The Fragile World Hypothesis: Complexity, Fragility, and Systemic Existential Risk

David Manheim

Abstract

The possibility of social and technological collapse has been the focus of science fiction tropes for decades, but more recent focus has been on specific sources of existential and global catastrophic risk. Because these scenarios are simple to understand and envision, they receive more attention than risks due to complex interplay of failures, or risks that cannot be clearly specified. In this paper, we discuss the possibility that complexity of a certain type leads to fragility which can function as a source of catastrophic or even existential risk. The paper first reviews a hypothesis by Bostrom about inevitable technological risks, named the vulnerable world hypothesis. This paper next hypothesizes that fragility may not only be a possible risk, but could be inevitable, and would therefore be a subclass or example of Bostrom's vulnerable worlds. After introducing the titular fragile world hypothesis, the paper details the conditions under which it would be correct, and presents arguments for why the conditions may in fact may apply. Finally, the assumptions and potential mitigations of the new hypothesis are contrasted with those Bostrom suggests.

1 Introduction

The risk of humanity's possible extinction could take various forms, from diseases to runaway climate change to asteroid strikes. It is widely agreed that either extinction or global catastrophic risks leading to most humans dying would be bad, even if the exact relative importance of avoiding extinction is unclear. [40] It does seem to be the case, however, that technology or other modern developments are far more likely to cause human extinction than natural events. [42]

Bostrom has argued that the prevention of potential extinction is a current moral priority. [5] Because of this, Bostrom has researched potential sources of extinction risk extensively. Some past technologies are understood to potentially have such devastating consequences, such as nuclear war. Newer concerns include malicious artificial superintelligence, [35] malicious use of biological engineering, [18] and technologies that could destabilize the international balance that makes the use of nuclear weapons less likely. [16] In each case, the risks of these technologies are being investigated by an active research community.

Bostrom recently suggested another possibility, the ”Vulnerable World Hypothesis,” [6] where a technological advance occurs that will by default lead to extinction. He suggests that this is concerning, but notes that it is unclear if there are in fact such technologies in humanity’s future. It is far more clear, however, that there are already technologies or events which would lead to global catastrophic risks short of extinction, and it seems implausible that more will not be found. Such events would create widespread devastation, could at least severely limit humanity’s potential future advancement, and would at the very least cause tremendous human suffering and death in the short term. [31, 4]

The class of existential risks addressed in this paper is related to several of the previous models, especially Bostrom’s vulnerable world hypothesis, but focuses on systemic risks rather than individual technologies. Systemic risks are those that emerge from complex system failure, where the failure of a single component leads to systemic knock-on effects. These are potentially devastating, and we will suggest that they might lead to an inevitable extinction of the type Bostrom suggests. Not only that, but because of the way these risks emerge, they seem more difficult to classify or estimate than typical risks. Tonn and Steifel suggest methods for estimating existential risks which are primarily suitable for risks that are not systemic or unknown. [45] These methods are unfortunately appropriate for neither type of systemic risk discussed here, and work on understanding and estimating these risks is an open and important problem. Still, Baum suggests that “threats are rarely completely unknown or unquantifiable,” [2] and while he notes that risk-based analyses are limited, resilience as a paradigm is useful for cases like this where the risk is underestimated, unknown, or unquantifiable.

Daniel Schmachtenberger suggests that there is a dichotomy of two generating processes that can lead to human-induced existential risk. [39] The first, where rivalrous and uncoordinated actors combined with “exponential technology” lead to either collapse, or anti-rivalrous solutions, parallels Bostrom’s Vulnerable World Hypothesis scenarios. The second, where complex systems become increasingly complex and fragile, is the case discussed in this paper. We will seek to expand on this, and more clearly describe why these systemic failures are particularly dangerous.

In order to discuss this, the paper first reframes Bostrom’s argument by replacing the analogy he uses with a more complex but also more complete explanation. This reframing is used to show how the systemic risk cases discussed in this paper are similarly worrying. Following this, the paper presents arguments for why systemic risks and fragility could make complex sociotechnical systems into fragile worlds, and then presents an additional sub-hypothesis about such systems which would entail Bostrom’s vulnerable world hypothesis, focused on path dependence, as will be explained. Finally, the paper presents a number of ways in which the expanded view of vulnerable worlds leads to a number of conclusions which are strikingly different from, or even exactly contrary to, the conclusions Bostrom suggested.

2 Novel Technologies Reframed as an Explore/Exploit tradeoff

Bostrom's recent "Vulnerable World Hypothesis" lists several ways in which "there is some level of technological development at which civilization almost certainly gets devastated by default, i.e. unless it has exited the 'semi-anarchic default condition.'" He illustrates this with an analogy modeled on a classic type of explanation in probability theory, that of an urn from which we draw balls. In his analogy, each ball represents a technological advance, and the balls can be white, black, or any shade of gray – representing the impact on humanity. Since we do not know what the future of technology will bring, the distribution of the colors of the balls inside the urn is unknown. He then discusses four scenarios which posit the existence of a "black ball" technology drawn from the urn, which, unlike those found so far with positive effects or mixed effects of humanity, "invariably or by default destroys the civilization that invents it."

Bostrom admits that his analogy to an urn is simplistic, though it is compelling. However, for our argument it is insufficient. Instead of an urn containing discrete technologies, the process of scientific discovery that occurs when certain advances pose risks can be understood with a different parallel from probability theory, one more often used in machine learning, that of the explore-exploit tradeoff. To explain this, we can consider the multi-armed bandit problem, where a gambler possesses some finite number of coins, and is faced with a row of slot machines, each of which has a potentially different payout distribution. For each coin, the gambler is faced with a choice of which machine to play. A conservative gambler might "explore" a bit and try a few machines, and notice a machine that pays out 2:1 just over half the time, and stay there to "exploit" the machine, slowly but steadily making money, uninterested in further exploration. A more adventurous gambler would instead spend more time exploring the other machines, perhaps seeking a machine that pays out less often, but with a far higher reward. If no such machine exists, they lose out from their continued exploration - but if there is a machine they find which pays out, say, 50:1 25% of the time, they will make far more money.

Explore-exploit tradeoffs are also discussed in somewhat different terms in machine learning, by analogy to a landscape, rather than an urn or a multi-armed bandit. Here, imagine a robot, which in our analogy will parallel human civilization, in a potentially dangerous environment. There are rewards, that is, new technologies, scattered about. At each stage, the robot can explore in any direction to search the landscape for a new reward (novel technologies), or focus on places where it knows rewards can be found. Because the payoffs of future exploration are unknown, optimal decision making in explore-exploit situations is notoriously hard to analyze. In fact, when the multi-armed bandit problem was first proposed during World War II, allied analysts suggested that perhaps the most useful application was to give the problem to the Axis, so that they would also waste their time on the seemingly intractable problem.[17]

Unlike the gambler, whose worst outcome is sub-optimal returns on their investment, in our scenario we suppose that the robot might accidentally become trapped or be destroyed by some previously unknown danger. If Bostrom's hypothesis is correct, civilization's exploration of new technologies comes with some probability of a fatal result, and some probability of discovering a highly rewarding new technology. In this new context, the goal is to find a search strategy that maximizes the total benefit humanity ever receives, what Bostrom has called humanity's cosmic endowment. [4] Any strategy, however, risks what Bostrom called astronomical waste - either embracing "technological relinquishment," by being so risk-averse that advanced technologies that greatly benefit humanity are never explored or developed, or being so risk-accepting that Bostrom's postulated Vulnerable World becomes inevitable, losing the entire potential future benefit to humanity. The first critical question addressed by Bostrom - "Is there a black ball in the urn of possible inventions?" is, to reframe the question, about the existence of fatal dangers in our Robot's exploration landscape. If we could answer that question in the negative, this would seem to refute the informal hypothesis he proposes of a vulnerable world. This is not the only refutation he suggests, however. Bostrom's suggestion of "differential technological development" is, in our terms, to cordon-off or minimize exploration of sections of the landscape or directions most likely to contain fatal traps. If this can be done, it would refute the suggestion that technological development inevitably leads to increased risk of choosing a devastating technology.

Falsifying Bostrom's initial hypothesis does not, however, show that the world is not vulnerable in other ways. The fuller statement of the hypothesis in Bostrom's paper is that "[i]f technological development continues then a set of capabilities will at some point be attained that make the devastation of civilization extremely likely, unless civilization sufficiently exits the semi-anarchic default condition.“ This statement, I will argue, cannot be falsified even by proving the impossibility of "a technology that invariably or by default destroys the civilization that invents it"[1]. In addition to Bostrom's four vulnerable worlds, I suggest a fifth possibility - that the simple accumulation of white and/or gray balls drawn from the urn can itself lead to fragility and, without strong forces pushing in the opposite direction, collapse of civilization by default. To frame it in our explore/exploit terms, the fifth possibility is that as the robot explores, it may dig too deep, or alter the landscape itself incautiously so that further movement could cause a collapse. The primary way for this to happen which this paper discusses is fragility - in the analogy, our exploration could lead to the equivalent of an avalanche or mine-collapse brought on by incautious behavior. That is, the creation of fragility in the landscape means that a lack of caution when exploring further could end up burying us under the weight of accumulated technologies.

3 Fragility and Systemic Risk

As technologies develop, they often build on one another, so that the continued operation of the system depends on a growing set of other systems. The way that failure occurs and propagates in such systems is non-obvious, but it is largely dependent on the topology of the interdependence between components. [32] Simple dependencies, where a system

[1] Such a proof of impossibility itself seems implausible, but I grant the strongest possible case in order to show its insufficiency.

requires the functioning of every component, can make the resulting system of systems more fragile than the components. A very basic model of this shows that given a system-of-systems with N components, each of which independently can fail at the rate F_i the failure rate is $1 - (\prod_{i=1}^{N} 1 - F_i)$. While this grows more slowly than the sum of the individual failure rates as new systems are added, it is also far higher than the average failure rate of those individual systems.

As a concrete example, the peak of efficient farming once required family farms to depend on a family to manage the farm, a blacksmith to make horseshoes and plows, and draft animals to pull them. Losing any one of these would be enough to (eventually) make the system unable to continue, and there was some risk that this would occur. Still, the limited number of inputs and the substitutability of other inputs made such systems fairly stable - especially because many risks were uncorrelated across farms, and could often be addressed by borrowing from other farms nearby.

The farms of today, of course, are not nearly as simple. They require everything from satellite GPS systems to pinpoint locations, to the semiconductors and fabrication plants used to make specific integrated chips used in the farm equipment, to internet connectivity to run machine learning algorithms using collected data and satellite imagery on remote servers. [50] Modern agribusiness depends on everything from finding and hiring skilled laborers to manage complex machinery, to managing regulatory, financial, and other factors critical to farm operation. [25] These are often more tightly correlated across an economy, increasing risk. Beyond that, managing these farms requires understanding "human, technical, economic, financial, risk, institutional and social" issues, as Lewis et al. noted more than a decade ago. [28]

The risk is likely not yet critical, but it seems clear that dependence is growing, and the ability to use backup systems can be lost. For example, if remote servers become unavailable, local corn farms may lack the information needed to decide where to increase and decrease watering levels, or even lose the ability to run their computer-controlled irrigation systems. Similarly, decades ago supplies and ordering were managed on paper, and now, without the servers running Software-as-a-Service supply-chain-management software, the dairy farm down the road may not have access to an inventory of their supplies or know what amounts of products are needed or what they have historically ordered, and end up unable to feed their cows.

The inter-dependencies in such systems are more complex than the above model allows, but more complex analyses, such as those employing percolation analysis to understand mutual interdependency of multiple networks, [7] show the same trend. That is, interdependent systems where failures can propagate can be far more likely to fail than the average rate at which the individual systems' components fail. Worse, analysis of "high-value, technology- and engineering-intensive products or systems... used to produce consumer goods and services" has shown that the failure rates are nearly-additive, and worse, are hard to identify. [49]

It should be noted that modern computer networks do not display such fragility, but this is a function of intentional design. [29] Simple network structures, such as lines or rings, are far more prone to failure [11]. That is, unless a system was designed for resilience, resilience should not be expected. And when technological systems are made efficient and complex, they tend to be tightly coupled - meaning that failure in one place spreads [3].

3.1 Inevitable Technological Fragility Hypothesis

The proposal of this paper, to provide an addendum to Bostrom's hypothesis, is that if technological development continues indefinitely, systemic fragility will increase to the point that the possibility of a shock sufficient for complete collapse approaches certainty.

This hypothesis rests on a number of assumptions, but there are also a variety of reasons to find it plausible and concerning. To lay these out clearly, we will first consider the question of how and why individual systems are fragile, then make an argument that it is at least plausible that the multiple interconnected systems and systems-of-systems which are necessary for much of modern civilization not only fail to address this risk, but multiply it.

4 Single-System Complexity and Fragility

The key question so far is whether fragility increases over time as systems are built. The answer to that question depends on a combination of factors that can push in either direction. These include increasing complexity of systems, the economic incentives for efficiency over robustness and the resulting levels of investment in resilience, the failure rates of individual components and systems, as well as the way in which systems-of-systems (and systems-of-systems-of-systems) are interrelated, and the extent to which systems and their interdependencies are designed to be robust.

Even the claim of inevitable fragility in individual systems makes several assumptions about how fragility increases. Before looking at the systemic question of how fragility could lead to collapse, we will outline these assumptions. Note that these are in fact assumptions, rather than claims - if any one of them is false in ways that are outlined, it would refute the hypothesis. The third assumption is particularly critical, and will be explored further in the next subsection.

First, for fragility to matter, the current trend of efficiency-increasing and resilience-decreasing technologies must continue to apply to at least one critical system, such as agriculture, communication, or transport. If this is wrong, and future white-ball / safe exploration technologies are ones that favor robustness over efficiency in all such critical domains, the trend would reverse. For instance, distributed fault tolerant computing arguably increases both efficiency and robustness. Most new technologies move in the opposite direction, but if enough resilience increasing technology is found, the balance could shift.

Second, the argument for increasing fragility assumes that economic growth continues to absorb human effort in a way that does not lead to overabundant resources. In Eric Drexler's 'Paretotopia' scenario, increased resources are unmatched by increased demand. [14] In that future case, resources are abundant enough that robustness is easy to achieve. This second scenario also assumes the absence of supercharged competition that uses the newly abundant resources. This would not occur, for example, in Hanson's proposed default "Em" scenario, where human-based intelligences are simulated computationally, leading to a reduction rather than an increase in surplus that could be redirected to robustness for lack of other needs. [19]

Third, it assumes that fragility is relatively hard to identify, such that at least some failures will be unanticipated. This has been true historically, but it is possible that future developments would reverse this trend, making the search for increased robustness itself efficient enough to counterbalance the more general and destabilizing increased fragility that new technology allows. If failures do become easy to anticipate, more expensive general resilience can be replaced with more specific redundancies targeted to the exact failure modes identified.

4.1 Non-Obvious Fragility

As mentioned, hard-to-identify fragility is a key assumption. Broadly speaking, non-obvious fragility is the result of planning for efficiency, instead of designing for redundancy, fault tolerance, or even provable safety. This is a fairly general fact about any control system. [30] The concrete result of the current optimization shows clear signs of producing fragile results. One example is the proliferation of disposable technology, such as fragile smartphones designed to be replaced rather than fixed or upgraded. Failure of these optimized devices is normal, and while mitigating failure is important, it is often the case that risk must be accepted, rather than avoided. [33] This type of fragility is obvious and anticipated, rather than non-obvious and worrying. For example, individual computers are fragile, and components fail frequently. For this reason, in high-reliability computer systems, a variety of mechanisms are in place to compensate, including redundant online systems for data storage, [9] or methods to address other hardware failures. [46]

The fact that computer networks are not fragile, and the fragility that does exist is well understood, seems to be a counterexample. But the resilience itself is planned, in contrast to ecological systems where it is emergent - as we will discuss in detail below. This means that fault-tolerant designs are built to be tolerant of expected faults. Not only that, but resilience itself is optimized, for example, to minimize the number of backups or other costs needed to have a planned level of reliability. [36] This creates fragility to unexpected faults, and allows the systems to operate through anticipated contingencies, but not to anything beyond that point.

4.2 Sociotechnical Resilience

Fragility of systems is not based purely on the lack of resilience of technical systems. In fact, fragility of optimized technical systems is compensated for by the greater robustness of sociological systems. The combined sociotechnical system, then, is the level at which fragility should be considered.

To reduce the fragility of sociotechnical systems, organizations can attempt to build more resilience at the organizational level. This can involve information sharing, distributed decision making, and better risk assessment. If done well, these attempts provide a sociotechnical system that compensates for technical and operation risk, but is again very different from emergent resilience. [27] Unfortunately, the interaction between humans and technology can often multiply risks, rather than mitigate them. [49]

Another reason to think that sociotechnical resilience will not fully compensate for technological fragility is the reduced human involvement in technical systems. As automation increases, Danzig notes that humans are increasingly necessarily out-of-the-loop. [12] He further argues that when there is competition, this dynamic is a necessary result of continued optimization.

To conclude the discussion of single-system fragility, we note that inevitable fragility of systems is not actually required for the hypothesis presented. As this section argued, it does seem plausible that in expectation, new technologies will be more fragile than those they replace. However, systemic risk can exist given the much weaker claim that specific critical systems are relied upon, and technological improvements relevant to those systems alone exhibit sufficient fragility to cause a cascading collapse. Before discussing the interaction between systems, however, it is worth considering how these human, technical, or sociotechnical systems differ from naturally resilient biological systems.

4.3 Contrasting (fragile) sociotechnical systems with (resilient) biological systems

There is an extensive literature on resilience and complexity in biological systems that contrasts greatly with the arguments in this paper. The analogy between complex systems in biology and complex technological systems is apt in many ways. For example, free markets are often understood as evolutionary, with businesses engaged in Darwinian competition. We will explain, however, that this provides reason to think that the comparison reinforces some of the arguments, rather than solely contradicting them.

Resilience was arguably first discussed in ecology by Holling. [20] In his paper, it is clear that resilience itself is an emergent property of the complex ecological system. Notably, however, the premise of the paper is that at the time, ecology and related disciplines were analyzed along similar lines to technical systems. He suggests that this parallel is unfortunate, because ecological systems fundamentally differ.

Engineered systems are complex in ways similar to biological systems. However, even small differences lead to very different higher level properties, which is a general fact about complex or emergent systems. One such difference which is critical relates to the diverse redundancy of ecological systems, where many different species can fill each niche. Technology usually relies on a single system, with a backup that is nearly if not completely identical. Holling made the argument for the difference between the two classes of system clearly, calling for people not to confuse resilience as discussed in ecology and resiliency as discussed in engineering.[21]

To expand on this point, Holling has more recently argued, in Peterson et al., that resilience is a function of ecosystem diversity, with overlapping functions and redundancies across scales. [34] In contrast, engineered systems typically reduce redundancy to increase efficiency. For example, a key software design principle is that anything which performs the same task should be "refactored" to use a single function rather than having redundancy. [24]. This will increase maintainability, and likely lead to fewer bugs, but will also increase correlation of failures and vulnerabilities. While there are design patterns that take the opposite tack, [8] they are not in common use.

Another key difference is the type of redundancy, as mentioned above. While having multiple organisms to fill a niche, or even individuals of the same species can increase resilience due to redundancy, having multiple copies of software in a system typically has no benefits, [22] and if anything the necessity to replace or patch software in multiple places makes maintaining a system, or recovery after a failure, even harder.

It could be contended that the layer above that of sociotechnical systems is where resilience will emerge. Markets create competition that can provide another layer of redundancy, and it could be argued that they create efficiency in ways similar to evolutionary competition. Unfortunately, there are clear arguments that markets may not evolve towards greater resilience. [43] In addition to theoretical arguments, there are empirical facts that have emerged to support the claims. Global markets are consolidating rather than becoming more diverse, and firms are often collaborating (if not colluding), rather than competing. As banks showed in 2007, large firms with complex businesses can create systemic risks of interdependence, rather than robustness from diversification. [26] Recent research shows that this is true not only in finance, but more broadly. [48] This implies that systemic risks are in fact critical.

5 Larger Scale Risks from Fragility

Failures happen, and as outlined above, we can see at least one plausible future where failures due to fragility cannot be eliminated. The idea that they could lead to global catastrophes or existential risks is perhaps more surprising, and requires a stronger claim than inevitably increasing fragility of individual systems outlined earlier. For that reason, we will now lay out some preconditions for increasing systemic fragility that would imply that larger scale vulnerabilities are also inevitable.

- First, larger scale risks leading to collapse requires that technological advances enable or require complex systems that cannot be replaced with simpler alternatives, either because of path-dependent technological dependencies, or lack of availability of such alternatives.

- Second, collapse risk requires that failures are rare enough that technological growth allows building systems far more complex than are maintainable or replaceable before failure occurs.

- Third, collapse risk requires that failures either exhibit a domino effect, where systems are reliant on one another in ways that causes a chain reaction of failure, or that failures are correlated due to all being reliant on a single point of failure.

Despite the above required confluence of uncertain claims, it seems that each of the claims is not only plausible, but at least somewhat supported by evidence. To explain how, it is worth looking at each, and considering how the incentives in the system work, and past experience with how such systems develop. Finally, we will note some ways in which these factors interrelated and increase risk further via incentives for suboptimal mitigation, and the increased risk of moderate investments in resilience.

5.1 Path Dependence and Irreversibility: There's No Way Out But Forward

Path dependence leading to irreversible complexity is another oft-seen dynamic in complex systems[2]. As the earlier example in agriculture noted, most real-world systems have a limited ability to roll-back to earlier states. Older equipment is discarded, and prior techniques are lost when older workers leave.

Even for software, where neither experience or equipment is typically needed to go back to earlier systems, there are limitations in rolling back a single protocol or system once other related systems depend on them[3]. Such software obsolescence is a well-understood problem. [37] As a concrete example of how this occurs, software programs cannot always be run on newer versions of operating systems. If you want to run an older program, often the only solution is to run an emulation, i.e. depend on the newer system. Similarly, older operating systems cannot typically run on newer hardware, since they do not have appropriate device drivers for new hardware. And older hardware cannot easily be manufactured, as production lines were long ago scrapped.

5.2 Momentum and Cumulative Complexity

Technology evolves rapidly, and failures are not often seen. Despite this, problems often persist for years, and even critical software security issues remain undiscovered and can last through many versions and for many years. [1] Not only could a cause of failure be present in older versions, but it may last to the point where software without the problem is obsolete. This implies that a system might not fail a single step past where it is safe, where rollback to less fragile technology is only difficult, but that it instead ends up 4 or 5 steps past that point, so that replacing systems with earlier versions in time to prevent collapse is completely non-viable.

A related issue is that complexity and failures do not always result from a single system. Often, it is the interplay between systems that creates issues. The entire discipline of systems engineering exists in large part to ensure different systems can work

[2] Randall Munroe humorously describes an exaggerated case in which an attempt to enable a computer to dual-boot a second operating system leads first to difficulty restoring the status quo, then an attempt to salvage a second system that was modified in an attempt to fix the first, and ending with the two characters stranded in the ocean hoping to stay alive. `https://www.xkcd.com/349/` While this scenario is exaggerated, the plausible result of cascading failures in large enough systems is much worse.

[3] Sarna and Drexler note that progress in provable safety and formal methods for software development is a critical trend that can reduce security vulnerabilities[38]. It is possible these or related methods could be used to provide guarantees on limited interdependencies between systems, providing a counterbalance to this risk.

together, and it is rarely a trivial problem. The problems of integrating pairs of systems is only made harder for systems-of-systems with dozens or hundreds of component systems.

When these systems of systems are designed in concert with one another, the problems are managed. More often, the parts were designed fully independently, and engineers build processes that depend on the systems working together. Of course, these systems-of-systems are themselves often then connected. Tonn et al. discuss "Earth Scale convergence systems" [44] and note that maintaining such earth scale systems is "a constant and seemingly growing challenge." This is not coincidental, or only due to a lack of planning. Instead, the size itself creates challenges.

5.3 Growing complexity and interconnection

A single system-of-systems failure might not lead to collapse, even if it did pose a global catastrophic risk. However, the benefits of efficiency means that firms and individuals will also seek to subcontract or outsource to specialized firms despite the risks of dependency, and create additional interconnections and dependencies. This is already seen as a large source of risk in global supply chains, where companies are exposed to risks by suppliers of suppliers, and those companies' suppliers as well.

An example of this type of fragility which surprised the automobile industry was the fact that despite attempts to mitigate their supply chain risks for components, they discovered that all of the various suppliers of metallic color paints for their cars actually ordered their pigments from a single factory which was shut down by the 2011 Japan Tsunami. [41] While failures of this sort are uncommon, further research found that the dependencies which allow them are typical. Research looking at consumer goods, health care, and manufacturing firms found that between 12% and 13% of the firms' first-degree suppliers depended on the same second-degree suppliers. [47] Similar analyses in other complex domains like software package dependencies show similar shared reliance. [13]

The growing interconnection of the world, and the growing dependence on technologies, is likely only going to increase as more and more new technologies that improve human lives and economic success become available. The fragility of these systems, of course, will also grow as systems grow ever more efficient, and interconnected. If such systems are fragile, however, people will be motivated to reduce risks. As we will argue next, this is insufficient.

5.4 Economic Incentives for Failure

A penultimate point is that systems will not only continue to have some risk of failure, but will also systemically underinvest in reliability. A straightforward microeconomic argument provides a reason to expect this. Experience with safe engineering has led to a clear principle in engineering that for any given project, there is a cost/safety tradeoff. A system can always be made safer, but past a certain point, it is not reasonable to pay for marginal improvements in safety. As argued above, this tradeoff by default leads to not only unsafe individual firms, but also to increasing systemic fragility. This implies that systemic resilience is a public good, in the economic sense.

A sufficient level of investment in this public good is unlikely, as is typical for public goods. This parallels Bostrom's example of climate change, where individual actors each contribute to the problem. [6] To see why, we note that resilience and robustness in self-governed systems is limited by a local/global risk tolerance mismatch. Firms will obviously invest in risk mitigation, but the extent to which they do so is limited by their economic benefit.

This is lower than would be hoped for in general because the worst outcomes possible for an individual firm or individual, of failing completely, imposes additional negative externalities on the wider system. That is, because a firm or individual has downside limited to its own existence, the amount of mitigation that they find to be valuable in expectation will be lower than the level which is socially optimal. This points to the economic "bad" of complexity and fragility - i.e. an economic good which is harmful rather than beneficial, parallel to pollution in Bostrom's example of climate change. For this reason, unless we see that firms and individuals are more risk-averse about systemic risks than their local preferences imply are rational, we expect that local decisions towards sufficient robustness will undervalue robustness, and instead be more complex and fragile than ideal.

5.5 Failure Correlation

A number of arguments have now been put forward as to why systems would fail, and these failures would likely be large and pose at least global catastrophic risks. However, the risk of any one system-of-systems failure at a given time is high enough that cumulative complexity and momentum is somewhat unlikely, as long as these failures are uncorrelated, and the confluence of several such failures is then vanishingly unlikely. For that reason, the hypothesis of inevitable *complete* collapse depends on growing correlation of failures, or an increased risk of larger failures.

Correlated failures can result from having a single point of failure across systems. As noted above, outsourcing is more efficient than distributing tasks, and firms will attempt to find ways to outsource to the lowest cost, most efficient provider or system. Because of economies of scale, it will sometimes be the case that a natural monopoly will form for certain services or products, and when this occurs, any failures will occur across the economy rather than locally. This is not only speculative. Failures in key services already have ripple effects across many industries, as in the case of Salesforce, a widely used customer contact system, which led to a brief shutdown of sales and related activities across industries when it went offline for even a short time. [10]

A similar phenomenon occurs in computing, as noted in analyses of cybersecurity risk. [15] Operating systems for servers, for example, have a market dominated by Linux. Desktop computer operating systems are similarly dominated by Microsoft Windows. There are a variety of reasons that such near-monopolies exist, but they are found often. In the same way, many key services for computer systems are complex, and only a few implementations exist. Bugs or failure points that are found in such key services are therefore not always confined to a single system, but are actually universal, by virtue of the universal usage of a single point of failure.

Even where no monopoly exists, convergent needs and the use of related techniques across an industry can lead to convergent failure modes, as was found in the case of Spectre and related security vulnerabilities for computer processors. In that case, all processors which used the general technique of branch prediction for speculative execution were vulnerable to a set of very closely related flaws. [23] The shared failure mode was not due to a single system being relied on, but rather to the fact that a single method was used across the industry.

A final argument can be made for collapse risk rather than simply danger from individual failures that relates the systematic underinvestment to the risk of overly complex systems. In short, rare failures can create additional fragility. The argument is that if systemic fragility is identified due to failures, the larger risks will be apparent. Less frequent failures, as argued above, will lead to growth in complexity and greater fragility. The fact that firms are motivated to invest moderate amounts in reducing fragility is likely to make failures less frequent, but not eliminate the risk - so that when failures occur, they will be failures of more complex and more closely inter-related systems.

6 Conclusion and recommendations

The paper argues that there is a significant and growing risk of global catastrophe due to technological complexity, and the resulting fragility of systems. Individual actors (at the company, state, or regional level) may benefit from technological races that promote economic growth over systemic safety and robustness, but the growing interdependence of international systems makes this risky. This implies that continuing the current trend of investment based primarily on the promised advantages of new technologies is a significant concern. The paper then presents a hypothesis that this is an inevitable result of a certain type of technological innovation. If the hypothesis is true, it would mean that continued technological innovation leads to what Bostrom refers to as a "vulnerable world," one that inevitably leads to catastrophe.

As with Bostrom's other vulnerable world scenarios, the risks discussed here are plausibly greatly mitigated by restricting technological development, and effective global governance. Unlike the scenarios he presents, however, this risk is not reduced by minimizing the variability of goals and motives of those looking for new and dangerous technology, nor via effective preventative policing. Instead, the existence of fragility risk argues strongly for a different type of risk-aware research prioritization. Specifically, research should be prioritized more thoughtfully with explicit investment in technologies that promote resilience. For the majority of research, investigating non-robust technologies, there should be more consideration of the potential for failure, and the systemic implications of each technology.

The existence of technological fragility risks does not, however, contradict the hypothesis behind Bostrom's fragile world scenarios, and as noted, can be fully compatible. It is not only plausible but near-certain that there are multiple failures possible that would prevent humanity from claiming their cosmic endowment. A key question is how to investigate the relative importance, likelihood, and tractability of the different failure modes. While this paper proposes no answer to that question, it seems reasonable that it is worthwhile to promote recognition of the risk, and to pursue low-cost mitigation, including the simple expedient of attempting to identify and reduce systemic fragility where it exists.

Thanks are due to Gregory Lewis for encouraging my work on this, to Anders Sandberg and Eric Drexler for early feedback and encouragement, and to Carlos Parada, Eli G., and an anonymous reviewer for very helpful feedback on the manuscript.

References

[1] Lillian Ablon and Andy Bogart. *Zero days, thousands of nights: The life and times of zero-day vulnerabilities and their exploits.* Rand Corporation, 2017.

[2] Seth D Baum. Risk and resilience for unknown, unquantifiable, systemic, and unlikely/catastrophic threats. *Environment Systems and Decisions*, 35(2):229–236, 2015.

[3] Richard Bookstaber. *A demon of our own design: Markets, hedge funds, and the perils of financial innovation.* John Wiley & Sons, 2007.

[4] Nick Bostrom. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3):308–314, 2003.

[5] Nick Bostrom. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31, 2013.

[6] Nick Bostrom. The Vulnerable World Hypothesis. *Global Policy*, 10(4):455–476, 2019.

[7] Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.

[8] Liming Chen and A Avizienis. N-VERSION PROGRAMMINC: A FAULT-TOLERANCE APPROACH TO RELlABlLlTY OF SOFTWARE OPERATlON. In *Twenty-Fifth International Symposium on Fault-Tolerant Computing, 1995, ' Highlights from Twenty-Five Years'.*, page 113, 1995.

[9] Peter M Chen, Edward K Lee, Garth A Gibson, Randy H Katz, and David A Patterson. RAID: High-performance, reliable secondary storage. *ACM Computing Surveys (CSUR)*, 26(2):145–185, 1994.

[10] Thomas Claburn. Salesforce? Salesfarce: Cloud giant in multi-hour meltdown after database blunder grants users access to all data, 2019.

[11] David D Clark, Kenneth T Pogran, and David P Reed. An introduction to local area networks. *Proceedings of the IEEE*, 66(11):1497–1517, 1978.

[12] Richard Danzig. Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority. Technical report, Center for a New American Security, 2018.

[13] A Decan, T Mens, and M Claes. An empirical comparison of dependency issues in OSS packaging ecosystems. In *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 2–12, 2017.

[14] Eric Drexler. Pareto-topia, 2017.

[15] Dan Geer, Eric Jardine, and Eireann Leverett. On market concentration and cybersecurity risk. *Journal of Cyber Policy*, pages 1–21, feb 2020.

[16] Edward Geist and Andrew J Lohn. How Might Artificial Intelligence Affect the Risk of Nuclear War? 2018.

[17] John C Gittins. Bandit processes and dynamic allocation indices. *Statistics*, 41(2):148–177, 1979.

[18] Gigi Kwik Gronvall. *Synthetic biology: Safety, security, and promise*. Health Security Press, 2016.

[19] Robin Hanson. *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press, 2016.

[20] Crawford S Holling. Resilience and stability of ecological systems. *Annual review of ecology and systematics*, 4(1):1–23, 1973.

[21] Crawford Stanley Holling. Engineering resilience versus ecological resilience. *Engineering within ecological constraints*, 31(1996):32, 1996.

[22] Java T Point. Software Failure Mechanisms, 2018.

[23] Andrew Johnson and Ross Davies. Speculative Execution Attack Methodologies (SEAM): An overview and component modelling of Spectre, Meltdown and Foreshadow attack methods. In *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6. IEEE, 2019.

[24] Joshua Kerievsky. *Refactoring to patterns*. Pearson Education, 2004.

[25] Ross Kingwell. Managing complexity in modern farming. *Australian Journal of Agricultural and Resource Economics*, 55(1):12–34, 2011.

[26] Mr Luc Laeven, Lev Ratnovski, and Hui Tong. *Bank size and systemic risk*. Number 14. International Monetary Fund, 2014.

[27] Krista S Langeland, David Manheim, Gary Mcleod, and George Nacouzi. How Civil Institutions Build Resilience Book: Organizational Practices Derived from Academic Literature and Case Studies. Technical report, 2016.

[28] Paul Lewis, Bill Malcolm, and Graham Steed. Conservation Crop Farming: A Farm Management Perspective. Technical report, Australian Agribusiness Perspectives, 2006.

[29] Robert M Metcalfe and David R Boggs. Ethernet: Distributed packet switching for local computer networks. *Communications of the ACM*, 19(7):395–404, 1976.

[30] J Paattilammi and P M Makila. Fragility and robustness: a case study on paper machine headbox control. *IEEE Control Systems Magazine*, 20(1):13–22, 2000.

[31] Derek Parfit. *Reasons and persons*. OUP Oxford, 1984.

[32] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, aug 2015.

[33] Charles Perrow. *Normal accidents: Living with high risk technologies-Updated edition*. Princeton university press, 2011.

[34] Garry Peterson, Craig R Allen, and Crawford Stanley Holling. Ecological resilience, biodiversity, and scale. *Ecosystems*, 1(1):6–18, 1998.

[35] Federico Pistono and Roman V Yampolskiy. Unethical research: how to create a malevolent artificial intelligence. *arXiv preprint arXiv:1605.02817*, 2016.

[36] Luiz Henrique Rodrigues-da Silva and José António Crispim. The Project Risk Management Process, a Preliminary Study. *Procedia Technology*, 16:943–949, 2014.

[37] Peter Sandborn. Software obsolescence-Complicating the part and technology obsolescence management problem. *IEEE Transactions on Components and Packaging Technologies*, 30(4):886–888, 2007.

[38] Gopal P. Sarma and Eric Drexler. Formal methods can provide secure software foundations and support AI safety. 2018.

[39] Daniel Schmachtenberger and Daniel Thorson. Designing a non-self-terminating civilization, 2019.

[40] Stefan Schubert, Lucius Caviola, and Nadira S Faber. The Psychology of Existential Risk: Moral Judgments about Human Extinction. *Scientific Reports*, 9(1):1–8, 2019.

[41] Yossi Sheffi and Barry C Lynn. Systemic Supply Chain Risk. *The Bridge (Fall), 22*, 29, 2014.

[42] Andrew E Snyder-Beattie, Toby Ord, and Michael B Bonsall. An upper bound for the background rate of human extinction. *Scientific reports*, 9(1):11054, 2019.

[43] Clem Tisdell. Diversity and economic evolution: failures of competitive economic systems. *Contemporary Economic Policy*, 17(2):156–165, 1999.

[44] Bruce Tonn, Mamadou Diallo, Nora Savage, Norman Scott, Pedro Alvarez, Alexander MacDonald, David Feldman, Chuck Liarakos, and Michael Hochella. Convergence platforms: earth-scale systems. In *Convergence of Knowledge, Technology and Society*, pages 95–137. Springer, 2013.

[45] Bruce Tonn and Dorian Stiefel. Evaluating methods for estimating existential risks. *Risk Analysis*, 33(10):1772–1787, 2013.

[46] Guosai Wang, Lifei Zhang, and Wei Xu. What can we learn from four years of data center hardware failures? In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 25–36. IEEE, 2017.

[47] Yixin Iris Wang, Jun Li, and Ravi Anupindi. Risky suppliers or risky supply chains? An empirical analysis of sub-tier supply network structure on firm performance in the high-tech sector. 2015.

[48] Jonathan William Welburn, Aaron Strong, Florentine Eloundou Nekoul, Justin Grana, Krystyna Marcinek, Osonde A Osoba, Nirabh Koirala, and Claude Messan Setodji. Systemic Risk in the Broad Economy: Interfirm Networks and Shocks in the US Economy. 2020.

[49] K T Yeo and Yingtao Ren. Risk management capability maturity model for complex product systems (CoPS) projects. *Systems Engineering*, 12(4):275–294, dec 2009.

[50] Yury Zubarev, Denis Fomin, and Nikolai Zubarev. Using high-precision farming systems in the agricultural sector-the path to digital agriculture. In *International Scientific and Practical Conference "Digital agriculture-development strategy"(ISPC 2019)*. Atlantis Press, 2019.