Differential Technology Development: Speeding up to slow down disasters?

Anders Sandberg

Reuben College, Future of Humanity Institute, University of Oxford

(This chapter is partially based on a paper under review, which presents the DTD strategy from a RRI perspective[1].)

Introduction

One of the earliest attempts at steering and applying technology in a socially beneficial way may have been Mohism, the school of Mo Zi during the Warring States era in China. Being committed to a harmonious (if authoritarian) peaceful social order they tried to dissuade rulers from attacking other states – and to develop and spread knowledge of defensive warfare, aiding cities and states under attack[2]. The aim was to make aggressive war impractical. This may also be the earliest attempt to speed up certain technologies to reduce the harm done by other technologies.

Technology has numerous consequences, many of which are morally relevant. Inventions such as the airplane implies the potential for bombing. Indeed, there is an affordance (in the design sense of a perceivable action possibility[3]) for certain forms of aerial warfare. Similarly, a nuclear weapon, while in principle useful for massive civil engineering projects a la Project Plowshare[4], has an obvious use as a weapon and a vast range of downstream causal effects on international conflict and existential risk that are far less predictable. The introduction of genetic engineering has a similarly unsurveyably broad set of implications, including some normatively unacceptable ones. While the "valence" of

[1] (Sandbrink et al. 2022)
[2] (Yates 1979; Loy 2015)
[3] This is a usage due to Don Norman's influential *The Design of Everyday Things* (1988), drawing on James J. Gibson's earlier work that treated affordances as actionable properties between the world and an actor. Norman added the perceived aspect: what appears to be possible is often more relevant than the full set of what is actually possible. (Norman 2004)
[4] (Johnson & Higgins 1965)

net good or ill caused by these technologies can be debated, even before coming into being some of their uses and misuses could be envisioned[5].

The technology completion conjecture

Given the consequential effects of new technology there has been much interest in whether it can be banned or controlled.

One challenge to this is the "technology completion conjecture". In Nick Bostrom's formulation[6] it is:

> If scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained.

If something is doable, eventually somebody will do it. Leonardo da Vinci envisioned flight using airscrews; helicopters became real a few centuries later. Charles Babbage envisioned a mechanical programmable calculator in the 1820s, and by 1991 (most of) his device was built[7]. JBS Haldane noted the applications of genetic engineering to enhance human health in 1923; by now we have examples of carefully done gene therapy and genetic selection, but also individually undertaken somatic (Elizabeth Parrish[8]) and germline (He Jiankui[9]) gene therapy for enhancement purposes without state oversight.

In a strict sense not everything doable can be done, given that intelligent life only has access to finite resources (due to the expansion of the universe limiting the reachable universe) and finite time (due to the likely decay of protons). The set of potential machines grows exponentially with the number of components and it is not hard to see that there has to be even fairly simple machines nobody will ever build.

---

[5] Indeed, there were plenty of science fiction or future-oriented speculation prefiguring each of these pointing out obvious and less obvious consequences.
[6] (Bostrom 2009)
[7] Interestingly, the machine likely could have been constructed with then-current tolerances. http://news.bbc.co.uk/1/hi/sci/tech/710950.stm
[8] (Parrish 2015; Sandberg 2017)
[9] (Greely 2019)

Yet this is an uninteresting complaint. The relevant capability is whether an advanced civilization can achieve all significant technological capabilities or have the ability to achieve an as-yet unachieved one if needed or desired. This need or desire may not be civilization-wide, but just belong to an individual or small group. As a general rule, the ability to achieve technological ends has become easier and more widespread over time.

The conjecture is in a sense a denial of the path dependency of technology. While temporarily path dependency shapes our technological repertoire, in the long run we will achieve Francis Bacon's "effecting all things possible". While there has been much discussion about path dependency and contingency in technology, most of it has been limited to relatively short spans of history. The conjecture boldly considers the total span.

The worrisome implication of the conjecture is that among the set of technologies there are some that are potentially *world-ending*. We have been lucky that nuclear weapons require rare substances that need to be refined at great expense in facilities not suitable for other purposes. A priori there is no reason why nuclear chain reactions could not have been triggered using household ingredients applied in a clever manner[10] - something that very plausibly could have led to the end of civilization had it become widely known[11]. If the technology completion conjecture is true future civilizations will have very dangerous technologies at their disposal, and if they are not managed well they will not survive long.

This threat is a good reason to deliberately trying to prevent the conjecture from becoming true through coordinated activity[12].

Banning, controlling and relinquishing technology

We can find plentiful attempts at reining in technology and research across history. Their

[10] Indeed, there has been at least one case of a natural fission reactor where a chain reaction occurred in a natural uranium ore, in the Franceville basin in Gabon (Gauthier-Lafaye, Holliger & Blanc 1996)
[11] (Bostrom 2019)
[12] (Joy 2000), (Bostrom 2019) The idea has a surprisingly long pedigree, see (Zaidi 2011) for attempts at reining in the potential destructive power of aviation through global agreements.

success has varied.

Less successful attempts

A common pattern is incumbent industries trying to prevent competitors or disruption (often using public interest arguments). Margarine was heavily regulated and taxed in North America, and prohibited 1886-1949 in Canada, largely due to lobbying from the dairy industry[13]. In the United States Congress alone there have been numerous bills attempting to outlaw the phonograph, gramophones, and VCRs[14] with the music or movie industry as major supporters. The various recording bills failed or merely resulted in in taxes from the sale of recording media going to the content industry.

The US government also attempted to limit the diffusion of advanced encryption technology in the 1990s, motivated by concerns about criminal use and national security. However, the export control regime was largely stymied by the portability of algorithms and computer code, as well as the growth of global trade on the Internet.

File sharing technology has been heavily prosecuted globally (supported by the content industry), and yet still remains a viable technology. While technical innovation in the file sharing field appears to have slacked off in recent years due to competition from convenient and relatively cheap streaming services, there is still a high rate of innovation in "darknets" and various forms of surreptitious information sharing and trade.

A more successful attempt of control in the US may have been online gambling. The Unlawful Internet Gambling Enforcement Act of 2006 did not outlaw online gambling but did prevent U.S. banks and credit card companies from processing payments to gambling sites. While not preventing circumvention, it has prevented the growth of a US-oriented online gambling market.

Another way of preventing an unwanted technology is restrictions on funding. The 1996 Dickey-Wicker amendment banned US federal research in which human embryos are

[13] (Ball & Lilly 1982),(Dupré 1999)
[14] https://gizmodo.com/10-technologies-that-congress-tried-to-kill-5874655

created, destroyed or harmed, presenting most embryonic stem cell research. While a new policy allowing research on 61 existing cell lines of embryonic stem cells was announced in 2001, federal involvement in this research (and funding) has been curtailed. However, privately funded research continued, US states responded with funding to prevent relocation of research overseas, and the field continues in the US and worldwide.

In the case of recreational drugs, many are banned or heavily restricted with perhaps more legal attention than any other individual type of crime, yet production and use is widespread despite decades of global efforts to both reduce demand and supply in a variety of ways. What seems to have resulted is rather a drastic reduction of scientific research into the use of the drugs (a reduction that is currently rebounding in the case of psychedelic drugs due to a shift in attitude). Meanwhile innovation on the criminal side has been ongoing. While much of it has been related to smuggling and cost effectiveness, there has been some genuine innovation in production (e.g. the Siegfried fentanyl synthesis[15]) and types (e.g. designer drugs evading legal restrictions).

More successful attempts

In discussions about *successful* bans on technology or research it is not uncommon to hear references to the Chinese Haijin sea bans 1371-1727 that banned overseas trade and relocated coastal people inland, or the Ottoman ban on printing 1423-1727[16]. Both are highly debatable, both in content, actual length, and effect. The sea policies were intended as anti-piracy and anti-insurgency, but also involved the economics and politics of tributary trade, shifting in focus, rationale and intensity over the dynasties[17]. The evidence for comprehensive Ottoman bans is scant and may be more of an European narrative to explain the assumed backwardness of the Empire, while actual printing business and policy was complex[18].

More well documented examples include:

[15] (Pardo, et al. 2019, pp. 58-67)
[16] (Cosgel, Miceli & Rubin 2012)
[17] (Zhihong 2006; Li 2010)
[18] (Pektaş 2015; Schwartz 2017)

Consumer DAT players, where the content industry successfully restricted a technology believed to enable piracy so that consumers could not record high-fidelity sound.

A proactive restriction of technology was the 1977 Environmental Modification Convention (ENMOD) and the Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques that prohibits widespread, long-lasting or severe effects due to military weather modification. While cloud seeding had been attempted during the Vietnam War the technology was limited and unreliable, and to this day militarily useful weather modification remains unrealized.

Another case of a proactive technology ban is the Protocol on Blinding Laser Weapons, Protocol IV of the 1980 Convention on Certain Conventional Weapons, was issued by the United Nations on 13 October 1995 and came into force on 30 July 1998. It prevents parties from employing weapons aimed at causing human blindness. The International Committee of the Red Cross lauded it just for being proactive:[19]

> The prohibition, in advance, of the use of an abhorrent new weapon the production and proliferation of which appeared imminent is an historic step for humanity. It represents the first time since 1868, when the use of exploding bullets was banned, that a weapon of military interest has been banned before its use on the battlefield and before a stream of victims gave visible proof of its tragic effects.

The protocol has been largely supported and overall interest in such weapons appear low. However, other weapons aimed at dazzling humans or blinding technology have been developed and may partially circumvent the protocol.

Nuclear weapons are the case where perhaps the most effort has gone into not just controlling their proliferation, but in restricting the ability for various parties to innovate and reducing the incentives to acquiring them. While keeping the general "nuclear secrets" secret turned out to be impossible, the non-trivial practical engineering requirements made restricting access to raw materials, processes and key components

[19] (ICRC 1995)

doable. However, perhaps most noticeably, maintaining a non-proliferation regime has depended on political pressure rather than legal restrictions.

This can be compared to the situation with chemical and biological weapons. Chemical weapons have been banned in treaties and have a fairly effective treaty enforcement organisation. That does not mean states (and in a few cases, individuals) have not acquired or even used them, but clearly it is a technology that could be far more widespread and is kept at bay. Biological weapons on the other hand lack treaty enforcement, and appear potentially growing in risk in the near term. Still, the strong taboo and general political agreement against them appear to be reducing the proliferation, at least for now.

Human reproductive cloning, in the wake of Dolly the Sheep, was largely condemned and banned in most countries. Here the restriction appears to have been successful simply because the demand for the technology is very low.

Locally, many regions such as the EU, have banned or restricted GMOs and products containing them, strongly limiting their spread and development.

Another approach is to classify or use secret patents to make it hard or impossible to develop technologies[20]. In the US the practice began at least in the 1930s, but was formalized by the 1952 Invention Secrecy Act that allow secrecy orders for patents detrimental to national security and reinforced by several international bilateral agreements. This has been used in nuclear weapons development, for example in the case of SILEX (Separation of Isotopes by Laser Excitation), but also numerous other fields. The system appears to have been effective in preventing technologies from entering the public domain, and reduced follow-on invention and commercialisation even when rescinded[21].

[20] Normally patents make innovations publicly known, speeding further innovation, which is a common rationale for supporting the patent system of time limited monopoly in exchange for disclosure. However, there may be situations where it is both individually and socially useful to rely on secrecy, such as trade secrets (Anderson 2011).
[21] (Gross 2019)

When do attempts succeed?

Some patterns that emerge from these examples:

Bans can work when (1) the infrastructure needed for the technology is large, expensive and visible, or the use obvious. (2) When nobody or only a few want the technology. (3) There is a single point of decision and enforcement.

Conversely, bans tend to fail when (1) the technology is diffusible or "lightweight", able to be developed in small, cheap and hard to monitor environments. (2) There are strong incentives to get it. (3) Multiple stakeholders are involved, either as regulators or as developers/customers.

Research and commercialization is more sensitive to bans than technologies themselves. Criminal entrepreneurs are already criminal, while researchers usually work openly within organisations endeavouring to follow the law and gain funding and prestige from the rest of society. Businesses often shy away from potential lawsuits or regulatory uncertainty. There is also research occurring privately (e.g. hackers and Internet rebels) that is little restrained by bans, and can easily be disseminated.

Nascent technologies may be shaped by bans in ways that introduce path dependencies, something incumbents sometimes use to prevent disruptions.
The existence of strong interest groups tend to drive attempts: when there is nobody to push for them economically and politically, technologies develop until they become hard to stop. While this may appear to be helpful for proactive attempts to stop them, the Collingridge dilemma implies that stakeholders are rarely aware of the technology when they could have the most influence[22].

None of the bans have been watertight.

Overall, this does not bode well if one thinks the world is vulnerable to world-ending

[22] (Collingridge 1982)

technologies and the completion conjecture is even approximately true. The prediction is that sooner or later "lightweight" deadly technologies will be developed by somebody regardless of attempts to ban them. This does not assume technological autonomy, just that the human-technical system is hard to budge deliberately due to its distributed nature, and that large groups of people will always contain some outliers[23].

Technology ordering

Technologies interact with each other in various ways. They also are developed and arrive in a contingent ordering. Generally, the order in which technologies arrive can influence the development of subsequent technologies and create pathways for future technological progress. This may provide an important lever of influence.

A technology can be dependent on other technologies. For example, telescopes require ways of making clear glass, a telegraph requires electricity, computers switching components, vacuum tubes or transistors (but are, in principle, independent of which kind of component is used). Hence the dependent technology cannot be developed without the first. This is a property that can sometimes be foreseen; for example Emmanuel Swedenborg observed in 1714 that heavier-that air flying machines would require motors that had a sufficient force to weight ratio.

Technologies can interact by one technology changing what another technology does. Computers enable complex, programmable control of anything electrically controllable. Radio communication makes other communications technologies (e.g. phones) mobile.

Technologies can substitute for each other, as in the case of switching components where relays, vacuum tubes and transistors can perform the same job. Petroleum products made whale oil unnecessary.

Technologies can also act through the wider techno-social system:

[23] (Bostrom, Douglas & Sandberg 2016)

Determining the distribution of resources: Technologies that arrive first often have a head start in terms of investment and development, which can result in an unequal distribution of resources and opportunities for other technologies.

Shaping social norms and attitudes: The first technologies to arrive can shape public perception and attitudes toward new technologies, influencing the reception of future innovations.

Setting the standard: The first technology to arrive in a market can set the standard for future technologies and make it difficult for later technologies to gain acceptance or compete effectively.

Creating network effects: The first technology to gain a critical mass of users can create network effects that make it difficult for subsequent technologies to catch up or compete.

Determining the regulatory landscape: The first technology to arrive can influence the development of regulations and laws governing its use, which can have lasting effects on the trajectory of future technologies.

These considerations make the ordering of technologies salient. While sometimes technologies arrive more our less simultaneously from independent inventors because the "time is right" (the preconditions are met, there is a widespread interest in developing them) it is not uncommon for technologies to spend long time as known "sleeping beauties" awaiting the right investment or effort (e.g. semaphore systems for long-range communications). More importantly many technologies develop at fairly slow rates and independently, making their ordering highly contingent.

We can imagine counterfactual worlds where technologies arrived in different order, producing better outcomes. For example, there is no strong reason why vaccines (or antibiotics) would have to be invented *after* long-range sea voyages began[24]. Had this been the case, it would have prevented massive pandemics in the New World.

[24] Inoculation against smallpox is documented in China from 1549, but may have occurred in the Song dynasty several centuries earlier.

One can also imagine a counterfactual where radioactivity and nuclear weapons were discovered/invented after space rocketry. In that world there is at least some possibility of having a humanity with off-planet settlements, strongly reducing the risk of extinction due to nuclear war.

These counterfactuals hint at potential benefits of adjusting ordering. While the technology completion conjecture assumes all technologies will eventually be found, their ordering can strongly affect the risks posed to the world.

The principle of differential technology development: reorder technology for safety?

If potentially harmful technologies are developed more slowly than technologies reducing their risks, their benefits will become available with less risk. This leads us to the following principle:

> *Differential technology development (DTD): Leverage risk-reducing interactions between technologies by affecting their relative timing.*[25]

This principle aims to get relevant actors to preferentially advance risk-reducing technologies and delay risk-increasing technologies.

The concept of differential technology development was first articulated by Nick Bostrom[26]. Bostrom introduces "differential technological development" to explore how the timing of advanced artificial intelligence relative to other technologies may impact associated risks. However, this principle can be applied within the context of responsible innovation

Responsible innovation already provides frameworks to consider and shape the impacts of particular technologies and projects. Differential technology development provides an

[25] (Sandbrink et al. 2022)
[26] (Bostrom, 2014)

additional framework considering how to prioritise a portfolio of technology development, treating it as a system with internal interactions that can be helpfully exploited.

In the following I will explore how "safemaking" technologies can enable DTD.
This can happen by safety technologies, by defensive technologies, or by substitute technologies arriving early or instead of risk-increasing technologies. Obviously, there are approaches that also try to slow the risky technologies. While such "negative" approaches can be important, I think the "positive" approaches for speeding up relevant technologies deserve more investigation, especially to meet the limitations of control discussed above.

Safety technologies reduce the negative impacts of other technologies by modifying them. They can be exemplified by carbon capture and sequestration methods that reduce the emissions of fossil fuel power plants[27]. In the case of dual-use technologies they can act by reduce the possibility of accidental or deliberate misuse, for example the permissive action links that in the 1960s began to reduce the risk of misuse of nuclear weapons[28]. They can also take the form of tools for monitoring compliance, such as trustworthy verification methods.

Defensive technologies decease the risk from another technology without modifying the technology itself. Car safety technology does not change the basic car technology, but add features (seat belts, crumple zones, safety glass, airbags, etc.) that reduce their danger.

Substitute technologies achieve the same benefits or ends as a risky technology while featuring less risk. Clean energy technologies that can replace fossil fuels, and replacements for ozone-depleting substances are good examples. Transmissible viral vaccine vectors have been proposed as cost-effective self-disseminating vaccines in animal reservoirs but clearly pose biosecurity risks; non-viral delivery methods could

[27] (Chu 2009)
[28] (Caldwell 1987) It is however noteworthy that despite PALs being mandated, the implementation in the military environment lagged behind strongly for many years due to military resentment, and codes may have deliberately been set to '0000000' (Blair 2004). A common insight in security engineering is: "Security should be built in, not bolt-on." There is enough path dependence in projects that safety technologies often become impotent if added late, a form of microscale DTD. This is even more true for how security is built into an organisation.

achieve the same end with less risk[29], and if viral delivery is nevertheless used methods that make the modifications non-heritable would be a good substitute[30].

Factors affecting the feasibility of DTD

Acceleration

These considerations changes the focus from preventing bad technologies to influencing the speed of technology development. The history of slowing technology is mixed: could speeding up desirable technology produce better results?

Two of the standard examples of speeding up technology development by concerted efforts are the US Manhattan Project aiming at a nuclear bomb, and the Apollo Project aiming at a moon landing. In both cases they accelerated development by putting significant resources into it, likely by at least a decade or more. However, such examples make bad examples because they were motivated by national security arms-races that used a noticeable fraction of a superpower budget.

A more peaceful example is the development of solar power and other low emissions technology. They were explicitly intended as substitute technologies. The original motivations may have been more long-running concerns about the finitude of fossil fuels, but later pollution and climate change became potent motivators. While the initial research costs were high and the cost of the clean energy uncompetitive, the persistent interest in improving the component technologies and the socio-technical system they were embedded in[31] paid off. For solar photovoltaic power an improvement curve ("Swanson's law") showing a price drop of 20% per module for every doubling of cumulative shipped volume. This is fairly typical for industry experience curve effects[32],

[29] (Sandbrink et al 2021)
[30] (Sandbrink & Koblentz 2022)
[31] (Geels et al, 2017)
[32] Often called Wrightean learning after (Wright 1936), that noted this behavior for aircraft production. The causes are manifold: learning-by-doing in the workforce, standardization, specialization and improved methods, changes in the resource mix used, redesign, network effects, efficiency spillovers between similar products, sometimes survivorship bias and early picking of low-hanging fruit etc. (Auerswald et al. 2000; Lafond 2022)

but when combined with exponentially growing sales due to cheaper module, the result is a Moore's law-like exponential improvement over time[33]. Similar experience curves apply for other low-carbon energy, and are rapidly making them competitive with fossil fuels. Forecasting based on these curves suggests massive savings from a rapid energy transition, making decarbonization much more likely[34].

Overall performance curves often have a sigmoid (s-shaped) or convex shape as a technology matures and approaches its limits. Sigmoid technologies need significant investment before they take off in performance or adoption, but then show exponential growth until they reach maturity/saturation. They are the quintessential disruptive technologies that may fly under the radar of more mature technologies before they overtake them. Convex development curves show large marginal improvements in early stages. Research often shows this property: the first investigations into a new area gain large amounts of knowledge, while later work needs to be more and more extensive to uncover new knowledge. This makes them more sensitive to changes in overall investment.

Anticipating technology valence

When Havelock Ellis asked the question "Will this gun not make war more terrible?" to Hiram Maxim, the inventor of the machine gun, Maxim answered "No, it will make war impossible". Numerous other versions of this self-assured false claim that a new weapon makes war less likely or less bad exist.

Another impressive failure of anticipation was the work of Pyotr Ufimtsev in the Soviet Union in the 1960s on creating surfaces that did not reflect radar. He was told that this work was of no significant military and economic value, and when it came to publication the military censors did not see a reason not to put it in the open literature. His book was translated by American engineers and used as input to the US stealth bomber project[35].

[33] (Nagy et al. 2013)
[34] (Way et al. 2022).
[35] (Ufimtsev 1971)

There are also many technologies like the Internet, encryption or human enhancement where the valence of the technology is not settled, or deeply contested. So, if even military censors during the Cold War could not see the relevance of stealth technology, what hope do we have in determining if a proposed technology is likely positive or negative?

That this determination is not always doable and frequently unreliable does not mean it cannot be done. It is fairly obvious that solar power is roughly positive, or that a novel long-lived environmental toxin is roughly negative.

Actionable anticipation may generally be easier: when risks are linked to defining features, when technology improvements are incremental, or the technology diffusion is slow or controlled.

Nuclear weapons could be anticipated as more destructive than conventional weapons long before being constructed, by virtue of the massive difference in energy scale between chemical and nuclear reactions. Gain-of-function research where the transmissibility of potential pandemic pathogens is increased can obviously be anticipated to have risks of accidents and misuse[36]. In both of these cases defining features of the technology imply risk.

Incremental technology development allows experience with precursor technologies to be scaled up. Large Language Models in AI will very foreseeably enable new forms of scams and disinformation just by scaling up the abilities of natural language processing[37]. When the technology grows with a smooth trend (e.g. Moore's or Swanson's laws) this also allows estimating how misuses can scale. For example, the exponential decline of DNA sequencing and amplification cost anticipate future near-impossibility of keeping one's genome private.

Slow or heterogenous diffusion of technology also allows empirical observation of its effects, hopefully enabling safe-making technology development (as in the case of car safety).

[36] (Duprex et al, 2015)
[37] (Caldwell et al. 2020; Seger et al, 2929)

The unilateralist curse and DTD

The unilateralist curse occurs when there is an action that it is enough one agent undertakes for the consequences to become global (e.g. telling a secret publicly, releasing a new organism in the environment, launching a geoengineering program). Even if the agents are all well-motivated and agreeing on what the good is, their fallibility will mean their evaluations will vary. That means that if the action is a net negative and the rational decision is to abstain, there is some probability that they will do it by mistake. If they cannot coordinate or share information, as the number of agents increase the probability of the action happening approaches one.[38]

This phenomenon is another support for the technology completion conjecture: somebody is likely to develop any technology, even if it is a net mistake. To discourage risky technology we need unanimity, and this becomes unstable as the number of actors increase. The general conclusion is that better coordination and institutions are needed to rein in the unilateral action or at least pool information to make estimates more accurate.

However, the positive flip-side speaks in favour of DTD. If an action is net positive, the curse turns into a blessing. Even if several actors mistakenly avoid the action, there is freedom to perform it. If we consider funding or initiating a project aimed at producing a safe-making technology, not all funding bodies need to be in agreement to make it happen. It is enough that some, or even one, do it for development to happen. Naturally, we may wish for more effective support, but that not every evaluator needs to be right produces resiliency. It also shows that diversity of evaluators can be valuable.

Technological fixes

When can human problems be solved through technology? While "technological fixes" often have a bad name, there are conditions that in general make them likely to work[39].

[38] (Bostrom, Douglas & Sandberg 2016)
[39] (Sarewitz & Nelson 2008)

The first one is that technology must largely embody the cause–effect relationship connecting problem to solution. In the DTD perspective, this means that safemaking technologies should ideally act directly on the problem the other technology causes. If a technology does something irreversible and this is the cause of concern, adding a simple method of undoing this change clearly ameliorates the problem.

The second one is that effects of the technological fix must be assessable using relatively unambiguous or uncontroversial criteria. Many technologies with "moral risk" like human enhancement are hard to evaluate since there is normative disagreement, and even within a normative framework evaluation may be too qualitative to help.

The third one is that research and development is most likely to contribute decisively to solving a social problem when it focuses on improving a standardized technical core that already exists. This also applies to purely technological problems: if entirely new tools are needed, they may take too long to develop, fail to develop, or be so specialized that they will not become widespread.

Summary

To summarize, DTD is supported when:

- Ability to correctly evaluate is good. This happens when a safe-making technology mainly has direct effects. It is less likely to work when there are multiple order effects, or in the case of general purpose technologies (since these are unpredictable and often have a complex mix of valence).
- There are multiple evaluators: there is more likely *some* support for safe-making.
- There is some ability to speed up development. This especially happens for understudied or very tractable technologies, such as simple ones or with an existing technology core. That something is unglamorous is often a good sign. Vested interests can both help and hinder, depending on the socio-technological landscape.

- The technology also needs to be affectable by R&D policy: not everything can be realized by grants. Sometimes competition, shared visions, or recognition that something is important can be motivating, but as work on green technologies has shown, this is rarely strong enough on its own.
- The safemaking technology also needs to have the ability to actually reduce the risks of the target technology. Ideally it is a simple technology with a direct causal link to the problem, with high diffusability (allowing it to be around where needed, as well as decline in price, have high acceptability and other scaling properties). If there are many substitutes available this can enable substituting for the risky technology – but conversely, there may be appealing but risk-unchanging substitutes for the safemaking technology.
- The safemaking technology can also be made safe.

Applying DTD to global risks

Can the DTD approach help with large-scale risks?

One clear example of DTD is the encryption technologies being developed against the future threat posed by quantum computers. When fully developed, quantum computers will be able to crack nearly all current encryption systems. This will be very bad for anyone that has a secret, i.e. everyone. But knowing this foreseeable problem, we can make safe-making technologies – quantum safe encryption systems – that can withstand quantum computer attacks. The NSA, Google and others are racing to get such technologies in place before quantum computers become super-powerful.

Biotechnology is an area with well-known risks such as accidental or deliberate production of novel pathogens, ecological disruption, and other kinds of biological hazards. There are many actors that can produce dangerous biological innovations, and emerging technologies such as AI supported genomics and lab automation may make them far more accessible to groups of people willing to use them unsafely[40]. At the same

[40] (Sandberg & Nelson 2020)

time biotechnology holds great potential for improving human health, food security, environmental restoration and so on. The field has many dual use technologies that have foreseeable risks. This leads to many opportunities for DTD interventions.

Gain of function research, where pathogens are deliberately modified or evolved to exhibit new and dangerous properties[41], is controversial due to disagreements about the cost/benefit balance. The DTD approach is to try to boost technologies such as universal vaccines against the entire class of pathogens investigated, or safeguards like deliberately inserted weak points (while trying to ensure proper safety and biocontainment of the pathogens).

Online ordering of DNA synthesis may allow malicious actors to source material. The DTD approach is to research ways of improving the screening of the submitted genes, so that it becomes more effective at catching misuse and discouraging it. This will require innovating beyond merely comparing to blacklists of pathogens of concern[42]. Similarly advances in benchtop synthesis will also require innovation[43].

GMOs may be risky in various ways if they spread in the environment, and various approaches to genetic containment have been proposed[44]. These include "xenobiology", engineering them to use a different genetic code not compatible with the code in the rest of the biosphere and hence unable to perform lateral gene transfer)[45]. The DARPA Safe Genes program[46] has aimed at control and reversibility of gene editing, ways of inhibit gene editing, and genetic remediation – ways of mopping up "spills" of GMOs. Xenobiology itself has some worrisome sides where proactive steps can be made to ensure such organisms do not get a fitness advantage[47].

Other forms of safe-making technologies for biotechnology is improvement in effective detection and handling of biohazards, as well as speeding up vaccine development and

[41] (Duprex et al. 2015)
[42] (International Gene Synthesis Consortium 2017; Diggans & Leproust 2019; SecureDNA 2023)
[43] (Carter, Tassif & Isaac 2023)
[44] (Lee et al 2018; Kim & Lee 2020)
[45] (Schmidt 2010; Nyerges et al. 2023)
[46] https://www.darpa.mil/program/safe-genes
[47] (Nyerges et al. 2023)

deployment. These are technologies that are in any case valuable, but get extra urgent if biotechnology accelerates.

Most of these examples represent technologies that already exist. A case of proactive attempts at controlling a technology is Artificial Intelligence (AI). Due to considerations of emergent risky behaviour of very powerful AI numerous voices have argued for restrictions, control or even relinquishment of the technology in whole or part[48]. However, the most plausible way of avoiding the risk, given the immense desirability of AI (essentially multiplying human capital indefinitely, getting access to new powerful modes of cognition) would be to develop techniques of ensuring its safety.

Limiting behaviour (such as isolating industrial robots from humans) only work in certain applications and misses many of the most valuable applications (e.g. autonomous vehicles or open-ended question answering). In the case of controlling smart systems controlling their motivations appears far more feasible than trying to keep an unruly system boxed up[49].

Designing a system that is intrinsically motivated to be well-behaved and safe, or at least not actively pursuing goals adverse to human flourishing ("AI alignment"), is at present an active field of theory and increasingly practice[50]. Here the DTD approach is very clear: were powerful AI to arrive before there was is a method of ensuring its safety, disastrous consequences seem likely. Indeed, in the AI alignment research community there is a very clear interest in accelerating progress on safety methods over increased AI capability methods.

It has also been argued that (safe) superintelligent AI might be a powerful safe-making technology in regards to nanotechnology, biotechnology and nearly any other global risks by virtue of it being able to detect and solve the problems, and hence it is important to develop it before e.g. advanced nanotechnology. This also leads to an argument that transitioning to an AI world is a risky transition, but remaining in the current vulnerable

[48] (Joy 2000)
[49] (Armstrong, Sandberg & Bostrom 2012)
[50] For general overviews, see (Bostrom 2014; Russell 2019; Christian 2020) and https://vkrakovna.wordpress.com/ai-safety-resources/ . For more technical detail on some approaches, see (Everitt, Lea & Hutter 2018; Hubinger 2020). This is a very fast-moving field and references will soon be obsolete.

state implies ongoing risk. The longer this risk is not defused thoroughly, the more likely it is that humanity perishes: hence there is a reason to take a cautious but speedy approach to AI. Whether this is a correct argument, what evidence and theory supports it, and how to distribute risk and cost fairly remains widely debated in the AI safety community.

It is also entirely possible that there are risks that not even superintelligence can handle or that its mere existence now introduces new risks. After all, human intelligence is superintelligent compared to most animals, yet we are often surprised by complex emergent phenomena in our world or from our actions, as well as technologies we invent yet cannot handle well.

Discussion

The DTD approach looks feasible in some domains – by not means all! Like any strategy, it has limitations and costs.

It encourages forethought: considering what emerging or possible technologies are risky and then constructively ensuring that they do not do too much harm. People often disagree on valence and consequences, but often less on what makes things safer.

Differential technology development may not be a panacea for dangerous technologies, just as teaching siege defence is unlikely to abolish war. But it can possibly introduce path dependencies that reduce overall risk. That suggests that an important area of technopolitics is not just what technologies get developed and how they are employed, but how they are paced relative to each other.

Positive DTD, speeding up technology development, may have a higher political "sellability" than slowing things down or banning them (although the numerous examples of incumbents trying to prevent disruption should nuance this claim). Any threatened limitation of funding to a research field usually meets a great deal of resistance, while opening up new funding avenues for desirable technologies generates excitement. At the

same time DTD will fail as an approach in (perceived or actual) zero-sum situations: it will only be unopposed if it can push the rope in a resource tug-of-war sideways.

Especially in academia research readily (perhaps too readily) follows funding. Still, DTD retains freedom of innovation – a very important aspect of an open society. It also allows having a wide variety of kinds of tools for reducing risks rather than trying to find the one great solution. The more safemaking strategies that are known the better the opportunity for finding ones that suit a new situation.

Technology, even if the technology completion conjecture is true, is fundamentally choosable. The combinatorial space of parts that can be combined into an invention is so vast that there must always be skilled decisions about what to explore. Innovation is built on implicit and explicit decisions about what to prioritize. That gives opportunity for judicious path-dependency.

J. Jonas Anderson, "Secret Inventions," Berkeley Technology Law Journal 26, no. 2 (Spring 2011): 917-978

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. "Thinking inside the box: Controlling and using an oracle AI." *Minds and Machines* 22 (2012): 299-324.

Auerswald, Philip, Stuart Kauffman, José Lobo, and Karl Shell. "The production recipes approach to modeling technological innovation: An application to learning by doing." *Journal of Economic Dynamics and Control* 24, no. 3 (2000): 389-450.

Ball, Richard A., and J. Robert Lilly. "The menace of margarine: The rise and fall of a social problem." *Social Problems* 29, no. 5 (1982): 488-498.

Blair, Bruce G. "Keeping Presidents in the Nuclear Dark: The SIOP Option That Wasn't." *The Defense Monitor: The Newsletter of the Center for Defense Information* 33 (2004): 1-3.

Bostrom, Nick. "The future of humanity." *Geopolitics, History, and international relations* 1, no. 2 (2009): 41-78.

Bostrom, N., 2014. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, Oxford.

Bostrom, Nick, Thomas Douglas, and Anders Sandberg. "The Unilateralist's Curse and the Case for a Principle of Conformity." *Social epistemology* 30, no. 4 (2016): 350-371.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455-476.

Caldwell, D., 1987. Permissive action links: A description and proposal. Survival 29, 224–238.

Caldwell, M., Andrews, J.T.A., Tanay, T., Griffin, L.D., 2020. AI-enabled future crime. Crime Science 9,

Carter, Sarah R., Jaime M. Yassif, and Chris Isaac. "Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance." Report. Nuclear Threat Initiative (2023).

Christian, Brian. The alignment problem: Machine learning and human values. WW Norton & Company,

2020.

Chu, S., 2009. Carbon Capture and Sequestration. *Science* 325, 1599–1599.

Collingridge, David. "*The social control of technology*." (1982).

Coşgel, Metin M., Thomas J. Miceli, and Jared Rubin. "The political economy of mass printing: Legitimacy and technological change in the Ottoman Empire." *Journal of Comparative Economics* 40, no. 3 (2012): 357-371.

Diggans, J., Leproust, E., 2019. Next Steps for Access to Safe, Secure DNA Synthesis. Front. Bioeng. Biotechnol. 7.

Dupré, Ruth. ""If it's yellow, it must be butter": margarine regulation in north America since 1886." *The Journal of Economic History* 59, no. 2 (1999): 353-371.

Duprex, W.P., Fouchier, R.A.M., Imperiale, M.J., Lipsitch, M., Relman, D.A., 2015. Gain-of-function experiments: time for a real debate. Nat Rev Microbiol 13, 58–64.

Everitt, Tom, Gary Lea, and Marcus Hutter. "AGI safety literature review." arXiv preprint arXiv:1805.01109 (2018).

Gauthier-Lafaye, F., Ph Holliger, and P-L. Blanc. "Natural fission reactors in the Franceville basin, Gabon: A review of the conditions and results of a "critical event" in a geologic system." *Geochimica et Cosmochimica Acta* 60, no. 23 (1996): 4831-4852.

Geels, F.W., Sovacool, B.K., Schwanen, T., Sorrell, S., 2017. Sociotechnical transitions for deep decarbonization. Science 357, 1242–1244.

Greely, Henry T. "CRISPR'd babies: human germline genome editing in the 'He Jiankui affair'." *Journal of Law and the Biosciences* 6, no. 1 (2019): 111-183.

Gross, Daniel P. *The consequences of invention secrecy: Evidence from the USPTO Patent Secrecy Program in World War II*. Cambridge, Massachusetts, USA: National Bureau of Economic Research, 2019.

Hubinger, Evan. "An overview of 11 proposals for building safe advanced ai." arXiv preprint arXiv:2012.07532 (2020).

ICRC (1995) "Vienna Diplomatic Conference Achieves New Prohibition on Blinding Laser Weapons and Deadlock on Landmines". https://www.icrc.org/en/doc/resources/documents/misc/57jmlc.htm International Committee of the Red Cross. 13 October 1995. Retrieved 1 July 2023.

International Gene Synthesis Consortium, 2017. Harmonized screening protocol v2.0. International Gene Synthesis Corporation

Johnson, Gerald W., and Gary H. Higgins. "Engineering applications of nuclear explosives: Project Plowshare." *Reviews of Geophysics* 3, no. 3 (1965): 365-385.

Joy, B. (2000). Why the future doesn't need us (Vol. 8, No. 4, pp. 238-262). San Francisco, CA: *Wired*.

Lafond, François, Diana Greenwald, and J. Doyne Farmer. "Can stimulating demand drive costs down? World War II as a natural experiment." *The Journal of Economic History* 82, no. 3 (2022): 727-764.

Lee, Jeong Wook, Clement TY Chan, Shimyn Slomovic, and James J. Collins. "Next-generation biocontainment systems for engineered organisms." Nature chemical biology 14, no. 6 (2018): 530-537.

Kim, Donghyeon, and Jeong Wook Lee. "Genetic biocontainment systems for the safe use of engineered microorganisms." Biotechnology and Bioprocess Engineering 25 (2020): 974-984.

Li, Kangying. *The Ming maritime trade policy in transition, 1368 to 1567*. Vol. 8. Otto Harrassowitz Verlag, 2010.

Loy, Hui-chieh. "Mohist arguments on war." In *Chinese Just War Ethics*, pp. 226-248. Routledge, 2015.

Nagy, Béla, J. Doyne Farmer, Quan M. Bui, and Jessika E. Trancik. "Statistical basis for predicting technological progress." *PloS one* 8, no. 2 (2013): e52669.

Norman, Don. "Affordances and design." *Unpublished article, available online at: https://web.archive.org/web/20190711202554/https://jnd.org/affordances_and_design/* (2004).

Nyerges, Akos, Svenja Vinke, Regan Flynn, Siân V. Owen, Eleanor A. Rand, Bogdan Budnik, Eric Keen et al. "A swapped genetic code prevents viral infections and gene transfer." Nature 615, no. 7953 (2023): 720-727.

Pardo, Bryce, Jirka Taylor, Jonathan Paul Caulkins, Beau Kilmer, Peter Reuter, and Bradley D. Stein. *The future of fentanyl and other synthetic opioids*. Rand Corporation, 2019.

Parrish, L. (2015) My name is Liz Parrish, CEO of BioViva, the first patient to be treated with gene therapy to reverse aging, ask me anything. (Accessed 30 July 2023) https://www.reddit.com/r/Futurology/comments/3ocsbi/ama_my_name_is_liz_parrish_ceo_of_bioviva_the/

Pektaş, Nil. "The beginnings of printing in the Ottoman capital: Book production and circulation in early modern Istanbul." *Osmanlı Bilimi Araştırmaları* 16, no. 2 (2015): 3-32.

Russell, Stuart. Human compatible: Artificial intelligence and the problem of control. Penguin, 2019.

Sandberg, Anders. "Morphological freedom: what are the limits to transforming the body?." (2017). In *L'humain et ses prothèses: Savoirs et pratiques du corps transformé*. Ed. Cristina Lindenmeyer. CNRS Editions, Paris.

Sandberg, Anders, and Cassidy Nelson. "Who should we fear more: biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk." *Health security* 18, no. 3 (2020): 155-163.

Sandbrink, J., Hobbs, H., Swett, J., Dafoe, A., & Sandberg, A. (2022). Differential technology development: A responsible innovation principle for navigating technology risks. *Available at SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213670

Sandbrink, J.B., Watson, M.C., Hebbeler, A.M., Esvelt, K.M., 2021. Safety and security concerns regarding transmissible vaccines. Nature Ecology & Evolution 5, 405–406.

Sandbrink, J.B., Koblentz, G.D., 2022. Biosecurity risks associated with vaccine platform technologies. Vaccine 40, 2514–2523

Sarewitz, D., & Nelson, R. (2008). Three rules for technological fixes. *Nature*, 456(7224), 871-872.:

Schmidt, M. (2010) Xenobiology: a new form of life as the ultimate biosafety tool. Bioessays. 32: 322–331.

Schwartz, Kathryn A. "Did Ottoman sultans ban print?." *Book History* 20 (2017): 1-39.

SecureDNA, 2023. Random adversarial threshold search enables specific, secure, and automated DNA synthesis screening. https://secure-dna.up.railway.app/manuscripts/Random_Adversarial_Threshold_Screening.pdf

Seger, E., Avin, S., Pearson, G., Briers, M., Ó Heigeartaigh, S., Bacon, H., n.d. Tackling threats to informed decisionmaking in democratic societies. The Alan Turing Institute, 2020

Ufimtsev, P. Ya. (7 September 1971). "Method of Edge Waves in the Physical Theory of Diffraction". Translated by Foreign Technology Div Wright-Patterson AFB OH. AD0733203

Way, R., Ives, M. C., Mealy, P., & Farmer, J. D. (2022). Empirically grounded technology forecasts and the energy transition. Joule, 6(9), 2057-2082.

Wright, Theodore P. "Factors affecting the cost of airplanes." *Journal of the aeronautical sciences* 3, no. 4 (1936): 122-128.

Yates, Robin DS. "The Mohists on warfare: technology, technique, and justification." *Journal of the American Academy of Religion* 47, no. 3, Suppl. (1979): 549-603.

Zaidi, Waqar H. "'Aviation Will Either Destroy or Save Our Civilization': Proposals for the International Control of Aviation, 1920—45." *Journal of Contemporary History* 46, no. 1 (2011): 150-178.

Zhihong, Shi. "China's overseas trade policy and its historical results: 1522–1840." In *Intra-Asian Trade and the World Market*, pp. 4-23. Routledge, 2006.