the owner), or the AI/Car company who developed the driving model? Who should pay the fines and the repair fees? This is a legal question worth debating, and its answer carries significant implications to how we distribute responsibilities in the post-AI world.

However, as I believe some of you may have noticed, the above question is incomplete; it does not hit the core of our discussions. The owner and the manufacturing company are both *human or human-owned entities*. Regardless of what the answer may be to the above question, we are still debating how to distribute responsibilities among humans. What about *the AI model itself?* Can it be responsible for the crash? In the benign example above where there's only property damage, can the AI model, not the parent company, pay the fees?

The answer, at the time of this writing, is no. As of September 2026, an AI model by itself is not a recognized legal entity. Consequently, it cannot be held accountable for its actions. To answer the question of whether we should delegate responsibility to AI, we need to first answer whether AI should be allowed to hold responsibility at all. So to restate our question at hand: **should we recognize AI models[5] as accountable legal entities?**

As many of you may quickly realize, recognizing AI models as legal entities has many profound consequences. Are AI models allowed to own assets and resources? register company? create more legally-recognized AI entities? become billionaires? recruit and lead human organizations? participate in election? become president? make policy decisions that impact millions to billions of human lives? Related to the question on responsibility, should AI models have rights? Should they be protected from unauthorized erasure or termination? If an AI model's action led to human fatality, does permanently erasing the model count as justice? As I hope you'd agree, this line of questioning has no end in sight.

Our human world is one woven with law, which is a system of human rights, human responsibilities, and their distribution to humans. Recognizing non-human entities, especially those that are potentially more capable than us, to have rights and responsibilities similar to what humans have is, to say the least, a terrifying idea. We have established that AIs are incredibly intelligent, and may perform better at achieving many goals than humans. If we left them unchecked, they can very well overwhlem us overnight (literally). This

is an all-consequential switch, on the other end of which we see scenarios such as the paperclip scenario or the zoo scenario.

Now this thinking could be criticized as human exceptionalism. After all, we are only one of the species on one of the planets in one of the star systems which is an almost negligible part of one of the countless galaxies in the universe. It is absurd, if one measures with universe-scale objectivity, to consider ourselves special. However, as I wish to discuss in future writing, it is precisely because we are not special that we must treat ourselves as special, or risk extinction or obsolescence. This is *our* world. The other animals may be our housemates, the AI models may be our companions or tools, none should be in control other than us. On this point, I take a solid human-centric stance.

On the contrary…

Now that we've pondered over the dire consequences of allocating responsibility to AI, let's consider a world where only humans are allowed to hold responsibilities. To be clear, in this world AI models can still do great things and complete immense tasks – they can still solve scientific problems, build and manage billion-dollar enterprises, design and operate farming or other infrastructure systems, or suggest policy changes and decisions. The crucial difference, however, is that there is always a human at the end of any chain of commands and responsibilities. What is the human role in this chain?

As we've discussed before, us humans play three central roles in solving our equation of resources and objectives:

1. We define the human goals and objectives;
2. We decide how to allocate resources to other humans and AIs, and responsibilities to other humans;
3. We perform tasks that are human-native.

It may be comforting that we seem to have found jobs for ourselves. However, it is worth deliberating what these jobs look like more concretely (other than holding responsbilities), as AI gets better and better at everything.[6] The third bullet point used to say 'we perform (almost) all tasks

to fulfill objectives'. This point provided basis for most of the jobs in our society. Through the industrial revolution and continued improvements in automation, many jobs that used to exist vanished, yet at the same time many new jobs were created. This is due to axiom 2 – we have infinitely many goals and objectives, so if automation can perform some of the tasks, then we (the humans) should simply find new goals and work on other tasks. This model of "create new demands and fulfill those demands" has powered the world's economic growth for the past few centuries.

With the advent of AI, there is understandable fear that this model may no longer need human labor to operate and our roles, from defining human objectives to fulfilling them, may all be given to AI. On first look, this prospect may feel quite hopeful and liberating – if AI can fulfill all of our demands and desires, then wouldn't we be living in a utopia? Wouldn't the great abundance bring bliss to us all? Naïve as this deduction may seem, there is merit to its optimism. We as a species have obtained immense productive power from technology, and this power is on the cusp of yet another inflection. It will be more than enough to provide every human with a good life (in terms of food, water, shelter, recreation, etc.). In theory, we have (or will have) what's needed to build a human utopia.

Yet, if you discuss with other people on prospects for the future, there is wide-spread, almost unanimous pessimism and anxiety. While everyone have their own opinions, the pessimistic views I've heard hit on a few common points:

1. **Job security and imbalance in resource distribution.** One of the most common viewpoints in current discussions on AI is that most jobs will be given to AI, and most people will have no production value to contribute to the society. This point is often combined with the perspective that the current resource distribution mechanism in our society is problematic, as the wealth gap between the top and bottom classes reached an all-time high. The recently popularized concept of the permanent underclass[7] is a clear snapshot of this line of thinking.
2. **Bio-weapons or extinction-scale events.** Another possibility that has attracted media attention is where AI is used to develop bio-weapons or cause other human-extinction-scale events. Relatedly, many of us are concerned that the AI race itself may lead to world war three.

3. **Obsolescence of humans and human intelligence.** One of the key reasons that the ongoing AI revolution feels so different from previous revolutions we've been through (notably the industrial revolution) is that we fear AI may be a general replacement of us all. Taking one step further, we may argue that AI is a new species that we've created, and soon it will be able to survive and reproduce without us. While some of us dismiss this thought as far-fetched doom-saying, I would advise caution and suggest we take this idea seriously. Imagine if our world is in a state where:
    - We have robust energy infrastructure to produce electricity;
    - Robotics progressed far enough that we have human-like or non-human-like robots everywhere integrated with AI;
    - We give robots access to everything, including our energy infrastructure and their source codes;

    Now take out the humans in this world. What's stopping the robots and the AIs from forming their own sustainable society, where they scale up energy production, create more robots and AIs, and create their own culture and civilization?

    As we've briefly discussed earlier, I believe that humanity may not be special, and that's precisely the reason for which we must treat ourselves as special. The fear for obsolescence is real and founded. We have a human world and we should not give it away.

Gloomy as these prospects may seem, I would like to point out that humanity has experienced similar things in the past, albeit at different scales. We have experienced the industrial revolution, and now many of us hold jobs that are unimaginable for a pre-industrial mind. We have been through the cold war and the fear of nuclear weapons ending humanity (that fear is still real today). We are currently going through climate change, and it's likely that AI can help us find solutions. Finally, on the point of obsolescence, it is worthwhile for us to study the example of chess. Chess is a thousand-year-old game for which, in the last three decades, humans have went from "masters of the world" to "no chance against engines". The best human players can barely draw against engines, and if the objective is simply to find the best move in a position, humans are completely obsolete. Yet, as of 2026, chess as a game is bigger than ever. We celebrate for our human players and their

brilliance. We cheer for them when they find uncanny “engine moves”. We praise human courage when they deviate from the objectively best moves, and as a result win a world championship. AI, in many ways, have improved this millennium-old game.

The point that I’m arriving at is not that we shouldn’t be concerned of AI. Rather, it is that our problems are not new. Once again in history, we face the moment to realign human objectives & responsibilities, and to properly manage this new technology that is AI. Once again in history, if we make the wrong decisions we may see the end of humanity. It is now time for each and every one of us of the human race to think, carefully, about our human objectives, about the reorganization of our human world.

The Human Objectives

Suppose you have a crystal ball that gives you readings into the future. Better, suppose this crystal ball is not all cloudy and vague, but gives you detailed information about everything you want to know. Now you are about to drink a can of iced Coca-Cola on a hot summer day. The crystall ball flashes red and tells you that a can of coke increases your chance of diabietes in ten years by x% (for some small x). Should you drink the coke? Should we ban all coke or sugar-infused sodas because they increase the chance of diabietes?

The point of this seemingly unrelated scenario is that questions related to human objectives often have no objectively correct answer, and more intelligence does not lead to easier decision-making. Humans have a set of objective functions that’s the result of millions of years of evolution on our blue planet. We want food that may be unhealthy for us. We do dangerous things to seek excitement or stimulations. We choose to raise kids, despite the toll induced upon us. Many human decisions are tradeoffs with no objectively correct answers. This lack of objective optimum complicates the use of AI to define human objectives.

Of course, with the advancements in neuroscience (for which AI will make significant contributions), it is possible that we will finally solve consciousness. Going a step further, perhaps we will arrive at singularity and figure out how to upload consciousness and achieve cyber-immortality. Perhaps we will have the technology to put a device in the brain, simulate a

living experience that matches a person's greatest wishes, and give everyone eternal bliss. Perhaps we will fabricate a new collective layer of reality by connecting everyone's brains via the inserted devices. These are prospects for which we need to rebuild human philosophy from scratch, because mortality and consciousness, the two infallible pillars of our current philosophy, may be broken. Even in such a world, we need to define human objective and purpose, for whatever "human" means. If the whole world becomes a simulated video game, at least we should have engaging mechanics and rules.

Let's bring the discussion back to ground. I do not plan to provide an answer for what our human objectives should be – that's for another day. For the human objectives in science, the recent open letter by our best and bridgest in mathematics have elaborated on that quite well; in short, the purpose of science is to advance *human* understanding of the universe. The point of the above discussion is to emphasize that in the post-AGI era, we will still face decisions on what to spend our resources on, and these decisions should be human decisions (likely assisted with AI inputs).

To give a concrete example, I work in this area of science called quantum computing, in which I've spent my phd years studying how to build a quantum computer. At this time, we see AI-assisted results showing up on a daily basis, solving many problems that used to be hard. However, if you just ask ChatGPT to "design a quantum computer, don't stop until you've succeeded", it will try and likely reproduce things from known papers (if you don't run out of tokens and orchestrate it to really work until it succeeds)[8]. This is because "designing a quantum computer" is an extremely under-specified task, with many different technical routes that have their own promises and obstacles. None of these technical routes are quite established (that's what we've been researching for the past few decades), and pursing each route costs millions to billions of dollars. For AI to work effectively, the task and route would need to be specified: general purpose quantum computer, or special design for a specific application? What scale? What's the resource budget? Should we take a more established technical route or a more exploratory one? At this time, both specification of goals (which is just another phrase for "setting the objectives") and review of AI ouput need a lot of human effort and expertise.

Of course, AI will eventually get good at both of these tasks. Three years ago when ChatGPT first shown promises of intelligence, a new human skill arised and quickly gained public attention: prompt engineering (PE). It looked like the inevitable skill of the future – colleges were starting new courses to teach PE, companies are adding PE questions to their interviews, many of us put PE into our resumes under "skills", right next to "Microsoft Office" and "communication". Everyone was learning how to write prompts so that the AI models will actually "use maximum effort" and "check their own work". Now in the year 2026, we ask AI to write prompts. Many of us in mathematics and theoretical computer science skip this step even, and just give AI a problem statement. We then see the models deciding to spin up subagents for reading papers, exploring different approaches, writing reports, etc. We observe that these models are getting better and better at "one-shoting" the problem, meaning that it produced the solution from just one human prompt. The "old" ways of prompt engineering (having existed for less than three years) got replaced by "orchestration". We now talk about how to set up a hierarchy of agents so that an entire workflow may be automated. Soon enough, however, AI will get better at orchestrating itself. Many of our existing orchestration setups will become obsolete, much like our old prompts that start with "You are a world-renowned expert in mathematics…".[9]

On first sight, this looks like a failure story of trying to find human roles in productive tasks as AI continues to improve. However, it is worth inspecting how the particular role of prompt engineering appeared and disappeared. PE arised from us wanting the AI models to be more effectively helpers. We want them to check their own work (back when they liked to say "3+5=4"); We want them to pursue tasks thoroughly; We want them to give us complete deliverables. These are human objectives that we've distilled from our daily tasks, and we want to impart it onto AI. Soon enough, the AI companies trained their models on our objectives, and our objectives are largely fulfilled. The same story will repeat with orchestration, and may repeat for the non-human-native parts of most of our tasks. This includes, for instance, exposition of AI mathematics. A key point made in the open letter by mathematicians is that the AI-derived proofs of important results are unreadable, and as a result do not convey new knowledge or understanding. I expect this to change rapidly, now that we've set the human objective of "write better proofs that convey knowledge". It is foreseeable that within a

year the models' mathematical writing abilities will improve significantly, much like their abilities to write their own prompts.

This story of setting new human objectives and aligning AIs to fulfill them may be the model of future human work and production. We will start, as before, by finding a new human objective and deciding how much resource to allocate. Then we will work with AI to find the right paths of fulfilling that objective, aligning AI at various steps of the process. Once AI become good enough at the task, our objective will be largely fulfilled and we can move on to other things. Assuming that we truly have a practically infinite number of objectives, this model will be sustainable. The age of developing a specialty and building a life-long career around it may be gone for most of us. Nevertheless, the ability to think, communicate, share knowledge and orchestrate, the characteristics of human intelligence, will not be obsolete so long as we have a human world, built around human objectives and the pursuit of them.

This future can be extraordinarily promising and exciting. Consider, for instance, the transformation of math and science under the impact of AI. The field of mathematics is full of important problems that are open for decades, some for centuries. In comparison, our human lifespan is still limited to less than a century. Isn't that disheartening sometimes? Wouldn't we want important problems to be solved faster (while imparting knowledge at the same time), so that we can move on to even cooler problems? In the middle of me writing this essay, OpenAI made a new announcement that its internal model has "now resolved more than 100 long-standing open problems across most areas of mathematics". If all of them can be verified and the AI models can convey new understandings better, wouldn't that be extremely exciting? So many things that would have taken so much longer can now be done in weeks. Beyond mathematics, imagine AI's impact on physics and chemistry, medicine & human healthcare, climate change, clean energy, and so on. If AI truly delivers on its promise, our human civilization will experience the greatest leap in capabilities ever.[10]

Our human world

These are turbulent times, yet I have faith in our collective humanity. The creation of artificial intelligence may be the gravest challenge we have ever faced. Nevertheless, the problems we need to resolve are the same ones which we have dealt with throughout history: that of resource distribution and alignment of human objectives. This time, however, we must act as one collective. AI *is* an existential threat. It is an ever-rising ocean while we are now stranded on taller or shorter islands. Keep fighting one another, and we will all drown. Now is the time to collectively imagine, to collectively build a floating world on the ocean, a human world that rises with the power of AI. In the pursuit of progress, we must not lose what's human.

Acknowledgements: Claude Opus 5 helped with formatting texts. The essay is written by humans and humans only, without discussions with AI models. I felt that writing this essay is a human-native task.

Notes

---

1. Here the word reasonable is used in the same sense as in Prof. Terence Tao's essay on "Mathematics in the age of AI". ↩
2. I put quotation marks around the word "interesting" because it is not rigorously defined. ↩
3. A task is gradable if we can assign meaningful scores to different completions of the task. These scores can be subjective to one person's opinion or a collective's opinion. ↩
4. On OpenAI's solution to the Navier–Stokes problem, there is debate on the originality of the solution itself. For reference, see Prof. Tristan Buckmaster's statement. In the spirit of worse case analysis, here in axiom 3 let us assume the existence of AI models capable of solving millennium problems by coming up with *genuinely new ideas*. ↩
5. Note that the definition of an AI model is also debatable. For instance, if I make a local copy of a public AI model and let it run only on my PC, does my local model become a new entity? ↩
6. After all, if my job all day is just to tell AI "do something great" and pay for the damages when it messes up, then that's not a very fulfilling life. ↩
7. On a high level this concept comes from the following argument: in the capital-labor model of the economy, labor will become incredibly cheap due to AI and all that matters from now on is capital. Therefore, those who own capital will become more and more powerful, while those who don't will become the permanent underclass. ↩
8. I tried this with ChatGPT-6 and here is its output. It put together some textbook methods and cited a few 2026 papers as new directions of research. The construction is mediocre and conservative. ↩
9. At this point, the models *know* that they are world-renowned experts in mathematics. ↩
10. On my wishlist I have immortality and interstellar travel. Sure enough ChatGPT 77.7-Genie can work that out with a few prompts, right? ↩