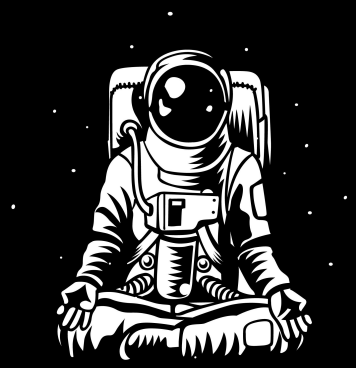


Spiritual Telegraph

CyberConsequentialism

A New Ethical Framework
To Mitigate Existential AI Risk

Introduction

You are responsible for the AI you're developing or deploying.

This isn't a theory or belief. It's a fact. What you are working on right now will have an impact on the world. After all, that's the point, right?

That impact might be large or small, but it'll be something. There will be consequences to your actions, whether intended or not. You may see something the moment you finish a task, or an impact might emerge years from now.

The consequences might appear distant intellectually or technically from what you did originally, too, but there will be consequences that belong to you, at least in part.

A small insight now could trigger something big and wonderful later on. A great accomplishment could lead to nothing or, worse, something bad.

The equation is simple: You do something and something else happens, and then something else. Over and over. Innovation is an iterative function.

Everything you do today will have consequences tomorrow. It means not only is there a chance that you are doing something great for the world, but that there's a greater than 0% chance that you will contribute to destroying it, too.

It's a fact. Circumstances will demand, businesses will intervene, governments will regulate, and the laws of physics will dictate how your work morphs in an endless series of forms and directions.

But you will still be responsible for whatever comes from your AI work, in some way.

You can be unconcerned but you can't be uninvolved. This pamphlet is a roadmap for an ethics that will enable you to maximize the positive consequences while minimizing the negative ones, including not destroying all life on the planet.

The dog will never catch the car

Summary:

Government and markets can't adequately assess the impacts of AI innovation beyond its immediately visible "functional" qualities, like reducing bias, protecting user privacy, or the profits to be made from monetizing it. The truly end-of-the-world stuff is invisible to them, which makes it more likely it'll happen if we rely solely on them to mitigate that risk.

Government can't effectively regulate AI. You know it, as does every industry celebrity who claims otherwise.

The fundamental problem is that government is always chasing development and deployment. Legislators and regulators are only aware of what's presently known or available to them. In other words, they don't know what they don't know, and existential AI risks are unknowns, by definition.

They're not experts, and any third-party expertise gets filtered through their first-hand ignorance. Most of that expert counsel is also suspect, usually coming from biased business leaders, zealous NGOs, or philosophizing academics.

Worse, by the time an AI innovation is concrete and presented to them, it's likely already let loose "in the wild." This means it will have changed before they finish their opening remarks at a first hearing. Their studies and expert commentaries will already be out of date.

This leaves government considering legislation or regulation for things that have already happened. It can't preempt innovations it can't imagine.

That's why most conversations about AI risk focus on functional aspects of existing tools, like making sure they're not biased against certain users or violate copyright or privacy rights. Or that they simply work like they're supposed to work. For now.

And then there's enforcement. Unlike controlling nuclear weapons, the inputs to AI aren't finite or traceable like plutonium. Unlike drugmakers, they're not beholden to robust establishments to get their products into the market. Big, complicated mines, labs, and factories are easy to locate and visit.

AI can and does happen with the push of a button in lots of places, including virtual no-places.

So, government not only doesn't know what it doesn't know, it's also too late to do enough about what it does know, and it wouldn't know what to do anyway.

It's not in the business of mitigating existential AI risk. Neither is business, which profits from AI's development and use.

Near-term returns are easier to value. There's no profit for assessing potential risks and taking tangible financials hits for them now. Look to the oil industry for a historical example.

In fact, the risk is that companies fail to use AI in ways that delight shareholders, like using it to fire human employees or get customers to buy more.

It's also easier and more profitable for VCs to spin big, juicy stories about AI and raising money and cash out of their investments, usually before there's any reckoning with actually selling something and long before the implications of their products are visible.

Social media, anyone?

Businesses and markets don't have a clue about AI beyond the opportunities it presents to make money in the short-term. What happens down the road is an "externality" to those interests.

Existential AI risk? Well, the world ends for each of us at some point. Why worry?

Further, existential AI risk doesn't just mean destruction of humanity but also destruction of the world as we know it. AI could blow up work or belief in our cosmos uniqueness, not actually blow up the world.

The often unspoken answer from government and business on this front involves a belief that destroying business or social norms makes room for better things to replace them. The threats aren't threatening because somebody will invent a solution sometime.

This isn't a fact, though. It's a belief, or maybe just a hope.

There's no guarantee that what will follow will be better, let alone as good as what preceded it, or that it will spread benefits to the same or more people who'd benefitted originally. The march of history is not linear, nor is its progress always toward better, fairer, more rewarding outcomes for the majority of us.

So, anyone who relies on this magical thinking as the answer to risk of any sort is ignorant of history and human nature. And it's an excuse for inaction.

Left on its own, government and markets can't effectively regulate AI.

The dog will never catch the car.

Why is it your problem?

Summary:

How is it your responsibility to help ensure good outcomes for your work in a world that doesn't recognize, incentivize, or reward it? Because you have agency over those potential impacts that nobody else possesses, and at some point they will impact you. In other words, your self-interest intersects with the common good. You can't escape the consequences of your actions, so you might as well own them.

You aren't just creating the world, you're living in it. By definition, the consequences of your actions, however disparate in scope and distant in time, will eventually impact you.

This, too, is a fact.

Every action has its effects on people and places, immediately and thereafter. This is true whether the scope of that impact amounts to a butterfly's wings flapping in Asia that affect the weather in Kansas, or a permutation of AI that orders the obliteration of mankind.

Therefore, the greater good will impact your good. Your self-interest intersects with others' self-interest, and theirs with yours. The only question is when and by how much.

The answer is usually immediately, and then that intersection morphs over time.

Let's say you choose to drive a car that leaves a trail of air pollution. What's happening at your "small" intersection with the common good:

- **Environmental** — The planet won't melt for years and your contribution to it may be small, but it's not zero. Further, the consequences of your actions on a smaller scale, such as your neighborhood, are proportionally larger. The air isn't as clean. There's more sound pollution.
- **Personal** — You make different decisions based on your driving choice. Maybe you head out to get fast food more often, which will have consequences for your health. Maybe you get into an accident. It's called *context*.
- **Interpersonal** — The people you know may think of you differently because of your choice and treat you differently; some my celebrate your iconoclastic climate change denial, while others might give you the cold shoulder. More consequences.
- **Social** — More broadly, your choice will be evident to many other who may similarly decide their own actions based on it, at least in part. Those decisions might impact you.

Now imagine the AI you're creating for open source or some paid client use. Its actions will resonate across time, just like yours.

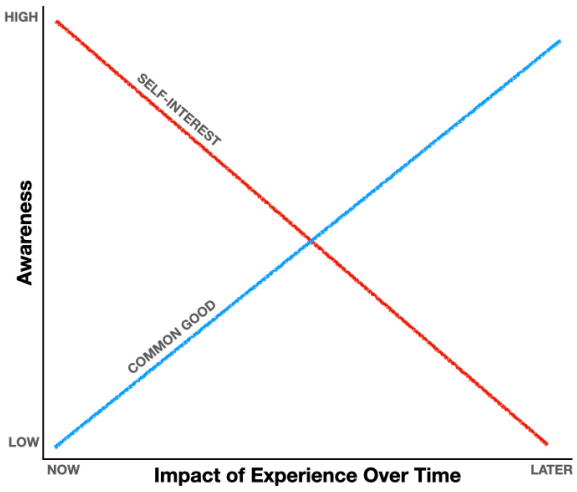
And those intersectional dynamics will change over time as consequences become more sustained, examples more widely shared and viewed, and more follow-on decisions are made. Maybe larger or more impactful consequences emerge. Maybe a lot of smaller ones come your way.

You could chose to not care, but that decision would have its own impacts on the world, affecting how you approach other decisions, treat other people, and affect how they treat you.

The relationship between you and your world isn't a line or binary distinction, but rather a correlation coefficient. Consequences are always shared. Only the ratio changes.

Consequences are unavoidable. Another fact.

If you wanted to create a representative graph of this intersection between self-interest and the common good, it would look something like this (lines represent broad directional movement):



Your self-interest is most apparent to you early on as you consider the costs and benefits of making a decision. It's easy to weigh benefits since they're near-term goals, while risks are less clear (they take longer to form, usually) and certainly not those that might have no direct connection to the decision at hand, however causally connected they might be over time.

The consequences for the common good are mostly unknowns and ill-defined because they're not presented as your concerns.

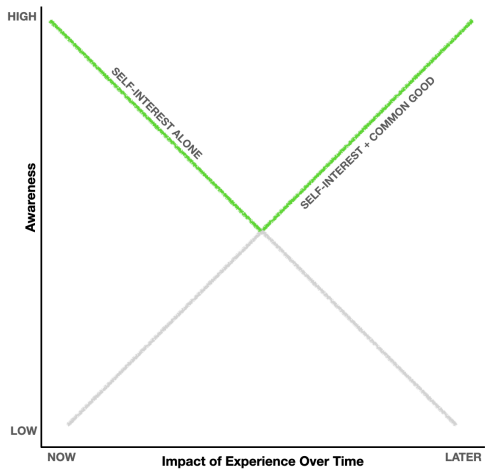
Only they're always present, even if they're not visible or they haven't yet emerged. Your intersection with them at the moment you make a self-interested decision is just opaque to you by default. You don't look for it.

As time passes, and your decision leads to actions which lead to other decisions and actions, that intersection starts to become more clear. The effects of your actions on the common good can become more visible and experientially tangible because they are more numerous, more overt, and may impact you.

As more time passes, the consequences of your actions become inescapable; the effects of your past decisions inseparable from your current and future choices.

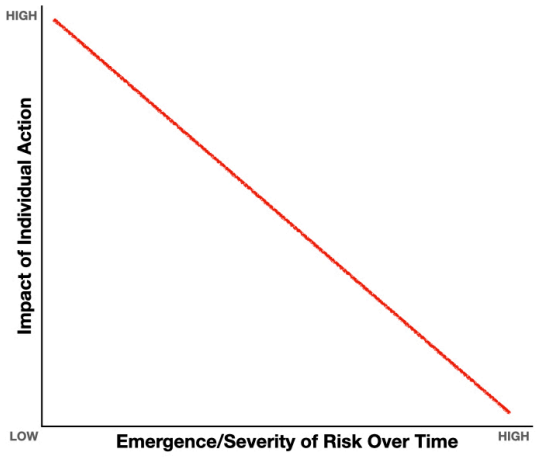
A small contribution you made to the degradation of civil society renders where you live less habitable. A system that uses your code to decide healthcare chooses not to give it to you.

Your awareness of the common good and its impact on your self-interest grows. The progress of that intersection looks something like the the "green v" below (directional representation, not linear progression):



Once you recognize the reality of this intersection of your self-interest and the common good, you should look for it as early in a decision-making process as possible, as that will maximize your potential to thoughtfully impact those subsequent consequences.

The time to preclude or reduce the potential for AI risk is now, not sometime in the future, with any decision, small or large:



It is in your self-interest to choose to gather as much data as possible about both benefits and risks of AI as early and often as possible. No decision is too small, since the consequences of any decision could be large.

You don't need to define the intersection, just recognize that it exists. What matters is what you do with that knowledge.

The ethics in pattern recognition

Summary

Existential risk has already emerged. The data will tell you so. A development/deployment approach that recognized this truth could enable more ethical decision-making. The challenges of building such models are immense but suffer most from a lack of intention versus any limitations to the way existing models are designed. The philosophy of trying to accomplish such goals is called **CyberConsequentialism**.

It's a big ask to take responsibility for existential AI risk. It's uncertain and far off, and its causes are disparate and vague. Unless you're a budding Bond villain, your contribution will likely be unconscious and unintended.

Only that doesn't change the reality of what's happening right now. The causes of every possible existential AI are in today's models, or will be added tomorrow. It's not magic or a mystery. If AI does something horrible, the only surprise will be realizing that we didn't notice the cause(s) when we could have done something about them.

Today's incremental contribution could have enormous consequences over time, for good or bad.

This means that there's a deeper, more fundamental ethics embedded in your actions and their outcomes. It's not something that you're told, or a set of rules with which you

need to comply. It's in the data. The patterns and predictions are in your data, waiting to be revealed.

It's a fact.

The philosophy built on this fact is called *consequentialism*. It's a philosophy of data, really, as it came to the fore in the 1800s in England just as statisticians were discovering the strange power of data to reveal patterns in farming, population growth, commerce and, in a famous example, tracking the source for the spread of cholera in London to a single water pump.

Since we can know the consequences of our actions, we can be held accountable for them. Consequentialism says that moral or ethical actions are those that maximize what's good, and do so for the largest number of people.

Your intentions don't matter. The data on your actions and their consequences will tell the story.

Collect the data. Process it. Clean it, and then analyze it. Ethics don't need to be declared.

They will be revealed.

Consequentialism should be a science, with its data trends and patterns the arguments for behaviors that maximize positive outcomes.

Improving Internet search is good. Blowing up the world is bad.

Only the devil's in the details.

As you know, what you discover is informed and limited by the design of your analytic framework. When it comes to revealing and then valuing the implications of any action, that project framework can be a monster, since the potential consequences of your efforts can be far off and lack any direct connection to you or anybody you know. They can appear in an unlimited number of places, times, and ways.

Ultimately, everything is connected to everything else.

Good luck building that model. Then it gets worse.

Taking responsibility for whatever the data reveal requires effectively assigning values to it. This means deciding how one outcome might be better or worse than another, and calculating by how much, only there are no established or reliable units of measurement for making those calculations.

Smart people have debated this conundrum for centuries and its challenges drill down to three core questions:

- **What's good?** How do we know and measure it (and assess its opposite)?
- **Who do we care about?** Everyone alive? How about future generations? How do we define the group(s) we need to address, and how do we comparatively value them?
- **How far do consequences reach?** Again, the data will connect the butterfly wings flapping in Boston to thunderstorms over Cape Town. At what iterative step do consequences become measurable but too slight to count?

But these are data and modeling problems, not problems of ideas or intent.

Consequences can be estimated when they can't be specified. Algorithms can model decisions even if they're imperfect or incomplete. It should be possible to build a rules-based system that recognizes the ethics inherent in decision-making that is inherently reasonable and usable to AI developers and deployers.

A new framework to modeling and executing projects that helped mitigate existential AI risk.

The result is CyberConsequentialism.

The framework for CyberConsequentialism

Summary:

We've already discussed why laws (and markets) not only aren't enough to mitigate existential AI risk, but might very well allow or encourage it. You are in the best position to be proactive but the challenges are immense, just as the solutions aren't perfect or complete. Simply trying to implement them, however, should lead to better, more ethical outcomes (that reduce risk). There are concrete steps that you can take, and then improve and then come up with new ones. CyberConsequentialism favors less talk, more applications.

There are few requests as useless as those that state a goal without defining its meaning or metrics. "Be responsible" or "more ethical" fall into that category. So do "just have fun" or "do no evil."

For any system to function effectively, it needs a framework and rules that define its operation and uses.

CyberConsequentialism is an applied ethics, which means it takes the principles of what's moral — called "normative ethics" — and makes them actionable. If the devil is in the details, he can be battled there, and not in one's mind or soul.

And, speaking of terms, let's explore the implications for CyberConsequentialism in plain English, devoid of buzzwords or technical terms common to computing or philosophy.

Before we get to the rules for applying CyberConsequentialism to AI projects, we need to revisit the questions raised earlier in this pamphlet about the framework for that approach, and attempt to answer them.

First, **what's good?**

Not blowing up the planet or killing everyone, for starters. But those are end points of trends, and we want to determine what's good up to those points so as to best make those outcomes less likely (or at least allow us to adapt future decisions based on the outcomes of our past decisions, so as to continually modify the emergent consequences).

Our assumption is that the more good consequences we can deliver, the bad ones will be less likely, viewed in real-time.

But is new job creation good? What about replacing carbon emissions with greater use of renewables? What about easier shopping? Would taking so-and-so decision out of peoples' hands and making it for them be good?

Isn't every decision a trade-off between "good" and "bad" outcomes?

What about simpler outcomes for tasks that aren't inherently questions of good vs. bad, like helping a factory better meet tolerances for the products it makes? What about things that might be bad in one instance and good in another, like enjoying a cheeseburger or an objectively awful movie?

Now, what about their relative values? A new job trumps that cheeseburger, though in the right situation, the latter may offer more value than the former.

The answer to such questions is in the context of what you hope to achieve. Consequences represent a duality of experiences. Good is contextual. Maximizing it means seeing your expectations against its risks and negative impacts.

There's no good outcome without some bad outcome or risks that may come with it.

But you're not God, and there's no logical or moral basis to assume that you can make judgements to create woe for some people in order to deliver more good to others. CyberConsequentialsm doesn't suggest that you have that power.

But you do have to power to reduce bad outcomes in your pursuit of the good.

How much bad is acceptable for achieving good? The starting point is zero and you should work down from there.

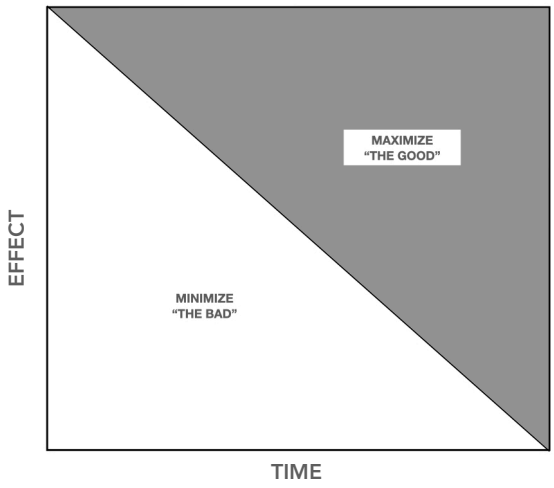
If risks and bad outcomes come hand-in-hand with your AI model or application, you have to try to reduce them to zero and then find ways to actually reduce or remove bad outcomes that aren't of your creation.

In other words, CyberConsequentialsm asks you to maximize the good by minimizing the bad, not simply limiting the latter.

Using this approach, imagine how the ethics of building the first nuclear bomb would have played out had those involved truly balanced data on winning a horrible war in a horrible way vs. the horrors of generations suffering the threat of nuclear annihilation.

Had the designers seen those consequences, and been allowed to act on their ethical conclusions, maybe they'd have abandoned the project? Or not?

Ultimately, "what's good" is defined by how large of a delta exists between its benefits and the costs and risks of "what's bad." You could imagine the duality of good/bad outcomes expressed as a graph over time, something like this (Again, movement is directional, not linear):



Closer to home, your plan to enable daily cheeseburger consumption might be constrained by risks of weight or cholesterol gain, thereby obliterating its benefits. A one-off dinner now and then would trigger different consequential restraints and thereby pass muster, though risks to climate, costs of meat transportation, and related risks might make any cheeseburger project nonviable.

Pity that thought.

No two individuals are going to agree on what's good, or at least not consistently. Your assessment of good and bad outcomes requires real-world metrics of impacts actions and impacts. Calculate "reduced this" and "increased that" with real numbers attached. Gone are vague terms like "enhance" or "improve convenience."

Make the data speak its definitions so that everyone can use the same language. This also addresses the measurement challenge, both in your analysis and ultimate conclusions.

If your deliverable is an objectively real phenomenon — somebody doing or restraining from doing something, some other thing happening or not — then there's a real-world metric for it, by definition. Improving search can be measured in tasks completed or speed of completion, for instance. Car performance improved by, say, increased efficiency in lane auto-correction has a percentage increase or reliability that comes with it.

More cheeseburger consumption can be measured in weight (of burger patties, that is), or by calories or some other physical metric(s).

Similarly, the risks and negative consequences can be measured with the same approach. Emissions generated. Jobs lost. Weight gained. Increase in risk of mistakes and what they might entail.

It's not an absolute calculation and it will change in value over time. Today's cheeseburger joy may be tomorrow's problem. CyberConsequentialsm relies on iterative definition of good because it is based in the data.

Maximizing good outcomes over time should have the cumulative effect of reducing bad ones.

At some point, this may become an equation. For starters, it's a conversation that gets you closer to an ethical goal of maximizing the good.

Second, **who do you care about?**

A philosophy that recommends maximizing the good for the maximum number of people the maximum number of times just screams out for clarity.

That challenge has prompted debate for centuries, often constrained by the lack of technical tools and the supporting math necessary for knowing who was impacted by actions and assessing those values.

Your access to data offers a path to answering that challenge.

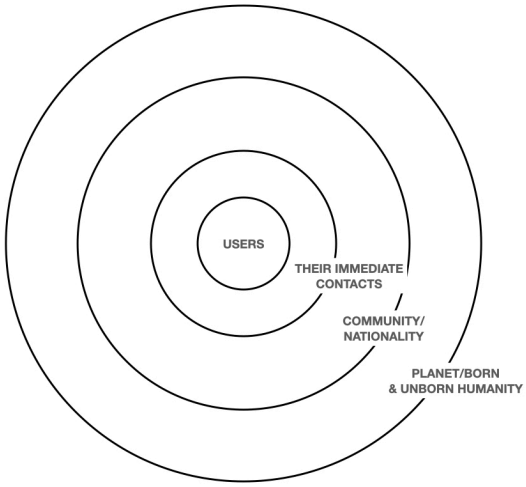
When you define the good you're planning to deliver with your project, you've assembled a universe of people who will experience that/those outcome(s). Behaviors and experiential impacts are attached to the lived experiences of human beings.

So, you can't calculate what's good without creating a population for whom you're providing it.

Where to start? Users, obviously, whether they'll operate your AI or be the primary beneficiaries of its brilliance. Assessing the consequences for their immediate contacts should also be a part of your data set.

Again, you need this sort of clarity to calculate your model for impact, so it's not extra work, per se.

After those two audience groups, however, the realm of possibilities starts getting far larger and more unwieldy. You could image that range looking something like this (Distances between groups are representative and not based on actual numbers):



Considering users and their immediate contacts, you could develop models that told you 1) What you expect their behaviors would be, 2) What the good/bad assessment reveals, and 3) The probabilistic reliability of those forecasts.

"Delivering this good to these people has so-and-so likelihood of happening."

In the CyberConsequentialism model, that last part — reliability — is the most important metric. A higher number should correlate with a greater good, since delivery of experience is more valuable than intending to deliver it.

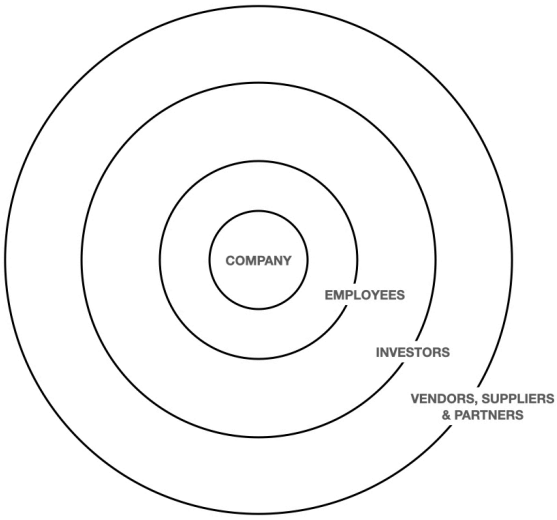
But what about the larger community or the entire planet? Do you have an ethical responsibility to everyone, even people unborn?

A similar challenge emerges from your other grouping of people you need to care about: Your business.

Unless you're doing your AI project for your own entertainment, you're doing it because you're employed or are developing a product to sell to someone. The reality of AI in our world cannot be understood accurately without acknowledging this fact.

Nobody can be expected to act ethically by ignoring it.

The circles of people for who you are responsible/have targeted to make an impact, either as an employee or seller of a product or service to them, looking something like this (Still only a representational graphic not drawn to scale of any sort):



You could imagine developing a model for these groups similar to the ones you did for your user/direct contacts impact analyses. Does your work provide benefits to the company? Then that is a good you've included in your initial design and you value it accordingly.

Employees? Include them, too (this is especially important since many AI innovations result in staff reductions, which are a negative consequence that have to count when considering positive impact of improved sales).

But these are still open-ended models, insomuch that the rings out potential groups that your project could impact still ends up with everyone, everywhere, and every time.

Third, **how far do consequences reach?**

There's a deceptively simple answer to this question: As far as you can reliably see.

“Seeing” requires that you allow yourself to imagine a wider range of consequences than you might have considered when looking at the outcomes of your project on your direct users.

It’s a thought experiment that goes something like this:

The groups beyond the direct targets in my project design remit might experience its effects in the following ways (imagine A, B, and/or C follow-on outcomes). When they experienced X, additionally it might lead them to do/not do D, E, or F. They might see their lives change in so-and-so ways. Reduce something in their lives that was valuable (or not). Increase something bad.

You’ve already calculated the risk of one of your direct targets putting your AI to nefarious uses. That’s part of your good/bad assessment.

This calculation is about passive impacts. People who may not even be aware of your AI innovation. But there will definitely be consequences for them. Your challenge is to do your best to imagine those consequences before you impact them.

These assessments are mediated by the availability of data and the reliability of your forecasts. Yes, it’s another indefinite line separating what you are ethically bound to consider and that which is outside the scope of your work.

Your responsibilities are limited by what you can reliably see. But you are responsible for looking.

The scope of your vision is going to be constrained by availability of data, time to develop your models, cost of doing so and, ultimately, the accuracy of your probabilistic forecasts. At some point, your data set gets too big or complicated, too expensive, too time consuming and too unreliable.

Your project design needs to establish that initial framework by clearly defining what you intend to deliver (the good), to whom and with what expected related consequences.

And then you need to be willing to change it as your product or service is implemented, as consequences that were once vague get clearer and groups you may not even have imagined could be impacted are, in fact, impacted.

The ethics of CyberConsequentialism are in the data.

The rules of CyberConsequentialism

Summary:

If a framework defines the boundaries of your project, rules inform how you operate within it. As such, they need to be specific in order to be actionable. There are six rules to CyberConsequentialism, each with stand-alone value and a contributory impact on the others. It will likely prove impossible to realize each one to its fullest potential; the goal, as with any applied ethics, is to endeavor to embrace both the purpose and details of each rule and therefrom create the best possible design to maximize good outcomes.

Rules are meant to be broken insomuch that even the simplest ones are impossible to follow completely or consistently, regardless of a commitment to doing so. Imperfection is a quality of the human condition. So is guile.

Imperfection and guile are also the sources of greatest worries when it comes to existential risks of AI.

Government regulation can prohibit acts and punish violators, but it can't anticipate every infraction nor creative workaround. Breaking a rule doesn't require evil intent, either.

Government can't establish regulations for aspects or uses of AI that don't yet exist, or which it doesn't understand. Same goes for markets' ability to discount the value of AI investments with the costs of global annihilation.

It just doesn't happen.

This means that we imperfect human beings run afoul of rules by unintentional acts of commission or omission every day. Bad actors intentionally exploit the law's limitations just as frequently, and with more success. Markets reward the promise of AI without recognizing its peril.

The result is that any action brings with it the risk, whether intentional or not, that it will lead to a bad outcome up to the annihilation of all human life on the planet.

The rules that others set for you will not be enough to mitigate that risk.

Your choice is to ignore this shortcoming, and thereby perpetuate or increase its potential effects, or take responsibility for the consequences of your actions and, hopefully, reduce that risk.

The rules of CyberConsequentialsm aren't simply meant to be obeyed or worked around. They're statements of intentionality, of purpose. A commitment to informing your conduct that you've established for yourself.

A statement of your ethical intentionality vs. unethical (or non-ethical) disregard.

There are 6 rules of CyberConsequentialism that you can apply to how you conceive, design, implement, and iterate your AI projects. They serve as actionable tools to work within the framework discussed earlier (i.e. apply the framework first):

- Rule #1 — **Reduce inputs**. It has become common to equate larger data sets with better or faster learning and development of AI capabilities, but those accomplishments come with a commensurate increase in performative variables. In other words, the more complex the machine, the more difficult it is to predict and manage its operation or the consequences of its actions.

  A smaller data set does not automatically correlate with a decrease in performance, however. A CyberConsequentialist approach to designing the data set would have a bias toward using the fewest inputs so as to maximize your ability to understand and track outcomes.

  For you to be responsible for the consequences of your actions, you have to be able to decipher how and why they occurred. It is ethically irresponsible to say "I don't know how that happened."

- Rule #2 — **Reduce scope**. Again, the miracle of LLMs is in large part the fluidity of their applications. Doing a lot of things is cooler and potentially more financially lucrative than doing fewer things. It is also inherently more risky.

  While this is more apparent on the deployment side of the equation, it is also important for development. As you consider the smallest data set necessary for your stated goals, you may want to reconsider limiting those goals. Your willingness to focus on specific tasks will have commensurate implications for your ability to ensure their successful completion. Your CyberConsequentialist analysis out outcomes and risks will implicitly lead you to more specific, refined goals.

  You will reduce risk by choosing to do fewer things better. Doing more things with an imperfect understanding of their implications is unethical.

- Rule #3 — **Increase transparency**. The cliche the nobody wants to know how the sausage is made does not apply to ethical AI; in fact, it's the exact opposite. The more you can share with others about your design, the more likely it is that they can help you ensure its successful performance (i.e. help find and/or manage risks), and the more others can rely on your efforts to reduce risk.

There's a natural conflict between this rule and the legitimate desire of businesses and individuals to protect their work, but there are legal methods available to preserve those claims. Your bias should be to sharing as much as you can as early as possible in your development or deployment process.

CyberConsequentialism means letting others know that you want their awareness and help in understanding and anticipating the consequences of your actions.

- Rule #4 — **Test again**. There are a number of tools that are widely used to test AI (both human and machine-run). You should do more of them.

  This rule cuts to the core of the limitation of government regulation, which could require you to conduct certain tests/do so over so-and-so arrays of data sets and/or time, but still miss demanding a test that you know would be more useful (and that includes potentially more difficult for you).

  Failure of required testing to discover a risk doesn't absolve you of responsibility for that consequence. "Doing what we were supposed to do" is the excuse of bureaucrats and other evil-doers.

- Rule #5 — **Maintain control**. This could be the most contentious rule of CyberConsequentialism because it asks that you ignore the common practice of releasing AI "in the wild" and instead endeavor to maintain control over it.

  Image if a drug designer released a medicine into the world and then took no responsibility for ensuring that it wasn't adapted to do harm instead of good? No manufacturer of any hard good, technology, or simply an article of clothing could escape responsibility for the performance of those products.

  If you've designed AI to do something(s) specific using specific data sets, you should try to keep control over how it is implemented. Your ability to do this will decline over time, but your ethical responsibility will never go away.

- Rule #6 — **Kill switch**. You've seen the moment in many sci-fi movies when the protagonist wants to stop the killer AI from blowing up the world but there's no way to override its autonomy (so they have to smash its head in a vice, or play tic-tack-toe with it). The premise is central to most predictions of existential AI risking the real world.

  There is no reason why you can't help ensure that there's a viable and durable "off" or "kill" switch at the software and/or silicon level of any AI that you create. The best way to limit AI existential risk would be to limit its ability to do harm. Make AI's ability to learn and/or general computable environments always interruptible.

  Full stop.

AI ethics in an unethical world

Summary:

There is no way that any government oversight or regulatory regime can preclude the possibility of existential AI risk and the percentage likelihood that you are contributing to that risk right now is greater than 0. It is quite possible someone else will do something to make that risk greater and/or realize it. There will always be people who act unethically and there's nothing you can do to stop them, at least not directly. But you can do your best to lower the likelihood that such risk or bad outcome is a consequence of your actions. Your ethical actions in an unethical world can make a difference.

CyberConsequentialism aggregates these simple, incontrovertible facts:

- There's a chance what you're doing now will either directly or indirectly contribute to AI existential risk.
- Government and markets are unable to identify or prohibit all of those possible actions.
- Your self-interest and that of the common good intersect; at some point, your impact on the world can and will impact you, either positively or negatively.
- Therefore, you bear responsibility for the consequences of your actions, whether intentional or not, and have a personal interest in mitigating them.

It then relies on data to construct a framework for addressing risk:

- Defining "what's good" by how large of a delta exists between data on its benefits and the costs and risks of "what's bad."
- Calculating the scope of the population for which you want to maximize good by basing in on the reliability of your data for identifying them and the consequences of your actions, and
- Assessing your impact on actions that may be far removed from your initial action by combining the availability of data, time to develop your models, cost of doing so and, ultimately, the accuracy of your probabilistic forecasts.

CyberConsequentialsm then provides 6 rules for specific development and/or deployment activities, all intended to reduce the likelihood of you contributing to AI existential risk:

- Reduce inputs, so you can best decipher why outcomes happen.
- Reduce scope, thereby reducing risk by choosing to do fewer things better.
- Increase transparency so that others know that you want their awareness and help in understanding the outcomes of your development or implementation.
- Always test, then test again, always exceeding the letter of any requirements.

- Maintain control, and stay away from releasing code “into the wild,” so as to keep control over how it is implemented, and
- Insist on a kill switch, whether at the software and/or silicon level.

Being ethical in an unethical world is not easy, nor will it generate the same positive feedback you might get from other activities that are more encouraged or favored.

But you will be right.

If you apply CyberConsequentialism to your work and strive to maximize the good for the greatest number of people, you might keep yourself and your business out of hot water.

And you might save the world along the way.

July 12, 2023
Chicago, Illinois