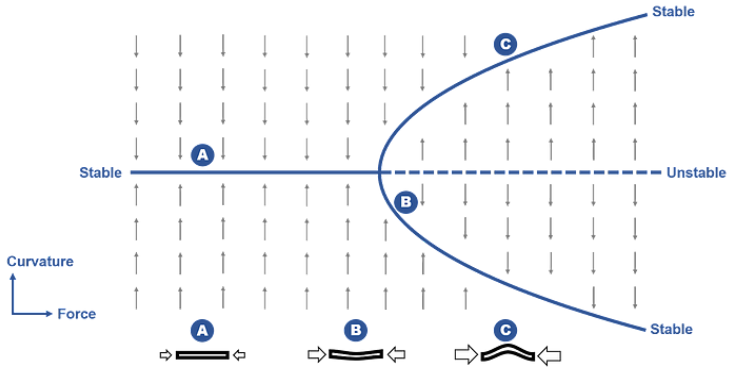
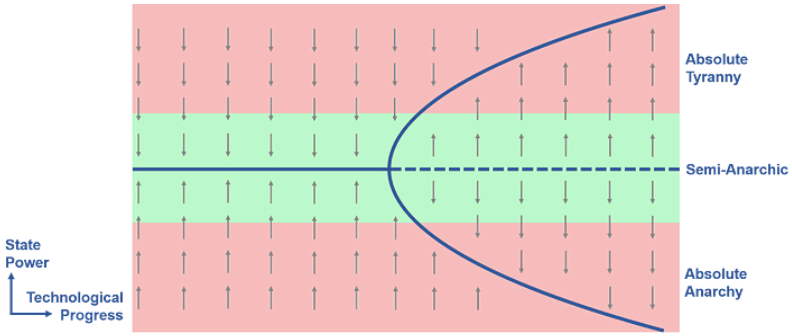
The Buckling World Hypothesis - Visualising Vulnerable Worlds

Motivation.
Mark Zuckerberg's notorious motto, "move fast and break things" [1], reflects a mindset shared by many of the most powerful entrepreneurs in Silicon Valley. This mindset rests on the assumption that the benefits of discovering advanced technologies will ultimately outweigh any disruptions (i.e., broken things) created along the way. However, Nick Bostrom's vulnerable world hypothesis (VWH) [2] presents a sobering alternative. It supposes that there exist certain advanced technologies whose discovery would break society to the point of devastation. The purpose of this article isn't to argue that we live in a vulnerable world, but rather to examine and visualise what such a world might look like. The intention is to provide a useful tool for policymakers governing advanced technologies, regardless of whether the VWH is proven true. In order to clarify the dynamics of a vulnerable world, we'll first explore a simpler phenomenon, that of a buckling ruler.

Introducing Buckling.
Imagine yourself holding a standard plastic ruler horizontally between your hands and applying some level of force to either end. For small forces, the straight ruler appears resilient (A). In this low-force regime, a small perturbation to the ruler's curvature is quickly resolved as the ruler snaps back into shape. However, there is a critical threshold at which the force becomes sufficient to provoke a dramatic change. At this point, the ruler succumbs to instability and buckles either downwards (B) or upwards (C). This illustrates how buckling can flip formerly stable states (a straight ruler) into instability while establishing new stable states (a curved ruler) that were previously unstable. Mathematicians would depict this in the diagram below [3].



Buckling as a model of the VWH.
The diagram presented below adapts our model of the buckling ruler to visualise a vulnerable world, replacing "force" with "technological progress[1]" and "curvature" with "state power." In this article, we employ a straightforward characterisation of "state power" as the capacity of the state to prevent and mitigate threats to its security. To qualify what may otherwise be an overly abstract diagram, we label three regions along the y-axis. In particular, we define regions of absolute tyranny, where a collapse of the existing order appears implausible, and absolute anarchy, where a return to order appears equally improbable. Finally, we label a semi-anarchic region that acts as a Goldilocks Zone between these two extremes. The model indicates that, for a moderate level of technological progress, there exists a stable level of semi-anarchic state power. However, beyond a specific technological threshold, the dynamics change, pushing states towards an excessive or diminished level of power. Regardless of whether this model accurately reflects reality, exploring potential explanations for this behaviour could offer valuable insights.



[1] In this article we consider technology in alignment Bostrom's definition in the original VWH paper [2]: "*We count not only machines and physical devices but also other kinds of instrumentally efficacious templates and procedures – including scientific ideas, institutional designs, organizational techniques, ideologies, concepts, and memes*"

Pre-Threshold Dynamics.
Why should there be a stable level of state power with modest technological progress? One potential explanation is that states with a modest level of technology are unable to tightly control the thoughts and actions of civilians, limiting their power. In the absence of advanced surveillance capabilities, a state that has drifted towards tyranny will struggle to anticipate the details of an inevitable revolt. Conversely, civilians without access to advanced technology may lack the insight, organisation, and resources to easily diminish state power. This prevents a perpetual threat to the state's existence, impeding the onset of anarchy and promoting a sense of order. In essence, the model predicts a self-correction mechanism that forces states with excessive or diminished power towards a stable baseline. Our hypothetical explanation for this stability is that the state lacks the technology to exert absolute power, and civilians (or foreign adversaries) lack the technology to easily threaten (or otherwise destabilise) the state. This mirrors the stability of the straight ruler in the low-force regime, where small deviations in its curvature are swiftly corrected as it snaps back into shape.

Post-Threshold Dynamics.
The stability described above only extends up to a certain technological threshold. Is there a hypothetical explanation for this behaviour? Advanced technologies enable individuals to gather, process, communicate, and act on information—all at a scale that perpetually challenges the power of the state. These challenges, or even the possibility of them, cast doubt on the state's authority, destabilising the semi-anarchic status quo. In order to reclaim authority, a state may attempt to control the individuals and technology that it perceives to threaten it. An illustrative example of these tactics can be seen in the U.S. response to the 9/11 terror attacks [4], marked by mass-surveillance, airport security, and the declaration of a "war on terror." Confronted with even larger threats, a government may respond in turn, exploiting advanced technology to weaken foreign entities or establish an all-encompassing surveillance state [5,6]. If a government established this level of absolute power, it would become difficult to overthrow, as civilian resistance could easily be predicted and prevented.

Nations opposing this tyrannic shift might find themselves on the brink of collapse, unable to manage the escalating threats enabled by advanced technology. Upon such a collapse, the transition from anarchy to order may be unusually challenging due to the hostile aftermath of an advanced society's downfall. In particular, consider the difficulties of rebuilding society in a world littered with dangerous technology or beset by an inhospitable climate. In summary, advanced technology expands the scale of human action, destabilising the semi-anarchic status quo and forcing states to consolidate power (absolute tyranny) or risk losing it altogether (absolute anarchy). Moreover, a technologically advanced state that reaches absolute tyranny or anarchy is inherently stable, as it becomes difficult to restore the semi-anarchic status quo. This mirrors the instability of the straight ruler in the high-force regime, where small deviations of the ruler's curvature are quickly amplified, causing it to buckle towards a curved, stable state.

Discussion.
Our model favours simplicity and, in doing so, overlooks some aspects of the VWH, rendering it a valuable but incomplete tool. Notably, our model reduces a complex range of socio-technological vulnerabilities into a single technological threshold where society buckles. This simple account of technological progress neglects the complex interplay between society and technology, such as the human capacity to pursue risk-reducing instead of risk-increasing technologies [7]. Furthermore, the model's characterisation of the dynamics of state power is perhaps overly deterministic. Unlike a ruler, whose shape is entirely determined by external forces, our civilization has the ability to resist the pull towards undesirable futures [8].

In summary, a more comprehensive consideration of perspectives on state power and technological progress would be a useful complement to our potentially reductive approach. Despite these limitations, we believe that our model offers a valuable visual tool to guide policy and evaluate humanity's macrostrategic situation. Whether or not our world is inherently vulnerable, technology will continue to exert a disruptive influence on people's lives that must be constrained by the state. Only through a concerted effort to anticipate and address these disruptions can we hope to prevent our world from buckling under the weight of its own progress.

References.
[1] Drake Baer. "Mark Zuckerberg Explains Why Facebook Doesn't 'Move Fast And Break Things' Anymore". Business Insider, 2019.

[2] Nick Bostrom. “The Vulnerable World Hypothesis.” Global Policy, 2019.

[3] Christopher Zeeman. “Euler Buckling.” Warwick Mathematics Institute, 1976.

[4] Robinson Grover. "The New State of Nature and the New Terrorism." Public Affairs Quarterly, 2002.

[5] Yuval Noah Harari. “Technology favours Tyranny.” The Atlantic, 2018.

[6] Ross Andersen. “The Panopticon is Already Here.” The Atlantic, 2020.

[7] Jonas Sandbrink, et al. "Differential technology development: A responsible innovation principle for navigating technology risks." SSRN, 2022.

[8] Markus Anderljung and Julian Hazell. “Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?” Centre for the Governance of AI, 2023.