Global Policy, in press

Metatechnological mapping and the Ye Wenjie effect: Mitigating civilizational vulnerabilities

Milan M. Ćirković
Astronomical Observatory of Belgrade, Volgina 7,
11000 Belgrade, Serbia
e-mail: mcirkovic@aob.rs

Abstract. In a widely cited study in this Journal, Nick Bostrom has posed the Vulnerable World Hypothesis: technological development, if occuring under conditions similar to those in the present, will make the devastation of civilization likely. In light of such drastic consequences, the hypothesis is worth seriously discussing in both bredth and depth. Two related proposals are hereby made and justified: creating a metatechnological map (or "tech tree") capable of telling in advance where exactly the dangerous technologies are, and reducing the size of the apocalyptic residual by antitotalitarian deradicalization and deprogramming. Both are modest proposals in the sense that they imply neither deep restructuring of human nature nor building instruments for potential totalitarian violations of civil rights and liberties.

Policy implications

- Mapping technological design space should be high on the list of priorities in order to be able to better assess the validity of the Vulnerable World Hypothesis, as well as a prerequisite to undertaking any action targeted at either increased oversight or curbing any threatening research directions.
- Mitigating the threat of totalitarianism, either local or global, should be given an additional layer of priority in light of both reducing the apocalyptic residual and preventing further radicalization due to the expanded surveillance and policing capacities (even if well-intentioned).

1. Bostrom's Vulnerable World Hypothesis

In an illuminating article in *Global Policy*, Nick Bostrom poses the Vulnerable World Hypothesis (Bostrom 2019, p. 457):

> VWH: If 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 semianarchic default condition.

He then proceeds to develop a taxonomy of civilizational vulnerabilities, as a useful tool for organizing further research, before giving policy-relevant suggestions for mitigating those vulnerabilities. Those suggestions have been already critically discussed from multiple viewpoints (e.g., Sears 2020; Manheim 2020; Liedo Fernández & Rueda 2021; Hobson & Corry 2023) in a lively debate. In this brief comment, I wish to focus on two aspects of the situation which are largely missing from the discussion so far: the *meta*technological task of determining which set of capabilities produces vulnerabilities covered under VWH and the need for elaborating properties of those actors (the "apocalyptic residual" in Bostrom's phrase) who would desire the devastation of civilization for whatever reason.

Although these two issues are clearly separate and independent, together they offer a deeper policy lesson: there are indeed things *apart* from establishing preventive policing and global governance which could be done to stabilize the vulnerable state of the world. Both current proposals could be best construed as additions to, rather than criticisms of, Bostrom's policy framework. It is also non-trivial that these things can and should be done *in the present*.[1]

The outline of the rest of this paper is as follows. Mapping of the technology space as a form of mitigation strategy will be considered in Section 2. Section 3 will focus on properties of the apocalyptic residual and their fateful link to totalitarianism. The concluding section will recapitulate the arguments and outline some directions for future research.

2. Metatechnological mapping: A missing approach?

Bostrom's study contains a dazzling wealth of important insights which inspire and will continue to inspire many researchers for years and decades to come. There is, however, one research direction – and a policy suggestion logically linked to it – which is almost conspicuously absent from the study, which starts and holds onto premise that technological evolution is essentially a random process (albeit constrained by both the laws of nature and what is available at any given epoch): understanding – and subsequently controlling – the evolutionary process itself. Bostrom starts with the "black ball" metaphor and almost never deviates from it (p. 455):[2]

> What we haven't extracted, so far, is a black ball: a technology that invariably or by default destroys the civilization that invents it. The reason is not that we have been particularly careful or wise in our technology policy. We have just been lucky.

Luck is notoriously difficult thing to prove in science. Even more remarkably, the concept of technology policy is mentioned in this paragraph – and nowhere else in the paper! Which is strange for a study which entirely revolves around the adverse consequences of technological

development. While the balls-in-an-urn may be a good model for philosophical thought experiments, it is quite suspicious for modelling the real world in anything but a very crude picture.

Here, one could argue that the premise of VWH ("technological development continues") is intrinsically ambiguous and hence the implication is invalid in strict sense even if other conditions stated by Bostrom are satisfied. In fact, one could argue that the implied picture of technological development is by default in itself "anarchic" or "semi-anarchic" in the sense very similar to what characterizes the present world order. That seems to beg the question: if semi-anarchic condition is an obstacle to proceed with safe (non-civilization-devastating) application of particular technologies, why should we wait for the threatening items to appear on the horizon and *only then* regulate their usage and access to them via surveillance, policing and governance? It makes more sense to work on the problem *at the source*; that is, in directing and controlling technological development itself, than on mitigating the risks with surveillance and policing. This would correspond to the distinction between *proactive* and *reactive* relationship appearing often enough in futures studies. Only proactive approach allows one to formulate the "technology policy" mentioned in the quote above.

All this applies to the definition of the "default semi-anarchic" conditions, which includes three elements (diverse motivations and limited capacities for preventive policing and global governance), all of which are relevant only *post hoc*. Ironically, in view of Bostrom's default view of technological evolution as essentially random, as mentioned above; one would expect that the "semi-anarchic" attribute covers this random element at the source, in the process of innovation itself. Hence, reducing the randomness at the source would act in and of itself to mitigate the default semi-anarchic conditions.

In other words, we should *look into the urn, identify black balls, and locate them* – and presumably leave them safely in the urn. This would amount to discovering structure of the real-world technology tree (or "tech tree"), usually known only from the world of gaming (see e.g., Ghys 2012; Heinimäki 2015; King 2021), but which is in fact a very powerful analytic tool in the "real world" as well. After we map the space of technologies, we can analyse the map in terms of risks and threats to civilization, and even the relevant capacities for stabilization. Finally, the measures could be taken to suppress the key threats one way or another.

Even if we leave out that last step (relinquishing or banning the "black ball" research directions) from our analysis, previous steps – understanding the "tech tree" and identify civilizational risks with each branch and bud – should be included in any mitigation program for stabilizing human civilization. This impacts many other conclusions of Bostrom's study, in particular those analyzing various policy approaches.

"In its general form, technological relinquishment looks exceedingly unpromising." (p. 462) My car broke down unexpectedly this morning. *In its general form, repairing a car looks exceedingly unpromising*. There are literally tens of thousands of parts in a modern automobile – how could

I ever hope to establish which ones failed in what combination and order? Not to mention the utter improbability of me knowing which part could be repaired and which one needs to be substituted for a new one. Should I get desperate? Should I seek radical, drastic solutions – like changing *all* parts, i.e. buying a new vehicle? Should I insist that – due to my lack of insight – the car manufacturer continuously monitor all the parts (using nanotechnology, for instance) and warn me prior to each failure?

Obviously not. The correct answer is (of course) not to dwell on my ignorance of car's composition and functions of each part but to consult an expert who *does* know functions and connections of each part or group of parts, what are likely causes of failure, and what needs to be done about those failing parts/groups of parts = a good car mechanic! What to me looks not just "exceedingly unpromising" but more akin to sheer wizardry is a routine part of daily life for him.

Why? The foremost reason, I submit, is that the defining characteristic of a good car mechanic is having a clear and sharp mental image of which groups of parts do what and which individual parts are most susceptible to wear, tear, and failure. This does not include precise description of each individual part in the sense of manufacturer's documentation (presumably running in tens of thousands of pages); no single human being, bar a few savants, cannot know that for any modern car. Instead, this mental image contains more compact, streamlined summary of the distribution of parts and their functional relationships – a *map* of the whole structure in terms of both morphology and function. Some pieces of the map can even work on purely intuitive, associative and vague level, which does not affect the outcome of the repair process. The better this map is – the more efficient and respected the car mechanic will be.

The example is not just accidentally linked to technology. In a very deep sense, technology *must always be mappable*, in contrast to highly abstract sciences or arts. While a particular technological solution may look like an entirely unexpected brainstorm of a single engineering genius working in her garage, there can be no branch of technology without the complex iterative process of developing, testing, optimizing, etc., where each discrete step is explicated and justifiable, usually in terms of already established or *adjacent* technologies. This does make description of individual technologies in the form of a map not just possible, but desirable and practical.

The same could be said about as not-yet-understood technologies studied under the heading of *exploratory engineering* (Drexler 2012; Armstrong & Sandberg 2013; among significant precursors are Dyson 1966 and Lem [1964] 2013): "the art of figuring out what techniques are compatible with known physics, and could plausibly be reached in the future by human scientists" (Armstrong & Sandberg 2013, p. 2). The very essence of exploratory engineering is to *map the landscape* of the space of possible technologies, as a precursor to any systematic research policy. The figure 1 of Armstrong & Sandberg paper (originally due to K. Eric Drexler) is a kind of zero-order trivialized outline of what we should strive for.

The idea has been envisioned – with plethora of thought-provoking details – in the science-fictional context by Canadian author Karl Schroeder in his fine 2005 novel *Lady of Mazes*.[3] Its protagonists, advanced posthumans centuries down the line, live in manifolds: communities of their choice defined by specific cultural settings (a pastoral version of early modern England, or a Native American tribe, a medieval village dominated by the clock-making guild, etc.). Their stability against societal upheavals is ensured by a metatechnology called the *tech locks*, which makes certain that each individual manifold stays at the appropriate level: Native American tribesmen use hybridization of maize, but not gene splicing; elven houses are built of oak and sandstone, but not ancient Roman (or modern) concrete, etc. The tech locks could not be possible, however, without full and complete database and a map of all human technologies. At the fulcrum of the story, the main protagonist accesses this database (Schroeder 2005, p. 255):

> Towers of data flickered into being around her. The arrow flattened out, broadened, became a plain. Thousands of other lines stood up out of that plain, like a forest.
>
> She moved her virtual body through the forest, checking the tiny labels on some of the lines: *Resistance, Capacitance*, said one; *Condensers, designs and uses*, said another. Instead of a forest, she imagined she was sailing across a sea of technologies, able with a gesture to pull any invention or principle to herself and, as if she was hauling a net full of fish, come up with all the other technologies that it necessitated.

Of course, this highly complex multidimensional database could be parsed and probed in many ways: "If she chose another view, she could see the anthropology and politics that spears, bows, and cannon each entailed." (*Ibid*., p. 256)

There has been some research in the real world along these lines. Several lines of attack have been proposed both in general methodological terms (e.g., Mishra, Deshmukh & Vrat 2002; Phaal, Farrukh & Probert 2004; Choi et al. 2012; Park et al. 2013) and in regard to particular technologies (e.g., Amadi-Echendu et al. 2011; Adamuthe & Thampi 2020; Heidary Dahooie et al. 2021). Clearly, the mapping should be accompanied by anticipatory ethical analysis, as pointed out in the case of nanotechnology by Philip Brey and many subsequent researchers.[4] We are obviously far from the comprehensive map envisioned by Schroeder, but the road *is* in front of us – to follow or not at our own peril. To be fair to Bostrom, this could be subsumed under his plausible move "be more cautious and do more risk assessment work" (p. 464).

There are multiple possible ways to proceed, which is impossible to adequately survey here. An interesting example is the methodology of the fuzzy multi-attribute decision-making (MADM) applied by Heidary Dahooie et al. (2021) to the aerospace industry. MADM is a kind of "meta" approach subsuming multiple individual techniques used to evaluate and prioritize technological alternatives, such as Monte Carlo models, Delphi surveys, patent analysis, relevance trees analyses, and others. Their evaluation and ranking is based on multiple mathematical criteria, incorporating the concept of fuzziness to handle uncertainty and imprecision in decision-making processes, as well as multiple criteria which can mutually

conflict. The evaluations and rankings are then aggregated in order to determine the overall preference for each alternative; this can be implemented in several ways, starting from rather simple weighted sums and proceeding toward more complex algorithms that account for the interactions between criteria. In general, fuzzy MADM methods offer a structured approach to map very complex problems, such as building a metatechnological map, while accommodating deep underlying uncertainties. In the age of AI, it is exactly such, computationally heavy models which are likely to play bigger and bigger role in forecasting of *any* kind.

A critic might argue that the complexity of the problem is too high and that we can never hope to map that huge multiparameter space. After all, the very fact that a particular technology has not been predicted so far tells us that its connection with the existing knowledge is tenuous and indirect at best. Moreover, the biggest threats may not come from some spectacular cataclysm (as in the "easy nukes" scenario of Bostrom), but from scenarios in which the risk takes a more socioculturally mediated form – e.g., a kind of detrimental biasing of our global discourse system enabled by radical advances in neurotechnology (think currently debated "deep fakes" taken to the n^{th} power). Predicting and modeling such extremely complex scenarios may be exponentially difficult, hence unfeasibly expensive, for a long time to come.

One should not despair in face of complexity, however. History is choke full of technological undertakings alleged too difficult or even outright impossible which turned out to be rather mundane on the timescale of decades or even years. From August Comte infamously arguing that we will never be able to establish chemical composition of distant stars to Orville Wright stating that "no flying machine will ever fly from New York to Paris", to a genius such as János von Neumann concluding in 1949 that we have reached the limits of computer technology.[5] So, the pronouncements of impossibility should always be taken with a tonne of salt; as should be assessments of the long-term impact of momentary fashionable and hip technologies. The latter is often encapsulated as Amara's Law, in honor of the great Roy Amara, one of the pioneers of technological forecasting: *We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run* (e.g., Amara 1988; Amara & Lipinski 1983). Long-run effects are usually consequences of the technological feedback loops and *synergies* between various branches of the tech tree. It is exactly the metatechnological mapping along some of the directions outline above which will enable us to better understand and even explain the empirical efficacy of Amara's Law and other similar regularities. While all this is not in and of itself an argument that the problem is solvable, it is indeed a useful reminder that the burden of proof that the envisioned mapping of the technological space is *impossible* sits upon the shoulders of the skeptic.

Note that all this is completely orthogonal to the issue of technological determinism, sometimes invoked irrespectively of discussions of existential risk (e.g., Smith & Marx 1994). Metatechnology map could invoke probabilistic as well as deterministic elements. Even if the map were deterministic, this is still *just a map, not the teritory* itself: it may show us where is relatively safe to go, but cannot guarantee that we shall really arrive there. The latter depends

on multiple complex entangled cultural factors, one of which is the nature and structure of the "bad actors" or the apocalyptic residual.

3. Totalitarianism as a double risk factor: The Ye Wenjie effect

The implications of VWH revolve around the notion of the "apocalyptic residual", which is Bostrom's term for "actors… who would act in ways that destroy civilization even at high cost to themselves." (p. 458) However, little attention is devoted to the properties of that set of actors, which is of obvious interest for risk analysis and relevant policy studies.

The apocalyptic residual is no more constant than the legendary "1%" of the richest persons on the planet, whose composition changes on the yearly basis, often dramatically. We need to avoid the error of *reification* of the apocalyptic residual, assuming that there is a fixed list of such actors; such an error is bound to lead to erroneous or confusing policy recommendations. The real question, paraphrasing Hercule Poirot, is not to ask *who* the suspect is, but to ask who the suspect *is*. In other words, the investigator is not primarily interested in the suspect's name, but in what properties and experiences make her a suspect; in the present case, what properties and experiences make one a part of the apocalyptic residual. Hereby I propose that **(i)** an important factor determining the properties of the apocalyptic residual is totalitarianism; and **(ii)** a way of mitigating the threat of this residual is to work toward limiting the extent and influence of totalitarian ideologies and movements. In other words, we need policies reducing – or at least not increasing – the *totalitarian potential* of existing and future societies.[6]

This could, to take again inspiration from an insightful work of fiction, Cixin Liu's novel *The Three-Body Problem* (Liu 2014), be called the *Ye Wenjie effect*. To recall the basic element of the plot of this extraordinarily successful – in both East and West[7] – book: during the Cultural Revolution in China, Ye Wenjie, an astrophysics graduate from Tsinghua University, sees her father get beaten to death during a struggle session by Red Guards. Ye herself is branded a traitor and is forced to join a labor brigade in Inner Mongolia. Due to her obvious intellectual brilliance, she is subsequently recruited by the military to work at the Red Coast base, a secret Chinese initiative to use high-powered radio waves to damage spy satellites. When their radio antenna detects a message from Trisolaris, an inhabited planet in the Alpha Centauri triple star system, urging humans to remain radio-silent, Ye does the contrary: disillusioned by her experiences and having reached the point of despising humankind, she responds anyway, inviting the Trisolarans to invade Earth to settle its problems. She even murders her husband, their commissar boss, and subsequently other people to keep the alien message secret, becoming a leader in what is essentially a terrorist fifth-column, the Earth-Trisolaris Organization (ETO).

Nobody will read Liu's masterpiece without shuddering at the skillful portrayal of Ye's ordeals and her cascading downfall into cynicism, misanthropy and, ultimately, apocalyptic terrorism.

Liu leaves no doubt whatsoever that the key impetus for turning a brilliant science student into immoral villainess was provided by excesses of the communist totalitarianism, which utterly dehumanized both its executioners and its victims. In contrast, Ye's main collaborator on ETO, an American billionaire named Mike Evans, was radicalized by deep ecological ideology, including a whiff of Rachel Carson and Peter Singer.[8]

This fictional portrayal is so deep, disturbing, and persuasive that it seems advantageous to use it as a scenario analysis in global risk studies. Thus, I propose what could be called the Ye Wenjie effect: *a significant fraction of the apocalyptic residual is radicalized by totalitarianism*. (The term itself is used in its most inclusive sense, as a shorthand for both repressive practices of totalitarian regimes and distopian tenets of totalitarian ideologies; cf. Arendt 1951.) It has important consequences for our in-depth understanding and mitigating the civilizational vulnerabilities.

Many scholars have noted the morbid ideological drive, so pervasive under totalitarian regimes (Nisbet 1943; Arendt 1951; Friedrich 1964; Aron 1968; Medvedev 1971; Gregor 2020; Desmet 2022). This is perhaps best encapsulated in the infamous motto of the Spanish fascist Falangists *¡Viva la Muerte!* (Long live death!). Reactions to such a drive are multifold, but for our present purpose two categories are of particular interest: accepting of – and often bizarre revelling in – the nihilist rejection of values in favour of empty slogans on one side, and forceful disapointment in the world itself, leading to cynicism, loss of empathy, and open misantropy on the other. In both cases, we encounter a psychological framework meshing very well with the role played by the apocalyptic residual under Bostrom's VWH. Obviously, this implies that devoting additional research and policy-making resources to the study of apocalyptic radicalization by totalitarianism makes perfect sense.

A particularly dangerous development may consist in a kind of vicious circle, when the *incrementally* strengthened surveillance and *gradually* implemented global policing create conditions of creeping totalitarianism. The latter would, in turn, lead to radicalization of more actors, thus strengthening the apocalyptic residual and therefore increasing vulnerabilities, especially those of Type 1 and Type 2-a.[9] This vicious circle must be broken if civilization is to be stabilized. Deradicalization of at least a fraction of the real and potential civilization-devastating villains is a pathway to be explored in this regard. Even if that does not drive the size of the apocalyptic residual down to zero, it will still be immensely helpful in "buying a little time" (Bostrom 2019, p. 463). And one should keep in mind that even in the case of Type 1 vulnerabilities, there is likely a logistic threshold requiring a minimal number of actors and their coordination.

Intuitively, a bit of good news could be that the prior state of being opposed to totalitarian government or ideology is already likely to reflect the underlying rational moral layer capable of making deradicalization easier and likelier to succeed. Further research will be necessary in order to better establish conditions and the optimal approach for such deradicalization – an important pathway for stabilization of civilization in the context of VWH.

4. Conclusions

While Bostrom's insightful and disturbing study is likely to remain an inspiration and a warning for years and decades to come, we should not refrain from critically engaging with it, especially in the domain of its policy prescription. Global surveillance and increased global policing, as policies with significant potential for totalitarian misuse, are not necessarily the only recourses available if VWH turns out to be true and in fact could be immensely counterproductive. Two policies suggested in this comment, appropriate to this rather early stage of our studies of this subject matter, are to undertake mapping of the technological space in order to be able to steer away from "black balls" in advance, and investing significant effort in reducing the role of the apocalyptic residual. The latter should be done in term of both reducing the number of people qualifying as such as per what I dubbed the Ye Wenjie effect, and their de-radicalization in ideological, psychological and moral terms. While such a mitigating effort may not be enough to ensure stability of civilization, it at the very least presents an additional measure which has not been much studied and elaborated so far. If anything, these ideas would help delineate how much could be done prior to recoursing to the "high-tech panopticon" and similar policies with significant totalitarian potential.

Future research is clearly necessary in order to reduce the probability of a highly destructive and unregulated "black ball" technology being used to devastate human civilization. In this area, we clearly need more interdisciplinary research; in particular, we need strong synergy between researchers dealing with technology forecasting, futures studies and global risk analysis on one side, and scholars of totalitarianism on the other. It goes without saying that research in this area should be quite high on our list of priorities, together with open-minded investigation of other mitigation options, such as extensive human space settlement.

References

Adamuthe, A.C. & Thampi, G.T. 2020, "Forecasting technology maturity curve of cloud computing with its enabler technologies," *Journal of Scientific Research* **64**, 239-246.

Amadi-Echendu, J., Lephauphau, O., Maswanganyi, M., & Mkhize, M. 2011, "Case studies of technology roadmapping in mining," *Journal of Engineering and Technology Management* **28**, 23-32.

Amara, R. 1988, "What we have learned about forecasting and planning," *Futures* **20**, 385-401.

Amara, R. & Lipinski, A. J. 1983, *Business planning for an uncertain future: scenarios & strategies* (Pergamon Press, New York).

Arendt, H. 1951, *The Origins of Totalitarianism* (Schocken Books, New York).

Armstrong, S. & Sandberg, A. 2013, "Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox," *Acta Astronautica* **89**, 1-13.

Aron, R. 1968, *Democracy and totalitarianism* (Weidenfeld and Nicolson, London).

Bostrom, N. 2019, "The vulnerable world hypothesis," *Global Policy* **10**, 455-476.

Brey, P. A. 2012, "Anticipatory ethics for emerging technologies," *NanoEthics* **6**, 1-13.

Choi, S., Park, H., Kang, D., Lee, J.Y., & Kim, K., 2012, "An SAO-based text mining approach to building a technology tree for technology planning," *Expert Systems with Applications* **39**, 11443-11455.

Ćirković, M. M. 2019, "Space colonization remains the only long-term option for humanity: A reply to Torres," *Futures* **105**, 166-173.

Dafoe, A. 2015, "On technological determinism: A typology, scope conditions, and a mechanism," *Science, Technology, & Human Values* **40**, 1047-1076.

Desmet, M. 2022, *The Psychology of Totalitarianism* (Chelsea Green Publishing, Chelsea, Vermont).

Drexler, E. 2012, "Physical Laws and the future of nanotechnology". Inaugural Lecture of the Oxford Martin Program, Feb,2012. https://www.youtube.com/watch?v=zQHA-UaUAeo.

Dyson, F. J. 1966, "The search for extraterrestrial technology," in Marshak, R.E. (ed), *Perspectives in Modern Physics* (Interscience Publishers, New York), 641–655.

Friedrich, C. J. (ed.) 1964, *Totalitarianism* (Grosset & Dunlap, New York).

Ghys, T. 2012, "Technology trees: Freedom and determinism in historical strategy games," *Game Studies* **12**, 143-52.

Green, B. P. 2019, "Self-preservation should be humankind's first ethical priority and therefore rapid space settlement is necessary," *Futures* **110**, 35-37.

Gregor, A. 2020, *Totalitarianism and political religion* (Stanford University Press, Redwood City).

Heidary Dahooie, J., Mohammadi, N., Daim, T., Vanaki, A. S., & Zavadskas, E. K. 2021, "Matching of technological forecasting technique to a technology using fuzzy multi-attribute decision-making methods: Case study from the aerospace industry," *Technology in Society* **67**, 101707 (11pp).

Heinimäki, T. J. 2015, *Technology Trees and Tools: Constructing Development Graphs for Digital Games*, dissertation defended at Tampere University of Technology (Tampere, Finland; https://cris.tuni.fi/ws/portalfiles/portal/4265122/heinimaki_1349.pdf).

Hobson, T. & Corry, O. 2023, "Existential security: Safeguarding humanity or globalising power?" *Global Policy* **14**, 633-637.

King, M. 2021, "The Possibilities and Problems of Sid Meier's *Civilization* in History Classrooms," *The History Teacher* **54**, 539-567.

Lem, S. [1964] 2013, *Summa Technologiae* (transl. by J. Zylinska, University of Minnesota Press, Minneapolis).

Liedo Fernández, B. & Rueda, J. 2021, "In defence of posthuman vulnerability," *Scientia et Fides* **9**, 215–239.

Liu, C. 2014, *The Three-Body Problem* (transl. Ken Liu; Tom Doherty Associates, New York).

Manheim, D. 2020, "The fragile world hypothesis: Complexity, fragility, and systemic existential risk," *Futures* **122**, 102570.

Medvedev, Z. A. 1971, *The Rise and Fall of T. D. Lysenko* (Doubleday & Co., Garden City).

Mishra, S., Deshmukh, S.G. and Vrat, P. 2002, "Matching of technological forecasting technique to a technology," *Technological Forecasting and Social Change* **69**, 1-27.

Nisbet, R. A. 1943, "Rousseau and totalitarianism," *The Journal of Politics* **5**, 93-114.

Park, H., Kim, K., Choi, S. & Yoon, J., 2013, "A patent intelligence system for strategic technology planning," *Expert Systems with Applications* **40**, 2373-2390.

Phaal, R., Farrukh, C. J., & Probert, D. R. 2004, "Technology roadmapping—a planning framework for evolution and revolution," *Technological forecasting and social change* **71**, 5-26.

Rhodes, R. 1986, *The Making of the Atomic Bomb* (Simon & Schuster, New York).

Schroeder, K. 2005, *Lady of Mazes* (Tor Books, New York).

Sears, N. A. 2020, "Existential security: Towards a security framework for the survival of humanity," *Global Policy* **11**, 255-266.

Smith, M. R. & Marx, L. (eds.) 1994, *Does technology drive history? The dilemma of technological determinism* (MIT Press, Cambridge).

Umbrello, S., Bernstein, M.J., Vermaas, P.E., Resseguier, A., Gonzalez, G., Porcari, A., Grinbaum, A. & Adomaitis, L. 2023, "From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice," *Technology in Society* **74**, p.102325 (11pp).

Wells, H. G. 1914, *The World Set Free: A Story of Mankind* (Macmillan & Co., London).

[1] There are other avenues to pursue for preventing the adverse outcome of VWH which are ignored or downplayed in the debate so far. Most notably, human space settlement is a straightforward and clear way of mitigating civilizational vulnerabilities, hence exiting the "semi-anarchic condition," not least due to the simple

fact that it will enable development and testing of risky technologies literally *far from civilization*. In general, space settlement is the strongest near-universal mitigation strategy for almost all existential and global catastrophic risks (e.g., Green 2019; Ćirković 2019). Unfortunately, it is not a near-term prospect which perhaps explains why it has not been invoked by either Bostrom or his critics.

[2] A mild deviation could be found on p. 463, the section on "Technological relinquishment".

[3] Some readers may object to using a science-fiction novel as referential example, but history of science will tell us that such cultural resources have been quite influential. In particular, the instructive episode told by Bostrom (p. 456) about Leo Szilard discovering nuclear chain reaction in 1933 has an additional, often untold layer arguably at least as important as Szilard's annoyance with Lord Rutherford: he was actively inspired by Herbert George Wells's prediction of nuclear warfare in his 1914 (!) SF novel *The World Set Free* (Wells 1914). Which is exactly why H. G. Wells features so prominently in the very *opening* chapter of Rhodes's magisterial history of the nuclear weapons (Rhodes 1986, pp. 13-28).

[4] Brey (2012); Umbrello et al. (2023) and references therein.

[5] Note that there is no symmetry here with technological forecasts which failed to realize due exclusively or mostly to political and social reasons. For example, we do not have widespread usage of cheap nuclear power (and perhaps even household modular nuclear reactors) as of 2024 only because unwise political and cultural decisions in the energy sector were made back in 1970s and 1980s.

[6] There is no implication here that Bostrom's study increases the totalitarian potential *tout court*. In spite of some misunderstandings (e.g., in Hobson & Corry 2023), it is abundantly clear that specific measures suggested by Bostrom increase the totalitarian potential *in proportion to* the magnitude of the threat. In other words, if we are unlucky enough, the world could be such that the "high tech panopticon" or something similar is necessary for avoiding the adverse outcome.

[7] Liu's novel became the first work of an Asian author to receive the Hugo Award for the best novel in 2015.

[8] Liu (2014), pp. 304-312.

[9] Those correspond to threat of technologies so destructive and easy to use that a small and otherwise disempowered apocalyptic residual could bring about the undesired outcome (Type 1) and the threat of technologies offering incentives to powerful actors to use it first (Type 2-a). Arguably, totalitarianism itself would be most successful (in a qualified sense of the word) in mitigating Type 2-b vulnerabilities, which are those requiring combined small actions of many actors to achieve devastation of civilization, since intrusive surveillance is likely to perceive any protracted and/or coordinated social process.