This is a rough draft from 2021 by Michael Aird, which I never got around to finishing or posting.

9 mistakes to avoid when thinking about nuclear risk

Right now, nine countries possess a total of roughly 13,000 nuclear weapons. In my view, the risk such weapons pose is one of the most important problems in the world today. But after a year of research on nuclear risk,[1] I've noticed several mistakes that I had previously made when thinking about this risk or that other people often make, and which I believe make it harder to effectively determine (a) *how much to prioritise* reducing nuclear risk relative to addressing other important problems, and (b) *how* to best reduce nuclear risk.[2]

Sometimes these mistakes take the form of people clearly, explicitly believing something that's incorrect. But often they instead take the form of people seeming to *implicitly* assume some false premise, or to *overlook some key point* - even if they might acknowledge the point if asked about it directly. Relatedly, several of these mistakes may strike you as rather obvious.

I wrote this essay not only to convince readers of things they might initially be skeptical of, but also to highlight to readers things that are obvious *once mentioned* yet aren't necessarily noticed or salient until then.[3] I also follow the description of each mistake with some information and links readers may find helpful for learning about issues the mistake related to.

Here are the mistakes I'll discuss:

1. Assuming substantial nuclear proliferation or substantial increases in stockpile sizes will very likely occur
2. …Or not even considering those possibilities

---

[1] Specifically, researching nuclear risk has been one aspect of my job (but not my whole job) since November 2020, and I had studied the topic a little before then. I've definitely learned a lot in this time, but nuclear risk touches on a vast array of topics and I'm aware I don't have deep expertise on any of them.

In this essay, I use "nuclear risk" as a shorthand for "risks posed by nuclear weapons". I do *not* mean to include risks posed by nuclear energy.

[2] People making these mistakes include (but aren't limited to) people associated with the effective altruism community, people who spend most of their time working on nuclear weapons risks, and well-credentialed academics.

[3] But what mistakes I've chosen to highlight is of course somewhat arbitrary; other mistakes may be more important and/or more common.

Note that my intention is to help people to avoid the mistakes I mention, rather than to criticise people who make these mistakes. Given that, and to save space, I have opted to avoid providing examples of specific people making each mistake.

3. Not distinguishing between smaller- and larger-scale nuclear conflict
4. Not distinguishing between countervalue targeting (e.g., targeting of cities) and counterforce targeting (e.g., targeting of ICBM silos)
5. Ignoring the possibility of major climate and famine effects following nuclear conflict
6. …Or overstating the likelihood/severity of those effects
7. Ignoring or denying that having nuclear weapons may have *some* positive effects
8. Assuming that eliminating nuclear weapons means *permanently* eliminating nuclear weapons
9. [[Ignoring possible technological advances]]

General lessons

There are a few recurring themes in these mistakes, which I think suggest some overarching lessons relevant to efforts to analyse and mitigate both nuclear risk and other global catastrophic risks:

1) **Be wary of very coarse-grained analyses; consider distinguishing a risk's importantly different subcategories, scenarios, stages, pathways, etc.**[4] E.g., notice that "How likely is nuclear war, and how bad would it be?" is a bad question, and replace it with something like "What are the important possible nuclear war scenarios, how likely is each one, and how bad would each one be?"

2) **Beware overconfidence; reality may be more complicated, scary, or stable than you think.** E.g., notice that there's no well-supported theory or wealth of data strongly suggesting nuclear conflict would remain limited or escalate massively, that it would or wouldn't cause the collapse of civilization, or that civilization would or wouldn't recover. We *can* and *should* make reasonable arguments and estimates, but we should also recognise our massive uncertainty.

3) **When prioritising between problems and prioritising between or implementing interventions, consider not only the *likelihood* but also the *severity* of a catastrophe.** E.g., focus more on preventing nuclear conflict scenarios that are more likely to involve large numbers of countervalue strikes (holding other factors constant), and consider interventions that reduce how bad nuclear conflict would be if it happened.

Mistakes 1 & 2: Assuming substantial proliferation or increases in stockpile sizes will *very likely* occur—or not even considering those possibilities

Some people seem to see substantial nuclear proliferation and/or substantial increases in arsenal sizes as more likely than seems reasonable to me. Other people seem to entirely ignore

[4] See also The Practice & Virtue of Discernment.

these possibilities and to thus underestimate nuclear risk[5] or fail to notice or appreciate the benefits of some options for intervening to reduce nuclear risk.[6]

Overall, I think **it's *unlikely* that in the coming decades there'll be proliferation to substantially more states** (e.g., to more than 2 new states) **or substantial increases in stockpile sizes** (e.g., any state doubling its arsenals). But **those possibilities do seem worth thinking about**, since they seem *plausible* - more than 1% chance, maybe more than 10% - and could substantially increase risk from nuclear weapons.

I find it useful to bear the following points in mind:

- There's no fundamental reason why there *couldn't* be nuclear proliferation to many more states over the coming decades, or why states *can't* multiply the total number of warheads or yield of their stockpiles.
- There's also no fundamental reason why those outcomes *have to occur*, or why there couldn't be *de*proliferation or *decreases* in stockpile sizes.
- It seems a wide range of future trends in proliferation and stockpile sizes are plausible, in light of history[7], the rising number of nuclear-latent powers[8], and potential future technological, economic, or geopolitical changes.[9]
- History also apparently suggests people tend to overestimate the likelihood or pace of nuclear proliferation.[10]
- [[Metaculus forecasts]]

---

[5] E.g., they may focus entirely on how likely extinction or societal collapse would be given *current* arsenal sizes, rather than also accounting for the low-probability, high-stakes scenario in which arsenal sizes rise.
[6] In particular, I think reducing the chance of *major increases* in arsenal sizes may be a relatively promising and neglected intervention option. E.g., one could aim to reduce the chance of a nuclear arms race between India and Pakistan or between the US and China.
[7] Consider that, in the decades after nuclear warheads were developed, the number of countries that possess nuclear weapons rose to 9 and global stockpiles rose to 70,000 warheads. On the other hand, in the decades since then, few countries have made serious efforts to develop nuclear weapons, and global stockpiles have declined to ~13,000 warheads.
[8] Nuclear latency is "the possession of many or all of the technologies, facilities, materials, expertise (including tacit knowledge), resources and other capabilities necessary for the development of nuclear weapons, without full operational weaponization" (Pilat, 2014). The number of nuclear-latent states has risen over time and seems likely to continue to rise in future (Pilat, 2014). I think this is probably mostly evidence that it's plausible there'll be substantial proliferation, but it could perhaps also be seen as evidence that it's plausible there'll be no or very little proliferation, since recent history suggests there can be an increase in nuclear-latency without an increase in the number of nuclear-armed states.
[9] For example, some plausible technological developments may reduce the cost or technological barriers of proliferation or of increasing arsenals sizes, or may threaten to undermine deterrence at current arsenal sizes and thus incentivise states to increase arsenal sizes.
[10] Disclaimer: I haven't thoroughly read that source I linked to.

Mistake 3: Not distinguishing between smaller- and larger-scale nuclear conflict

It's common - and understandable! - to just focus on estimating and reducing the chance of "nuclear war".

But "nuclear war" could theoretically - and, in my view, could *plausibly* - involve anywhere from a single nuclear detonation (with no retaliation) to tens of thousands of detonations or more (if arsenal sizes rise). And the number of detonations massively affects the expected consequences of the war (e.g., shifting the expected number of fatalities from the thousands to the billions).

Failing to distinguish between smaller- and larger-scale nuclear conflict can also lead to the following mistakes:

1. Implicitly treating *any* nuclear conflict scenario as being as bad as a scenario with hundreds or thousands of detonations, and thus overestimating nuclear risk and perhaps prioritising it too strongly relative to other important problems.
2. Not factoring the expected scale of a conflict into one's decisions about which nuclear conflicts to reduce the odds of (e.g., when thinking about whether to focus on potential conflicts between the US and North Korea or the US and Russia).
3. Not considering interventions that could reduce *how large-scale* nuclear conflict would be if it *does* happen (e.g., [[example]]).

Mistake 4: Not distinguishing between countervalue and counterforce targeting

Countervalue targeting is "the targeting of an opponent's assets that are of value but not actually a military threat, such as cities and civilian populations". In contrast, counterforce targeting is the targeting of assets which have military value, such as ICBM silos or nuclear command and control sites. Countervalue targets tend to have much higher population densities - and thus flammable material - than counterforce targets.[11] As a result, if a given number of nuclear detonations are on countervalue rather than counterforce targets, they'll typically cause *far* more immediate fatalities and loft *far* more soot into the atmosphere (increasing the risk of nuclear winter).

People often don't distinguish between these two types of targeting, or implicitly picture nuclear war as *certainly* a matter of strikes against cities. But **anywhere from 0% to 100% of**

[11] But note that counterforce targeting *can* involve detonations directly on high population density areas (e.g., capital cities, which are relevant to nuclear command and control), or detonations near enough to high population density areas that those areas suffer significant damage.

detonations in a nuclear war could be on countervalue targets,[12] **and that percentage massively affects the expected consequences of the war.**

As with failing to distinguish between smaller- and larger-scale nuclear conflict, failing to distinguish between countervalue and counterforce targeting can lead to the following mistakes:

1. Treating *any* nuclear conflict scenario as being as bad as one with mostly countervalue targeting, and thus overestimating nuclear risk and perhaps prioritising it too strongly.
2. Not factoring the expected proportion of countervalue targeting into one's decisions about which nuclear conflicts to reduce the odds of (e.g., this may suggest prioritising exchanges involving Pakistan or China [[add metaculus links and/or very brief reasoning]]).
3. Not considering interventions to reduce whether or how much countervalue targeting would occur given nuclear conflict (e.g., [[example]])

Mistake 5 & 6: Ignoring the possibility of major climate and famine effects following nuclear conflict—or overstating the likelihood/severity of those effects

When thinking about nuclear risk, people often focus on the immediate harms (e.g., from the blast) and the harms from radioactive fallout. And those harms could indeed be huge! **But those harms could be dwarfed by the harms from major cooling of the climate** - perhaps a nuclear winter, or perhaps a smaller version of the same effects. That cooling could perhaps cause huge numbers of famine deaths (plausibly in the billions, for some nuclear conflicts). And this seems the most likely way for nuclear war to cause an existential catastrophe.

...or maybe not! The effects depend on factors such as:

- how many detonations occur
- how much flammable material is in the targeted areas
- how much black carbon fires in these areas would produce and would reach high enough in the atmosphere to persist there for years
- how severely agricultural production would be reduced by various potential climate effects
- how people would respond to expected or occurring agricultural production issues (e.g., how well could they adjust what crops they grow, where, and how; how much would food usage patterns change; would international trade continue)

---

[12] [[Metaculus forecasts]]

[[See also other post]]

See also Ladish (2019).

- how likely civilization is to recover from a collapse

And, unfortunately, each of those questions are contested, complex, and under-researched.[13]

Ultimately, I suggest:

1. Recognising that major climate and famine effects are plausible, but that whether they'll happen and how bad they'll be is quite uncertain.
2. Seeing that as a key consideration when deciding (a) how much to prioritise nuclear risk relative to other problems and (b) which nuclear conflict scenario to prioritise reducing the odds of.
3. Considering interventions to reduce how bad the climate or famine effects would be (e.g., [[example]])

Mistake 7: Ignoring or denying that having nuclear weapons may have *some* positive effects

Some people who favour arms reductions or the elimination of nuclear weapons seem to ignore or deny that nuclear weapons can have *any* positive effects, even from the possessor's own perspective.

I think this is problematically black-and-white thinking.

As a general rule, I'd argue we should recognise that even things that are bad overall or in general are often good *in some ways*, *for some actors*, or *at some level*. Bearing this in mind can help us work out whether, under what conditions, and to what extent the thing in question really is bad. It can also help us understand (a) why an actor may resist our efforts to change

[13] I still consider myself quite confused on those topics, but here are some of my current bottom-line beliefs, in brief:

1. Climate effects severe enough to qualify as "nuclear winter" seem likely in scenarios in which there are thousands of nuclear detonations on high-population-density areas (e.g., cities or towns). In contrast, nuclear winter seems unlikely in scenarios with less than a hundred detonations on high-population-density areas. I really wish I had a clearer sense of the probabilities, especially for various scenarios between those extremes. (See, e.g., Toon et al., 2007; Reisner et al., 2018; Robock et al., 2019.)
2. If nuclear winter occurred, it would probably cause at least hundreds of millions of deaths. It's also *plausible but unlikely* that it'd lead to an existential catastrophe (via the resulting famine *combined with* other effects, e.g., further conflict triggered by the famine). (See, e.g., Aird, 2020; Beckstead, 2015; Ladish, 2020; Ord, 2020; Rodriguez, 2019; Rodriguez, 2020; Shulman, 2012.)
3. Climate effects that are similar to but smaller than "nuclear winter" could *plausibly* cause hundreds of millions or *perhaps* billions of deaths. But such effects would be much less likely to cause existential catastrophe.
4. Even a low probability of existential catastrophe is still really terrible and can be well-worth reducing further!

the thing (e.g., why nuclear-armed states may argue against the Treaty on the Prohibition of Nuclear Weapons), and (b) how to make effective compromises.

As for the specific matter of states having nuclear weapons, here are some possible positive effects of that (at least from a state's own perspective) :

- Having nuclear weapons probably increases *a given state's own* national security and geopolitical influence.
    - This is also a reason why it *might* be bad for the world if liberal democracies unilaterally reduce or (especially) eliminate their nuclear stockpiles. (And this is relevant in part because nuclear risk advocacy efforts are more likely to affect policy decisions in those countries than in other countries.)
- The fact various states possess nuclear weapons likely reduces the chance of *some* types of conflict (via deterrence), and *may* reduce the chance of armed conflict overall.[14]
- Possessing nuclear weapons may reduce a state's perceived need to have a large conventional military and/or to develop other offensive technologies (e.g., bioweapons), which could be a good thing.

(Overall, personally, I think it's very unclear whether it'd be good for various states or all states to massively reduce or eliminate their nuclear arsenals.)

Mistake 8: Assuming that eliminating nuclear weapons means *permanently* eliminating nuclear weapons

One somewhat common argument for eliminating nuclear weapons is, roughly, that "the only way to guarantee that nuclear weapons will not be used is to eliminate them".

But if a state - or *all* states - eliminated their nuclear weapons stockpiles, they or other states could potentially create new stockpiles in future.

It does seem *plausible* that a treaty to prohibit such actions - like the Treaty on the Prohibition of Nuclear Weapons - could be effectively verified and enforced. And such a treaty may be worth pursuing. But such verification and enforcement would probably be very difficult, and I'd guess that *some* state would still develop at least *some* warheads within a few decades if many states really wanted to - just as proliferation has occurred since the Treaty on the Non-Proliferation of Nuclear Weapons came into force.

(Furthermore, it doesn't even seem clear to me that temporary elimination of nuclear weapons by one state or all states would reduce the number of states that will have nuclear weapons in

[14] But whether and to what extent this is true is controversial. And of course, in any case, this benefit might be outweighed by the risk of *nuclear* conflict.

future, the number of weapons they have, or the likelihood that nuclear weapons end up being used.)

Mistake 9: [[Ignoring possible technological advances]]

[[TODO. Draw on and link to What technological developments could increase risks from nuclear weapons?]]

See also

- My post 20 things everyone (considering) working on nuclear weapons issues should know
- The series of posts on nuclear risk by me and other researchers associated with the think tank Rethink Priorities
- Jeffrey Ladish's One Hundred Opinions on Nuclear War
- 80,000 Hours' overview of the problem of nuclear security and how people can use their careers to address it

Acknowledgements

This research is a project of Rethink Priorities. It was written by Michael Aird. Thanks to Angela María Aristizábal, Avital Balwit, Damon Binder, Hamish Hobbs, Jeffrey Ladish, Ben Snodin, and [[REVIEWERS]] for helpful feedback. If you like our work, please consider subscribing to our newsletter. You can see more of our work here.

[Notes to self - not to be part of the post]

Archive

Notes to reviewers:

- This is a first draft. I plan to update it once I've gotten feedback, drafted my other nuclear risk posts, and opened more questions in our Metaculus Nuclear Risk Tournament (I'll integrate forecasts in this post where relevant). I'll then post this to the EA Forum in October.
- I consider this less important than my other nuclear
- **Things I'd particularly like input on are:**
    a. **Is there anything you disagree with?**
    b. **How can I make this more concise?**
    c. **Are there other misconceptions you think I should add?**
    d. In addition to posting this on the EA Forum and maybe LessWrong, I'm also leaning towards posting this on Medium and turning the content into a Twitter thread.[15]
        - Are you in favour of or against me doing those things?
        - If I do those things, what changes should I make to those versions?
    e. Who should I consider sending this to for feedback or further ideas?
    f. Should this be labelled as an RP post and listed on RP's site? Or should it be a personal post? If the latter, should it still be in the nuclear risk sequence?[16]
    g. Should I title this as a numbered list for the EA Forum/LessWrong version? For the Medium/Twitter version?[17]
    h. Should I bring "Error 7: Ignoring that progress has been made and risks have been reduced" back into the draft?
- This file is accessible by anyone with the link

To do

- Share for feedback, including from:
    - Peter
    - Other RP people

---

[15] Rationale for this:
- Lets me try my hand at that
- Could get more attention to our work and other good work via the links in this post
- Could draw more attention to nuclear risk and other longtermist issues
- Could help with broad dissemination of better beliefs on and ways of thinking about nuclear risk issues, which seems probably helpful in various little ways?

[16] Downsides of this being framed as an RP post:
- It's kind-of superficial and doesn't fully justify why I think people are wrong
- It's kind-of an arbitrary collection of hot takes
- It criticises common beliefs, maybe attributing them to particular people/orgs, which could make people angry at RP?

[17] I lean towards no for the Forum/LW version (instead saying "Some common misconceptions about nuclear risk") but yes for the Medium/Twitter version. See discussion of this general point here.

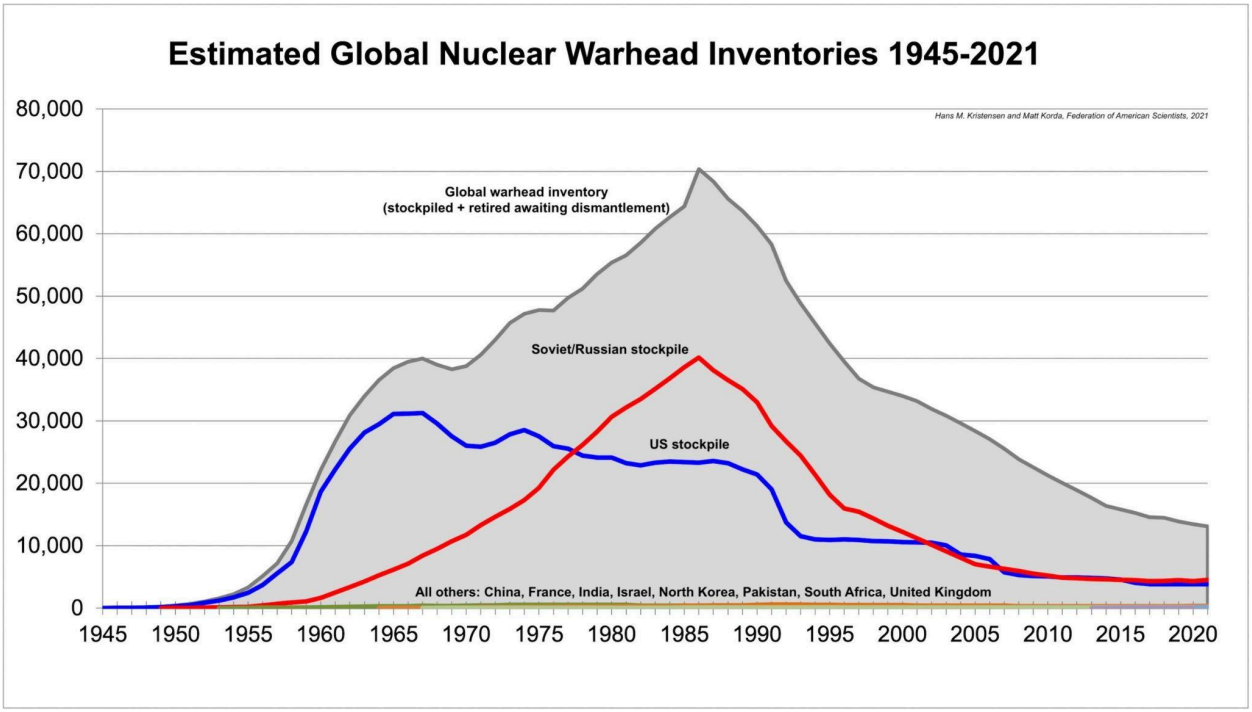
    - SERI fellows
    - Ladish
    - Jamie (I'm aiming for somewhat accessible and engaging to non-EA audiences)
- Update in light of changes to other posts, new forecasts, and feedback
- Share for feedback again
- Post

Sections I've tentatively decided to cut

Error 7: Ignoring that progress *has* been made and risks *have* been reduced

What I believe:

- Global nuclear warhead stockpiles trended terrifyingly upwards from 0 to ~70,000 warheads from 1945 to 1986, then fell to ~13,000 today.



- Odds of nuclear conflict between the US and Russia (which have always collectively possessed the vast majority of the world's nuclear warheads) seems lower now than it was during some particularly dangerous periods of the Cold War (e.g., the Cuban Missile Crisis)
    - That said, the odds do seem higher now than they were at some *other* times during the Cold War or recent history, and the odds of nuclear conflict between other pairs of states seem to have risen or to be set to do so.
    - (US-Russia nuclear conflict is especially important because those two states have always collectively possessed.

- [[Some evidence that close calls have become less frequent]]
- [[Something about fewer/no more efforts to make huge yield weapons?]]
- See also the excellent article [The world is much better; The world is awful; The world can be much better]
    - That isn't focused on nuclear risk, but the three titular points apply here too, as does this point: "What we learn from this is that it is possible to change the world. I believe that knowing that we can make a difference is one of the most important facts to know about our world."

Common errors:

- Some people confidently claim the risk of nuclear war is higher than it's ever been, that it's higher than it was during the Cold War, that it's rising, or things like that.
    - It's hard to be sure they're wrong. But I'd guess they're wrong that the risk is higher than it's even been or than it was for most of the Cold War, and I think the confidence with which these claims are sometimes made seems unwarranted.

Error 9: [[Close calls]]

What I believe:

- [[TODO]]

Common errors:

- [[TODO
    - Some people express what seems to be too much certainty that various "close calls" occurred or that they really did pose a substantial risk of escalating
        - I think Baum et al. 2018 point that fact out?
        - E.g., calling Petrov "the man who saved the world", rather than "the man who *might have* saved the world"
            - The latter statement would still arguably be too much, since even launching some nukes might not have meant launching loads, and even loads wouldn't necessarily have caused extinction or even collapse.
                - But I note the misconceptions related to those points separately above; they aren't what I'm talking about here
            - I think FLI are guilty of that?]]

Error 10: Ignoring that "eventually" is a very long time and that there are *many* ways the world can change by then

What I believe:

- [[TODO]]

Common errors:

- [[TODO
    - At least one person has stated/implies that the fact the probability is non-zero basically inevitable that nuclear weapons will eventually be used if not eliminated, implying that this is relatively urgent and demands a treaty
        - But "eventually" could be billions of years, if all we're talking about is that the probability is non-zero
            - And the same basic argument could arguably be used about literally anything else
        - And this ignores that situations might change relatively soon, in ways that makes use of nuclear weapons very unlikely or unimportant, for reasons very different from the introduction of a treaty
            - E.g., TAI
            - E.g., some technology renders nuclear weapons basically obsolete
                - Not sure what such technology would be like
                    - One possibility is extremely effective missile defence that's deployed so as to benefit all people
                        - That seems unlikely, though
                        - Maybe an altruistic billionaire could set that up in future?
        - The one clear case of this is Fihn: "But the risk will always be greater than zero, which means that given enough time, eventually will happen."
        - I think Ladish *maybe* makes a similar sort of claim here, but I'd have to re-read that with this in mind to check
            - Also here “However, the long term, deterrence will fail and nuclear wars will occur.”]]