Existential AI Risk

What It Is & Why It Matters

Introduction

Existential AI risk isn't something that's far off in the future or limited to some evil robot destroying mankind, contrary to what many of the leading lights in AI innovation have told us.

AI is changing your existence right now. It is affecting how you live, work, and see yourself and the world. The greatest risk is that you're not aware of its impacts and implications.

Of course, you know about efforts underway to preclude or mitigate risks of bias, bad data, hallucinated sourcing, and unreliable performance executing particular tasks in individual use cases. But even the most perfectly functioning AI presents risks that its success will lead to outcomes we didn't anticipate and for which we're unprepared, as even its smallest impacts on individuals will add up to massive impacts overall.

It's a huge problem because there are risks associated with these changes and outcomes that are both purposeful and unintended. There are risks that good things might lead to bad ones, or yield effects for which you're not prepared and may not like.

The ultimate risks aren't that AI fails, but rather that it succeeds and that we're shocked by the world it gives us.

To make matters worse, the changes that AI is giving us are irreversible, and they're happening as you read this sentence. Changes come with every new AI development or deployment. It's relentless and present tense.

If we don't acknowledge and understand the risks associated with these changes, we doom ourselves to suffering their consequences. If we only see the opportunities for AI to improve our lives, we risk missing the ways it might make our lives less, well, livable, or certainly less familiar to us.

Those changes could ultimately end up destroying us, too.

We don't need to worry about a future sci-fi moment. We need to focus on the changes to our existence right now.

This white paper decodes the three major existential AI risks we face, and then provides a three-step action plan that every consumer can apply to reduce those risks for themselves, their communities, and the world at large.

Existential risk #1: The unemployment tsunami

Summary:

We are wholly unprepared for the existential risk of vast unemployment or underemployment. The productivity revolution promised by AI can only be realized if it replaces work done by people, and since AI becomes iteratively smarter and more capable with every passing moment, the breadth and depth of jobs it can do will only increase over time. History gives little hope, as it took two or more generations before the massive job disruptions caused by the relatively "dumb" machines of the Industrial Revolutions in the 18th and 19th centuries were resolved.

Almost two-thirds of all jobs in the US and Europe could be reduced in scope and/or billable time by automation and generative AI, and one-fourth of them could be replaced entirely according to this study. Most other forecasts are equally stunning and they're estimates are based on AI's *current* capabilities.

It's an existential risk for which none of us are prepared; we're distracted by promises of increases in business productivity as work output goes up (robots don't need sleep) and labor costs go down (they're CapEx, not salaried, so investments can be depreciated instead of pay raised).

Investors and consumers will benefit while workers will suffer is a distinction without a difference. Different people on accounting ledgers. Same people in reality.

Another distraction is the claim that past technological innovations ultimately created more jobs than they took away, so we have nothing to worry about. This belief is suspect, at best, for at least three reasons:

- It glosses over the quality of those new jobs, which technology innovation doesn't specify, let alone guarantee.
- It ignores the time it takes for those jobs to appear, and
- It doesn't take into account of the associated impacts of such transformations.

The story of hand spinners in Britain is illustrative of these problems.

Around 1770, about 20% of the women and children in Britain spun wool and cotton into yarn for use in textiles. As machines replaced them, few to no new jobs were created for them, and they and their descendants remained unemployed well into the 1830s. When they did find work, the jobs were often of lower quality and pay.

Household incomes suffered, though most analyses of prosperity simply exclude the impact of women and girls. Same goes for impacts on specific types of workers and where they lived.

The productivity gains delivered by the spinning jenny and other industrial machines were unevenly distributed, to say the least. Living standards for many declined generally in the 1800s (their malnutrition measured by the reduced height of their children, for instance), even as industrialization created vast new fortunes for the few.

If the history of technological transformation is the best model for what AI's impact on our work lives will look like, we're in a lot of trouble.

If AI is different, will it be better?

There is no reason to believe AI's impact on our work lives and economies overall will be any smoother or fairer than past tech transformations, though there's a lot of hope. Usually, that hope is promoted by those who'll stand to profit most from it.

Its impacts may well be felt quicker and keep happening because of the nature of recursive learning. AI are machines that evolve over time. This has implications for the types of new jobs we'll see emerge in the wake of AI taking over old ones.

There'll be no such thing as a job that's "safe" from AI, and the half-life of those that are may be shortened. This will make it harder for people to train for specific tasks, since there'll be no way to assess if/how long they'll remain the purview of human workers. AI dominance in a particular job category will also make those human workers it displaces harder to retrain and find new work, as their skills will be useless (if switching industries will make their skills relevant again, it will only be a matter of time before AI catches up to them).

It will be much harder to imagine, let alone train for jobs when AIs are working harder, longer, more consistently, and more collaboratively to qualify for the same openings.

Productivity gains will appear because jobs for people disappear.

The math is brutally simple: Robots can be more productive than human workers, so the incentive for employers will be to hire people only when AI can't do the job, and then replace them as soon as an AI can take their places.

A permutation of the argument citing past technology transformations is that this process will create new jobs for people, especially ones that hadn't been possible before the advent of AI (and which we can therefore not imagine because we're not "there" yet).

There's some support for this line of reasoning. There were no coders before computers were invented, no pilots in the era before commercial aviation. Jenny operator jobs replaced hand spinners, and there were likely positions created for overseeing machines that hadn't existed before the machines did.

But it's also terribly risky to assume past patterns will repeat, not to mention that those experiences were not necessarily beneficial to the greatest number of people anyway. Add the fact that AI will slowly encroach on even the newest, most human-centric positions, and the argument for new job creation seems less of an opportunity than part of a problem.

The problem is expanded by arguments that the real benefit of putting people out of work is that they won't have to work anymore. AI, like slave labor in past eras, will toil so that a newly-created leisure class will be free to pursue their other interests.

History isn't too kind on this assumption, either.

Debates about consciousness and personal autonomy fade away when you consider that prior polities based on slave labor were quick to qualify their servants as "less than human." It was a convenient excuse for enjoying productivity gains and immense profits when people were treated like machines.

An economy based on AI would be little more than a slave economy, as

Generative AI has already given us machines that we treat like people. Recent tests like this one suggest that we are at the cusp of not being able to tell the difference between machine and human being (the Turing test establishing the premise that if we can't see the differences, the distinction is irrelevant).

This raises incredibly thorny questions about AI having rights and perhaps even the capacity to suffer pain (emotional duress, for instance). It is not science fiction. We risk already going down the road to creating and then enslaving intelligent beings. We don't talk about this existential risk.

Another aspect of the jobs problem is free time, or the lack of it.

Early last century, John Maynard Keynes predicted that innovation and productivity would raise living standards so much that people would choose to work less and enjoy more leisure. He estimated that a work week would total about two days.

Keynes got it wrong. The causes are complex but it's possible that well-off people work harder to stay competitive, strive for more income to pay for all of the new things that innovation and productivity provide, and might even simply like working because it gives their lives meaning.

People with less income work harder, too, only because it's more costly to survive, let alone thrive. There is significant evidence that productivity gains don't translate into leisure time, regardless of what workers may or may not choose.

Let's assume that enslaving AI and putting it to work in our steads wasn't risky and that it would somehow magically give us leisure time. What would we do with it? Proponents for AI have suggested it would free us up to pursue art. So, just think about a healthy portion of the people you know spending their days reading their poetry to you, or describing why the smudge they've painted is really a flower.

Yeah, bad idea.

The impact of AI on our work lives is an existential risk.

Long before an evil AI decides to annihilate all of humanity, the pursuit of productivity will forever change our existence. It will be a huge wave that washes over and through every aspect of how, where, and when we work, and then change it again and again.

We have nothing but fantasies and dimly remembered lessons from history on which to rely for dealing with it.

It's a textbook description of existential risk.

Existential risk #2: The end of uncertainty

Summary: The idea that existence is determinist and thereby knowably predictable is at the heart of data science and the promise of AI benefits: reducing car accidents, improving medical diagnostics, making political promises and marketing messages harder to resist will be just some of the ways our lives are changed by ever-better informed algorithmic management. What we'll give up — experiencing surprise, novelty, chance — comes with risks that are far harder to assess and value than the benefits of reducing those qualities in our lives.

Pierre-Simon Laplace was an 18th century French scientist who did a lot of groundbreaking work in math and statistics. He published a famous argument illustrating the promise of determinism in 1814, proposing that if someone knew the location and momentum of every atom in the universe, they'd be able to reconstruct the past and predict the future.

Quantum mechanics blows up his idea (you can't measure location without changing momentum and visa versa) as does the second law of thermodynamics, but the concept is profound and underlies the presumptions of data science.

More data means less variability, data quality issues notwithstanding. This makes surprises merely the lack of visibility into causes, mysteries are the absence of insights, and chances the result of unknown certainties. Novelty fills the space left empty when we don't know where, when, and how to look for causality.

Uncertainty is the gap between possibility and probability.

This is the gap that AI promises to close, and the overly-hyped benefits of autonomous driving are a good example of how.

Just imagine if the person sitting behind the wheel possessed the driving experiences of millions of drivers collected in every situation imaginable. Then think of that driver connected to hundreds of sensors that gave it real-time awareness of performance status, and kept it connected to every other AI driver currently on the road. Now, connect all of those drivers with the road infrastructure conditions, pedestrian position and movement, and weather conditions.

Random traffic becomes choreography.

Accidents become impossible because there are no surprises, no lapses of skill or attention since the integrated system predicts changes and adapts immediately to the occasional exception, should one occur. And the system collects, analyzes, and shares information so that it gets smarter with each passing moment.

It will be bad news for auto insurers, and such a perfect system is the endpoint not just for vehicle autonomy but most any other industry (think AI intelligence connected to omnipresent sensors/surveillance and other AI):

- **Financial markets** will function efficiently because there'll be no delta of opinion when all facts are available to AI at every instant (i.e. no arbitrage within or across markets). Money will be made based on business performance.
- **Health** diagnoses will yield predictable, reliable findings that enable proactive treatment. Money will be spent on the highest likelihood successes (more will be spent on those lower down the scale).
- **Job recruiting** will match people with positions with 100% reliability of performance outcomes (assuming jobs are available to people). Money will be spent on recruiting and then training only the recruits with the potential for the greatest productivity and longevity.
- **Education** will be outcomes-based, too, so students will be matched at an early age to programs that maximize their qualifications for particular jobs (again, should they exist). Money will skew to students who exhibit the highest potential.

- **Entertainment** will mean that every consumer is presented with content guaranteed to entertain them. Money will be spent to provide content that migrates unique, personal interests into shared expectations (reusing content will lower production costs).
- **Consumer marketing** will mean you'll never see an ad that doesn't drive you mad with interest, or get an offer for something that doesn't have a high likelihood of compelling you to purchase it. Money will be spent to maximize these outcomes.

Living becomes a predictable system.

It may also be bad news for us, insomuch that we will serve as resources for the same technologies and services used to grant us benefits (safer driving, better shopping deals, etc.). Giving up uncertainty in our lives will be what we "pay" for the enrichment of others.

But that's a good thing, right? Better markets, more efficient expenditures on education and employment.

No, it's a risk. A huge one.

First, it doesn't value uncertainty as anything more than a cost in our lives that can be reduced (and therefore monetized).

But what about the value and role of novelty? Do surprises have any positive value? Is there benefit in our lives to not knowing what the next moment will bring? Since the reduction of uncertainty in our lives can never be 100%, what is the value of those rare but potentially life-changing moments of happenstance or kismet?

Not everything will be predictable, but the vision for AI in our lives has determined that the value of such indeterminacy is zero. Instead, it will take those moments away, or make them far less likely, in exchange for making our lives more efficient. What's one person's loss of a truly random moment in exchange for an increase in driving safety overall?

The problem is that the value of our subjective lived experiences will be subsumed by the objective measures of statistics.

The whole will be more valuable than the sum of its parts.

Second, the promises of autonomous AI functioning in our lives depends on everyone opting into it, since any outliers will interject uncertainty into the system. This will require additional costs to mitigate those effects.

You may opt-out but the system will still include you.

We may run the risk that we'll be charged for such insouciance, paying a premium to take control of the steering wheels of our cars (presuming they still possess them) or electing to each an extra fry and therefore throw off an insurance company's assessment of risks to our health. The freedoms we take for granted, however imperfect and self-destructively we exhibit them, will be monetized.

We could end up paying for freedoms.

One could argue that there are always costs associated to decisions and that the bad choices we make incur costs not just to ourselves but to the infrastructures in which we operate. Health care costs go up to include coverage for the sickest, and car insurance rates reflect the fact that many of us take no responsibility for driving safely.

But it's not a binary or on-and-off determination, is it? Who's to say a little uncertainty or risk isn't just a good thing in our lives, and that much of it poses little to no impact on others? One more fry? Really?

Well, the companies that use AI will decide it for us, yielding incomprehensibly huge implications for our conceptions of personal autonomy and free will, none of which are discussed publicly these days.

Third, no AI system will work perfectly, which means we'll trade "predictable" risks for those that the system can't foresee and, therefore, we'll still suffer them.

Remember Laplace. No system can be 100% reliable because not only does it have to incorporate every conceivable input but also any impact from triggers that are, well, inconceivable. Known unknowns are one thing. Unknown unknowns will always be a potential problem lurking just outside the authoritative view of AI.

That means we'll risk being surprised by surprises.

We'll expect the car to safely drive itself until it doesn't (remember all of the autonomous car tests that kill drivers and/or pedestrians when something that "shouldn't" happen happens?). Health screeners will fail to include some obscure fact and yield an imperfect or even dangerous diagnosis. Markets won't function wholly efficiently because there'll always be insider information that stays inside.

Companies and governments will reap the benefits of AI-informed services while we bear the brunt of its inefficiencies.

The impact of AI on our well-being is an existential risk.

Removing uncertainty from our lives will fundamentally change our existence, for better and for worse. It is a particularly troubling risk because the analyses are wildly weighted toward valuing the objective statistics of well-being at risk of sacrificing the subjective value of lived experience. Worse, its promise will never be fully realized and will bring additional and different risks into our lives.

We don’t talk about these risks, which makes them more existentially threatening.

Existential risk #3: The robot in the mirror

Summary: What happens when you can’t take your own humanity for granted anymore? Our senses of self and our shared religious beliefs depend on presumptions that we are unique in the world because we possess qualities unavailable to other living things. Once AI proves itself capable of doing what we do in ways that appear identical to the way we do them, it will risk undermining the core tenets of our psychology and theology. Are we prepared to give up who we are to realize what we will become?

There are at least 5 beliefs from which we human beings choose to define our humanness. They are things we can’t prove let alone explain, but we believe one or more of them with such deep conviction that they feel like facts:

- **We’re conscious**. All human beings have some awareness of their own existence as discrete entities, so this attribute is hard to refute. Philosophers call it a “hard problem” because science can’t pinpoint where it resides or describe how it works. The thing is, it’s fact that we know that we know things. There’s an “I” separate or above (or whatever) whatever I think, say, or do, so a mind is something more than just the summation of my sensory inputs and electric flashes in my brain. Or not.

- **We have souls**. The idea that we have spirits that animate us is thousands of years old and we rely on it for our assumptions about our earthly authority and hopes for eternal life. It differentiates us from other living things (it has been long assumed that animals don’t possess souls, which would make Heaven a kinda sad place) and both separates and elevates us from non-living things (reducing the Earth to a collection of inert resources that we soulful types can exploit to our advantage). Yet proof of its existence makes theories of consciousness seem like settled science.

- **We have free will**. This has always been a contentious topic, as the dynamic of individual intention in a universe in which some Higher Authority has prescient knowledge of all things that have and will happen yields some thorny and gymnastic philosophical accommodations. We human beings certainly believe that we are cognitively in charge of our own actions, the influences of hormones notwithstanding. If it’s an illusion, it appears wholly and consistently real.

- **We can love**. What’s love? It’s something different for every person in most every circumstance, but the experiences share are some broadly consistent qualities of affection, empathy, and altruism (both positively and negatively, so more variability there). It’s something we believe surpasses our base instincts to include aspects of our consciousness and souls, which means it’s a vague idea based on our imagined selves. But no other things on the planet can imagine it like we do.

- **We’re aware of our own mortality**. Knowing that death is inevitable brings varying degrees of meaning to our lives. It informs our perceptions of ourselves, others, and the world around us, influencing our beliefs about purpose and even our ideas about time itself. No other living thing has this awareness — house cats don’t know that their days are numbered, let alone that they are individual beings — which make us not only unique but, along with our attributes, special.

All of these beliefs are misplaced or untrue, according to data science and the experts building AI.

There’s no hard problem with consciousness because consciousness doesn’t exist, at least not as some stand-alone feature of mind. Consciousness is an artifact of the biomechanics of our brains; there’s nothing else on which it can be based. It is an outcome of physical properties, maybe just an artifact our physical systems create to help manage their integration. But our our perceptions of mind (or our minds perceiving) are nothing more than a mirage. We aren’t just *like* machines, we *are* machines.

This belief obviates the need for souls, which also can’t be traced to an internal organ or biological process. It’s an easy lift to consider writing code for perceiving and describing the presence of a spirit. Mimicry, not creation. A soul is an idea, not a thing.

Code and data are stand-ins for intention and personality, or rather the latter are responsible for the former. Our belief in free will is the result of our inability to name, track, and correlate the internal and external influences on our decision-making. There is no “us” beyond the aggregation of those influences which are processed by our biomechanical brains.

We think we’re in control of our choices but our choices control us. This includes our experiences of love and our awareness of mortality. More collections of data pushed through the algorithms hardwired into our brains and bodies.

According to the folks behind AI development, the question we should be asking isn’t how to prove why we’re different from every other thing alive or inert, but rather how we’re actually the same. They don’t need to invent an AI with these so-called human qualities because those qualities don’t exist. They’re imaginary.

This worldview is what animates the pursuit of an AGI — for Artificial General Intelligence — which presumes an AI will operate just like a human being, limited by no constraints on what it can perceive, understand, or do. The bar for achieving it is far lower than you’d think, since people are already so much like AI.

Welcome to the world as a robot.

The implications for our loss of our uniqueness are immense. Each of us will have to debate the nature of our very existence, along with what we're supposed to do with it. It will change how we see and treat one another, and bring into question the institutions we've created to nurture and support our uniqueness.

It will also challenge us to discover how to treat machines that are effectively identical to us.

Will it be acceptable to build an aware AI and then doom it to a lifetime of servitude? What about building soldier bots? What types of responsibilities for actions will we assign to civilian AI? Will they possess rights, such as an ability to own property or borrow money? Will they vote and run for public office?

What will these questions do to our personal conduct, our shared spaces, and the markets and systems on which we rely for functioning societies and economies?

These aren't outlandish questions; what's outlandish is that we are aggressively pursuing a future in which robots and people will be all but interchangeable and we're *not* asking these questions.

This existential risk is greater than the risk of our annihilation because we'll have to learn to live with its implications. We are the architects of that risk (and cause its increase) because we are inventing a future in which we're blind to how it will change us.

One person can't make a difference

Summary: When something is presented as a *fait accompli*, it's usually a purposeful effort to close a deal that has yet to be completed. In business, it's called a *presumptive sale* and the progress of AI innovation is just that sort of topic: It's too fast, too complicated, and already too far along for any of us to do anything about it, so just get used to it. We should see this as a warning sign and a dare to stand up to it. Fortunately, there are at least 3 things you can do to make the existential risk of AI more transparent, addressable and, in doing so, survivable.

CyberConsequentialist philosophy suggests that individuals involved with AI should be responsible for their actions. It's not a radical idea, obviously, but it's usually ignored in conversations about regulating AI development and managing risk, as if the development of newer and more powerful AI tools is an unrelenting tsunami, or a force of nature bigger than any of us can imagine. So, no individual can do anything about it.

Tech happens. Get used to the risk.

Of course, this is a self-fulfilling prophecy, which is why it's promoted by the businesses that hope to make the most money from selling and/or using AI. It also amounts to a "get out of jail free" card for individuals directly involved in that development. Coders, modelers, planners, testers, and anybody else to touches the backend of AI tools can't be held accountable for what outcomes their work might deliver. They can't control it and shouldn't have to anyway.

That's because their work is agnostic to any labels of good or evil, so if something they do proves to be problematic, regulations will reign it in. In the meantime, it's full speed ahead!

Tech happens. Get used to the risk.

But what happens if that's not true, or it's not the best way to address the challenges of existential AI risk? Considering the implications of the uses of AI recounted in this white paper, which by no means should be considered comprehensive, there's good reason hope that we can act differently and see different outcomes.

This is especially true for users, since the valuations of many of the business models on which AI relies are based on user adoption and engagement. Consider these actions:

First, **always remember that understanding new technology is never about understanding new technology**. AI is already changing how we work and live, and it's going to transform everything we know and believe about ourselves and the world. So, are you disqualified from talking about those topics because you don't know how to write code?

The implications for AI are far more important than the functional attributes that will produce them.

Every current and potential user should ask questions about those implications, both directly and indirectly. Dare to pose questions that are incomplete or imperfectly worded. Reach out to others to get their opinions and add them to your understanding. Ask your elected officials why they aren't asking the same questions.

Most importantly, every user needs to be willing to challenge themselves and their understanding of how AI may or may not contribute to existential risk. Push back when you are told "that's not how it works" or your internal editor tells you to give up an inquiry.

Consider using this white paper as a a reference guide for questions to ask when you are intrigued or worried by a potential risk.

Second, **don't be a willing guinea pig**. AI, like other digital tools that have come before it, relies on usage to learn and improve. When you browse or buy something online, the technology gets smarter. Same goes for using Internet search or driving your car. Data is generated, collected, analyzed, and put back to work providing services.

When you play with AI, it's playing with you, too.

Every time you use an AI tool, either directly or via a service in which it works in the background, you are potentially adding to the risk it presents to you and the world. That doesn't mean you need to shy away from it, but at least consider the real utility to you for that interaction and, when the cost/benefit equation for you is equal or simply too vague to assess, consider stepping back.

Or ask more questions, as explained earlier.

Another cliche that gets thrown around when anyone resists the onslaught of some new issue or phenomenon is that "one person can't make a difference." That is a lie, statistically speaking. There is no question that every interaction makes a difference. The only variable is how and when the consequences will become apparent.

Third, **hold the makers and promoters of AI accountable**.

It feels like distant history, but it wasn't too long ago that a passionate Greek Chorus promoted the benefits of social media: it was going to change the world for the better and every person and company should embrace it with open arms.

Where are those outspoken advocates today? The businesses are established. Profits are minted on top of profits. What was visionary has become standard, along with all of the negative consequences we never heard about.

You could imagine something similar happening with AI. By the time some of the risks noted in this white paper come true, its promoters will have absconded behind corporate and institutional benefits with their business contracts and consulting fees, leaving the rest of us to contend with the risks that have become reality. Once AI's existential risks become reality, we will have to adapt to them.

The time to take action is now.

We consumers have a powerful tool at their disposal: *our wallets*. Each of us has the capacity to support or retard AI's development, depending on how we see the risks it poses.

You don't have to buy from companies that replace human employees with AI. You don't have to invest in them. Better yet, you could demand more transparency on their activities and make your decisions thereby (i.e. combine this activity with the other two suggestions in this section).

You can get more involved with organizations that focus on human relationships, whether religious, work-related, or social. As AI becomes more like us, it should encourage us to get to know ourselves better and perhaps find deeper meanings in the assumptions that we once took for granted.

Are we really no different than machines? Are there new ways to experience and share what makes us uniquely human? Can we reaffirm what and who were are instead of simply letting technologists do it for us?

We can lessen the potential for existential AI risk by increasing our certainty about ourselves.

It won't be perfect. Nothing ever is (well, except an all-seeing and knowing AI, presuming one could ever be built and maintained). But awareness combined with action will help mitigate existential AI risk.

Those risks are now, ranging from AI changing if and how we work and how we interact with our world, to how we see ourselves. We don't have to wait for it to annihilate us before we experience its existential risk.

Existential AI risk isn't about the end of the world…it's about ending our world as we know it right here, right now.

We risk everything if we don't do something about it.