Systemic Fragility as a Vulnerable World

David Manheim

December 16, 2018

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 a new hypothesis that complexity of a certain type can itself function as a source of risk. This "Fragile World Hypothesis" is compared to Bostroms "Vulnerable World Hypothesis", and the assumptions and potential mitigations are contrasted

1 Introduction

Existential risks take several forms, as discussed by Bostrom, and their prevention is a moral priority[Bos13]. These risks go beyond widespread devastation, to the point that they lead to not only astronomical waste[Bos03], but potentially to forfeiting humanity's potential cosmic endowment[Par84]. Some risks, include Artificial Superintelligence and Biological risks are at least being investigated by an active research community. Baum suggests that "threats are rarely completely unknown or unquantifiable," [Bau15] but notes that riskbased analyses are limited, and resilience as a paradigm is useful for cases where the risk is underestimated, unknown, or unquantifiable. A recent paper by Bostrom discusses and classifies black-swan technological advances that are by their nature existential risks, onces that would by default lead to extinction[Bos18]. The class of existential risks this paper addresses is related to several of the previous modes, but extends Bostrom's list of Vulnerable Worlds.

Systemic risks are those that emerge from complex system failure, where the failure of a single component leads to systemic knock-on effects. Because of the way these risks emerge, they seem more difficult to classify or estimate than typical risks[1]. Daniel Schmachtenberger suggests that there is a dichotomy of two generating processes that can

[1]Tonn and Steifel suggest methods for estimating existential risks which are primarily suitable for risks that are not systemic or unknown. [TS13] 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. In Kuhlemann's categoriation of GCRs [Kuh18], the class of GCR discussed here is epistemically messy and involves a collective action problem, but is a sudden and technologically driven risk, placing it somewhere in the middle of the sexy/unsexy risk dichotomy.

lead to human-induced existential risk. The first, where rivalrous and uncoordinated actors combined with "exponential technology" leading to either collapse, or anti-rivalrous solutions, parallels Bostrom's Fragile World Hypothesis scenarios. The second, where complex systems become increasingly complex and fragile, parallels the cases discussed in this paper. We will seek to expand on this, and more clearly describe why these systemic failures are particularly dangerous.

2 Fragile worlds reframed as Explore/Exploit tradeoffs

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." His four scenarios all posit the existence of a "black ball" technology drawn from the urn of potential advances which is devastating. These advances, unlike the ones that have positive effects, "invariably or by default destroys the civilization that invents it."

The push to discover new technologies despite risk can be understood as an explore/exploit tradeoff in a potentially dangerous environment. At each stage, the explore action searches the landscape for new technologies, with some probability of a fatal result, and some probability of discovering a highly rewarding new option. The implicit goal in a broad sense is to find a search strategy that maximize humanity's cosmic endownment - neither so risk-averse that advanced technologies are never explored or developed[2], nor so risk-accepting that Bostrom's postulated Vulnerable World becomes inevitable. Either of these risks astronomical waste. However, until and unless the distribution of black balls in Bostrom's technological urn is understood, we cannot specify an optimal strategy. 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 negative singularities in the fitness landscape. If we could answer that question in the negative, it would seem to refute the informal hypothesis he proposes of a vulnerable world, butBostrom discusses other potential refutations, which can also be discussed in these terms. 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, this too would refute the suggestion that technological development inevitably leads to increased risk of choosing a devastating technology.

The ways proposed for falsifying tinitial hypothesis does not, however, show that the world is not vulnerable in other ways. The fuller statement of the hypothesis is that "[i]f technological dhe evelopment 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

[2]Bostrom suggests that "technological reliquishment" is possible, which corresponds to reverting to a very cautious strategy in the explore/exploit terminology.

or by default destroys the civilization that invents it"[3]. 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 explore/exploit terms, the says that if the total footprint of the exploration becomes large enough, collapse may occur.

3 Fragility and Existential Systemic Risk

As technologies develop, they tend to build on one another, leading to increased fragility. The peak of efficient farming once required a mine to extract iron, a blacksmith to make plows, and a supply of draft animals. Modern, more efficient farming now depends on everything from futures markets to hedge risks in order to get loans, to 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-server farms. These dependencies can become complex, and hard to identify. For example, if remote server farms become unavailable, local corn farms likely lack the information needed to decide where to increase and decrease watering levels, and without the server farm running Software-as-a-Service supply-chain-management software, the daily farm down the road may end up lacking supplies to feed their cows.

Existential risk from fragile systems requires a stronger claim than inevitably increasing fragility. First, it depends on fragility increasing it in a way that is non-obvious or unaddressed until the fragility is irreversible. Second, the failure mode requires that technological advance leads to systems fragile in ways that replacing too-complex systems with somewhat simpler alternatives isn't viable, because of path-dependent technological dependencies. Third, it requires that failures are rare enough that technological growth allows building several layers past where there is a margin of safety before failure occurs. Because of the dependency, it is possible that technological interdependence does not fail a single step past where it is safe, where rollback to less fragile technology is only incredibly hard, 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.

Despite the required confluence of uncertain claims, it seems that each of the claims is reasonable. To explain, it is worth looking at each, and consider how incentives and experience seem to support them. First, it seems plausible that in expectation, each additional technology developed and deployed increases systemic fragility. This is not required, however. Instead, this form of systemic risk can exist given the much weaker claim that there are specific critical systems in which technological improvements exhibit sufficient fragility to cause a cascading collapse.

Non-obvious fragility is the result of planning for efficiency, instead of designing for provable safety. The observable expense of safe engineering has led to a clear principle in

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

engineering that for any given project, there is a cost/safety tradeoff, which by default leads to increasing systemic fragility. Increased fragility occurs in part because firm or individual has downside limited to its own existence. The worst outcomes possible for an individual firm or individual, however, almost certainly imposes externalities on the wider system. This strongly implies that resilience and robustness in self-governed systems is limited by the local/global risk tolerance mismatch. For this reason, unless we see that firms and individuals are near-universally more risk-averse about systemic risks than their local preferences, we expect that local decisions towards sufficient robustness will under-provide the public good of stability that robustness provides.

Path dependence leading to irreversible complexity is another oft-seen dynamic in complex systems[4]. Tonn et al. discuss "Earth Scale convergence systems" [TDS+13] and note that maintaining them is "a constant and seemingly growing challenge." This is not coincidental, and the issues that arise with maintaining human-structured systems are different than the often self-stabilizing natural complex systems[May01].

As the earlier example in agriculture noted, most designed systems have a limited ability to roll-back to earlier states. Even for software, where such features are relatively straightforward, there are limitations in rolling back a single protocol or system once other related systems depend on them[5]. As a notional example, it is plausible that networking protocols could be specified inadvertently in a way that precludes rollback or restoration[6]. We could imagine that the current version of a transmission protocol relies on a routing protocol, and the network for the routing protocol must be set up using that transmission protocol. Because the initial setup of the system may have preceded the dependency, if the system ever needed to be restarted, re-creating the system may be impossible without returning to earlier versions of the system. If there is not copy of the earlier version present at each site, the rollback itself may be impossible without re-sending the earlier version - which requires a working routing and transmission system. Obviously in this case it would be possible to physically mail the older version, as long as everyone involved knew what was going on. That means we must hope that the phone and physical mail systems that would be used to coordinate and deliver goods are not themselves dependent on the availability of the network that must be set up.

It could be argued that small failures will prompt improved resilience in designed systems. Because firms have incentives to reduce systemic risk to some extent, many such small failures will be identified and fixed even before such failures occur. Counter-intuitively, the resulting rarity of failures is itself a vulnerability, rather than a positive.

[4]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.

[5]Sarna and Drexler note that progress in provable safety and formal methods for software development is a critical trend that can reduce security vulnerabilities[SD18]. It is possible these or related methods could be used to provide guarantees on limited interdependencies between systems, providing a counterbalance to this risk.

[6]While the example is notional, it is not obvious to the author that the description does not describe the current international communications network.

This is because failures prompt re-assessments and identification of vulnerabilities, as well as prompting funding and promoting attention to fixing them. If small failures are rare, the system is able to become increasingly complex and prone to ever-larger cascading failures without the associated large-scale evaluation, identification of systemic vulnerabilities, and work to avoid them.

This "Complexity Overhang" problem alone, however, is insufficient to create existential risks. Even given a near-complete collapse of food supplies, it seems implausible that some sufficiently self-sufficient group, with enough genetic diversity to maintain a vialable future society, would not survive. Unfortunately, collapse from such fragility is likely to be gradual-enough to lead to shortages. The more interdependent and connected the world becomes, the less chance that such groups will be able to escape undetected and unmolested. Unlike the current semi-stable international order, there would be little disincentive not to take huge risks in the hopes of capturing enough supplies for small surviving groups, and the negative-sum competition could lead to further collapse.

3.1 Inevitable Technological Fragility Hypothesis and Conditions

The proposal, to provide a modification of Bostrom's hypothesis, is that "if technological development continues indefinitely, systemic fragility will increase to the point that ever-smaller shocks which can cause partial failures are increasingly likely to cause complete collapse, and the probability of the occurrance of such shocks approaches certainty."

The fragile-vulnerable-world scenario outlined here makes several assumptions. First, that the current trend of efficiency-increasing technologies is sufficiently broad to encompass 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 critical domains, the trend could reverse. For instance, distributed fault tolerant computing arguably pushes the efficiency-robustness Pareto frontier in a way that favors robustness. Second, it 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[Dre17], where advancing technology increases available resources in a way unmatched by increased demand, it seems likely that robustness becomes less expensive in relative terms. This second scenario also assumes the absence of the type of supercharged competition in Hanson's proposed default Em scenario, where trivially easy expansion of Em population leads to a rapid and unlimited need for resources.[Han16]. Third, the hypothesis assumes that complexity overhangs are relatively hard to identify. This has been true historically, but computational simulation make reverse this trend, making the search for increased robustness efficient enough to counterbalance the more general and destabilizing increased efficiency that new technology promotes.

4 Conclusion and recommendations

As with Bostrom's other vulnerable world scenarios, the risks discussed here are plausibly greatly mitigated by restricting technological development, and effective global

governance. This risk, however, is not reduced by minimizing the variability of goals and motives of those looking for new and dangerous technology, nor via effective preventative policing. On the other hand, the existence of fragility risk argues strongly for a different type of risk-aware research prioritization. While individual actors (at the company, state, or regional level) may benefit from technological races that promote economic growth over systemic safety and robustness, the growing interdependence of international systems makes this risky. The economic bad of complexity is overprovided without intervention. Additionally, continuing the current trend of investment based primarily on the promised advantages of new technologies is a significant concern. For that reason, it seems likely that research funders should be prioritizing more thoughtfully in view of both risks, benefits, and the broader set of technologies being researched.

The existence of technological fragility risks does not, however, contradict the hypothesis behind Bostrom's fragile world scenarios. There can be 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 lowering the risk of 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.

References

[Bau15] Seth D Baum. Risk and resilience for unknown, unquantifiable, systemic, and unlikely/catastrophic threats. *Environment Systems and Decisions*, 35(2):229–236, 2015.

[Bos03] Nick Bostrom. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3):308–314, 2003.

[Bos13] Nick Bostrom. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31, 2013.

[Bos18] Nick Bostrom. The Vulnerable World Hypothesis. 2018.

[Dre17] Eric Drexler. Pareto-topia, 2017.

[Han16] Robin Hanson. *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press, 2016.

[Kuh18] Karin Kuhlemann. Complexity, creeping normalcy and conceit: sexy and unsexy catastrophic risks. *foresight*, 2018.

[May01] Robert McCredie May. *Stability and complexity in model ecosystems*, volume 6. Princeton university press, 2001.

[Par84] Derek Parfit. *Reasons and persons*. OUP Oxford, 1984.

[SD18] Gopal P. Sarma and Eric Drexler. Formal methods can provide secure software foundations and support AI safety. 2018.

[TDS+13] 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.

[TS13] Bruce Tonn and Dorian Stiefel. Evaluating methods for estimating existential risks. *Risk Analysis*, 33(10):1772–1787, 2013.