Info hazard guidance for biosecurity discussions

Chris Bakerlee and Tessa Alexanian
Modified for the Biosecurity Fundamentals: Pandemics Course by Will Saunter with permission from the original authors.

Last updated May 2024

This guidance is intended for those facilitating or participating in the BlueDot Impact Biosecurity Fundamentals: Pandemics Course. It aims to provide a basic understanding of info hazard concerns and offer some norms for managing them responsibly, particularly in the context of discussions with people relatively new to biosecurity.

A lot of people (including ourselves) sometimes experience fear and uncertainty about how to handle info hazards; it's really challenging to balance concerns about inadvertently causing harm against our desire for open discussions and strong epistemics.

This document does not constitute an official policy. It is our work-in-progress attempt to offer guidance on info hazard issues and it's likely to change over time (e.g., in response to feedback). Note that the FAQ includes discussion of potential arguments for being relatively less cautious about info hazards, as do several of the recommended resources.

Summary

We believe that discussions of biosecurity and global catastrophic biological risks (GCBRs) are important to have, and should be approached with thoughtfulness and caution, because:

1. Generating or spreading information hazards related to biological risks can cause harm.
2. The costs of discussing biological threat models will often be larger than the benefits.
3. It's sometimes difficult to know when discussing GCBRs will be net-positive.

We offer the following recommendations for biosecurity discussions on the course:

1. Avoid brainstorming novel ways to cause harm with biology.
2. Ground discussions of threats in well-known historical examples where you can.
3. Focus on broad-spectrum interventions with limited downsides.
4. Consult others before taking actions towards reducing GCBRs.
5. Intervene kindly if discussions veer towards info hazards.
6. Use available resources, including reaching out for help.

This document also contains a list of concrete examples that seem safe to talk about and FAQ.

If you're grappling with how to handle info hazards, please feel free to reach out to discuss. You can contact Chris, anonymously or non-anonymously, through this short form. You can also reach out to Andrew Snyder-Beattie here or Tessa Alexanian (who is not a grantmaker in this space) here. If you're raising concrete infohazard concerns, please don't detail them – we can follow up more securely. This guidance is a work in progress, so please have a very low bar for sending us feedback.

Why be concerned about info hazards in biosecurity discussions?

Discussions of biosecurity and global catastrophic biological risks (GCBRs) are important to have. These discussions should be approached with thoughtfulness and caution, because:

1. **Information hazards[1] can cause harm**
   The historical record shows that state and non-state actors have pursued the capability to cause large-scale harm with biology, and likely continue to do so.[2] In light of this, generating or spreading information relevant to biological risks can have unforeseen and unintended negative consequences. To give a few real and hypothetical examples:
   a. According to an email sent by Ayman al-Zawahiri to a compatriot in 1999, al-Qaeda only became aware of and began pursuing biological weapons when "the enemy drew [al-Qaeda's] attention to them by repeatedly expressing concerns that they can be produced simply."
   b. While uncovering and sharing vulnerabilities in our biosecurity infrastructure (e.g., existing medical countermeasures, detection systems) can galvanize efforts to fix them, it can also lead bad actors to take advantage of them. For example, we're uncertain whether this disclosure about DNA synthesis screening was net-positive (probably yes).
   c. Hypothetically, raising the alarm about the concept or details of a novel bioweapon could help the world develop stronger defenses to that threat, but also could increase the probability this bioweapon will be created (e.g., by state and non-state bioweapons programs, biodefense programs, academic or industry biologists).

2. **The costs of discussing biological threat models will often be larger than the benefits.**

   An info hazard becoming known a few years early could be very costly. A contrived but illustrative example (that attempts to be quantitative): Imagine there exists a novel bioweapon design that represents 5% of catastrophic biorisk over the next 30 years (a timeframe relevant for convergent technological risks like AI). Assume the design is re-discoverable and will be widely discussed in exactly 10 years' time if we do nothing. In this scenario, widespread common knowledge of the design *today* would increase

[1] Here, we use a broader definition of "info hazards" than has been suggested by others; not information that is hazardous to know or think about, but true information that increases the probability of a GCBR if it falls into the wrong hands.
[2] See information on historical biological weapons programs by states; US intelligence agencies reported in 2022 that Russia and North Korea maintain offensive bioweapons capabilities. We also know that past non-state groups, such as Aum Shinrikyo in the mid-1990s, have attempted to build bioweapons.

our time exposed to risks from it by 50% (10 years on top of 20). Assuming risk increases linearly with exposure time, this could add 2.5 percentage points to the total 30-year risk of a global biological catastrophe, which is a lot.

Meanwhile, the potential benefits of learning about info hazards may not be as great as one thinks.

- Some threats could be very difficult to block. When people find a cybersecurity vulnerability, it can usually be patched with some effort. Patching specific vulnerabilities identified in biological systems will often be *vastly* harder.
- Knowing about specific threats wouldn't necessarily change our priorities. For many threats, the best things to do could be the things we planned to do anyway to block a wide range of threats (e.g., pursue better indoor air quality, PPE, rapidly reconfigurable vaccine platforms).
- That said, there will likely be situations where very carefully disclosing an info hazard is the right call, for example disclosing to a small number of people who would be able to develop countermeasures.

3. **We should be humble about which discussions will prove net-positive or -negative**
   It's plausible that smart, security-minded researchers who sounded the alarm about nuclear weapons in the 1940s accelerated the nuclear arms race in a harmful way. Like many communities, the biosecurity community faces these sorts of security dilemmas, and we should be humble about this.

   This guidance is addressed particularly at discussions that focus on the most extreme, catastrophe-scale risks. Since many people working on biosecurity and pandemic preparedness do not share this focus, we worry about finding ourselves in a world where the vast majority of the brainstorming of GCBR-relevant info hazards is coming from well-intentioned discussion groups such as those in this course.

   Moreover, while any particular discussion of potential info hazards is unlikely to cause significant harm, we should uphold norms that guard against widespread info hazard brainstorming and discussion, since the effects of such discussion could inadvertently end up very harmful.

Recommendations for discussions

In light of all this, we offer the following recommendations for biosecurity discussions on the course:

1. **Avoid brainstorming novel ways to cause harm with biology (e.g., via red-teaming).**
   In particular, don't create an environment where everyone is sitting around trying to figure out how to kill people with biotechnology.

2. **Ground your discussions of biological threats in well-known historical examples where you can.**

We recommend drawing on examples of biological threats from the historical record, which sadly includes many well-documented cases of natural and anthropogenic biological threats (e.g., smallpox, 1918 influenza, the Soviet biological weapons program). Beyond that, if it helps frame the conversation, we think one could also pretty safely consider the following threat, *without speculating about the underlying biological details*: a hypothetical respiratory virus (let's say "SARS-CoV-X") with a 99% fatality rate, measles-level contagiousness, resistance to all current vaccines, and/or a 6-month pre-symptomatic infectious period. (See the concrete examples below for more things that seem generally safe to talk about.)

3. **Focus on broad-spectrum interventions with limited downsides.**

   There is a huge range of technical and non-technical interventions with the potential to broadly reduce GCBRs and relatively little downside risk, from building metagenomic detection systems to strengthening the Biological Weapons Convention. These "broad-spectrum" interventions could be pursued without much need to think about the specifics of potential risks. We don't have great answers yet on how to *actually* do these interventions well – the details remain murky, and there's a lot to dig into. (Note, however, that it's still worthwhile to stay cautious about info hazards and dual-use potential[3] even when exploring these interventions.)

4. **Consult others before taking actions towards reducing GCBRs.**

   Be an *agent*, not a *unilateralist*. Even if twenty-four people have independently thought of an info hazard and kept it to themselves, it only takes the twenty-fifth discoverer to irreversibly spread the information. Rather than making judgment calls unilaterally, we encourage checking in with others—ideally more senior members of the biosecurity community, see below—before taking significant independent actions toward reducing GCBRs, such as publishing a blog post or launching a hackathon.

5. **Intervene kindly if discussions veer towards info hazards.**

   We have lots of practice in our daily lives with "filtering" what we share with others. Still, it can be tough to manage info hazard concerns in real time. It's not easy to strike a balance between openness and conscientiousness in these kinds of interactions. If you're anything like us, you're going to make mistakes, and your colleagues will, too. We all make mistakes, there is a lot of uncertainty here, and we need each other's support. See examples of how you might intervene kindly below.

6. **Use available resources, including reaching out for help.**

   A list of recommended resources is provided below.

   Beyond that, I (Chris) personally see it as a core goal of mine to support people in working responsibly and effectively toward GCBR mitigation, and I'm happy to troubleshoot issues or questions that come up. This could include questions around

[3] For example, some vaccine platform technologies present dual-use concerns.

general info hazard management, or questions around more specific info hazard concerns. The best way to contact me, anonymously or non-anonymously, is through this short form. (Alternatively, you could reach my colleague Andrew Snyder-Beattie here.) ***Importantly****, please do not include potentially sensitive details of info hazards in form submissions or emails to me* – if necessary, we can arrange more secure means of follow-up communication, anonymous or otherwise (e.g., a phone call).

You can also feel free to reach out to Tessa for advice on general questions related to navigating info hazards at hello@tessa.fyi.

There's a *tremendous* amount of work to be done toward mitigating GCBRs, and you as participants on the course are well placed to make a major positive impact. This document is neither the beginning nor end of the story on managing info hazards. Norms and guidance may shift over the coming months and years. But in the meantime, we hope these recommendations help you have productive conversations in this area.

Recommended resources

- Information hazards in biotechnology (article*, especially case studies section*)
- *Biosecurity Dilemmas* by Christian Enemark (book, audiobook, *especially "Secure or Stifle" section*)
- *Restricted Data* by Alex Wellerstein (book, audiobook)
- Horsepox synthesis: A case of the unilateralist's curse? (article)
- Avoiding the Biological Security Dilemma: A Response to Petro and Carus (article)
- Information Hazards: A Typology of Potential Harms from Knowledge (article)
- Information Hazards: A Very Simple Typology (blog post)
- Bioinfohazards (blog post, *including section "Risks from Secrecy"*)
- How can we improve info hazard governance in biosecurity? (blog post)
- Slate Star Codex: The Virtue of Silence (blog post)
- BBC Radio 4 - Apocalypse How, Death by DNA (podcast)
- Conjecture's Information Hazard Policy (article)
- Hear This Idea - Esvelt and Sandbrink on Risks from Biological Research (podcast)

Concrete examples that seem safe to talk about

While one should be cautious about attention hazards (e.g., amplifying the signal of public but relatively obscure information), it's generally pretty safe to discuss things that are already well known. On this basis, we personally would not be concerned if topics from the below *non-exhaustive* list came up in discussion (We've run these by some experienced colleagues who agree).

We hope that people who are relatively new to biosecurity will be able to refer to this list to identify examples they can discuss while experiencing less fear, uncertainty, and doubt than they would otherwise.

- As mentioned above, without speculating about the underlying biological details: a hypothetical respiratory virus (let's say "SARS-CoV-X") with a 99% fatality rate, measles-level contagiousness, resistance to all current vaccines, and/or a 6-month pre-symptomatic infectious period
  - Speculating about the underlying biological details of such a virus *would* pose info hazards
- Terrorists with the capability could in principle release multiple pandemic pathogens at once, and this is one respect in which deliberate attack scenarios may differ from natural spillover scenarios in some ways (though probably not in other ways)
- The lethal dose of botulinum toxin
- The documented mechanisms of action of various antibiotics, antifungals, and antivirals
- This notable paper, one of two that sparked the 2012 "gain of function" research debate
- This (in)famous paper about synthesizing the extinct horsepox virus
- The researchers who inadvertently developed a mousepox virus which evaded immunity
- The recent example of an AI drug discovery model being used to identify potentially lethal toxins
- The general idea that you could in theory make a biological weapon that targets a specific person or group of people, e.g., using some sort of DNA-recognizing system like CRISPR
  - Specific mechanisms that could do this *would* be an info hazard
- The fact that not all providers of synthesized DNA adequately screen their orders for dangerous DNA or malevolent customers , inc. the fact that journalists were able to order viral genome fragments in 2006
  - A list of specific companies that fail to screen their orders *would* be an info hazard
- The list of organisms that were developed as weapons by 20th century biological weapons programs, and the federal US list of select agents and toxins
- Items on the Australia Group's export control lists
- The fact that the genome of the variola virus (which causes smallpox) is freely available online
- The fact that accidental releases of select agents from labs are fairly common
- Kevin Esvelt's work on gene drive technology and countermeasures
- Current capabilities of LLMs for providing information on how to develop classic biological weapons agents, e.g., anthrax, including publicly-disclosed methods for measuring those capabilities
  - Information related to LLMs' ability to generate ideas for novel biological weapons could, however, pose info hazards
- Almost anything discussed in the materials in this resource list

Examples of how you might intervene kindly

It seems important to distinguish a few different situations in which you might want to intervene:

- **Course correction**: Sometimes our conversations drift towards info hazards, and we end up speculating about biological threats without having thought much about the relative costs and benefits. In these situations, deftly changing the subject and nudging the conversation back on track—with or without explicitly acknowledging the issue, depending on the circumstances—is an act of kindness.
- **Genuine disagreements**: In other cases, costs and benefits might have been considered, yet there's underlying disagreement about the risk profile of certain info. We don't yet have great generic advice to offer for such situations, except to point to the non-exhaustive list of "generally safe" items in the concrete examples above and once again underscore the need for 360° patience and kindness.
- **Repeated forays into dangerous territory**: If there are recurring or systematic issues that you feel are taking conversations into dangerous territory, we encourage you to speak up to fix them, or reach out for help (can contact Will Saunter or other BlueDot Impact staff on Slack, or alternatively you can contact Chris here).

Some potential ways you could intervene are:

1. **Interrupt.** If you're not feeling comfortable with the direction the discussion is taking, you can jump in and say this! In fact, we'd recommend setting explicit norms in your group that anyone should feel free to interrupt a discussion and check in about info hazard concerns before proceeding.
2. **Reframe.** As we've said above, it's often the case that the potential benefit of discussing info hazards simply doesn't outweigh the potential cost. You can often get a better cost-benefit trade-off by checking whether it's possible to instead use historical examples or a safe "black box" example, focusing on decision-relevant cruxes (vs. spiraling into spicy brainstorming).
3. **Redirect.** Sometimes, even if you're not feeling comfortable with this discussion, you won't want to flag your discomfort for fear of drawing attention to a potential info hazard (i.e., via the Streisand effect). What you might want to do in that case is see if you can move the discussion to a different topic without calling specific attention to your concern.
4. **Act later (or do nothing).** Community norms grow out of many interactions, not just a single discussion. Sometimes it's better not to do anything in the moment, but to then follow up afterwards with your facilitator, the person who was speaking, or reach out to someone senior to talk through concerns (as mentioned, Will Saunter and other BlueDot Impact staff on Slack, or Chris is happy to be a resource here). It may be better not to intervene at all; you have to pick your battles.

As we said above, there could be rare situations—usually further into a biosecurity career—where carefully discussing or disclosing an info hazard is the right call. You might need to talk through specific threats in order to make a decision about what to work on, or you might want to reach out to people who could develop countermeasures against a threat you're worried about. We talk about this a bit more in the FAQ.

FAQ

If you're grappling with any of these questions, or have any feedback on this doc, **please feel free to reach out to discuss**. You can contact Chris, anonymously or non-anonymously, through this short form. You can also reach out to Andrew Snyder-Beattie here or Tessa Alexanian (who is not a grantmaker in this space) here. If you're raising concrete infohazard concerns, please don't detail them – we can follow up more securely.

What if it's just a small group of highly-trusted people who are really careful about information security? Doesn't red teaming in that context seem worth it?
Talk to us about it! Generally we discourage people from doing this, in part because vetting people for trustworthiness is actually really hard. It can feel a bit silly to bring this kind of caution into small groups of collaborators, but recent history is riddled with examples of vetted personnel going off the rails (e.g., Aldrich Ames, allegedly Bruce Ivins) or just being reckless.

I'm doing a project in an academic lab that touches on some knowledge that I think has dual-use potential. What should I do?
We should expect most life sciences research to be dual use at some level. Chris's own past academic work is no exception. Generally speaking, you probably shouldn't panic if your project has dual-use potential comparable to the median academic life sciences thesis project. It's good to manage the misuse potential where you can, but don't drive yourself crazy stressing out about a marginal contribution to biorisk.

What about all these groups red-teaming LLMs and other AI models around misuse potential in the life sciences?
There is a lot of red-teaming of AI models for biorisk right now. The published work we're familiar with (as of March 2024) strikes us as fairly responsible from an infohazards perspective. This is an area where it's possible to causally connect red-teaming to security improvements, either through motivating policymakers during an active policy window or through directly motivating changes from model developers (e.g., red-teaming can be part of moving to a higher AI Safety Levels under Anthropic's scaling policy). It's also the case that red-teaming has focused on the potential lowering of barriers to *known* avenues for misusing biology, rather than discovering *novel* ways to cause large-scale harm with biology. Overall, these red teaming studies are conducted in a very different setting from the discussions for which this guidance is intended. We encourage people to engage on the subject of the potential AI-enabled biological risks bearing in mind the general suggestions in this document.

How will we motivate people to defend against specific threats if no one talks about them?
In some circumstances, in order to motivate defensive work, intentionally disclosing specific threats to select people will be the right thing to do. But figuring out whether to do so is a decision that really should not be taken lightly, and it's unlikely a reading or discussion group is the right venue for this. As discussed above, it may be surprisingly difficult to defend against a specific threat, or it may be that broad-spectrum approaches like metagenomic wastewater detection and excellent personal protective equipment are the best one could do anyway.

Isn't threat modeling really important for deciding what to prioritize?
We both feel like we've been able to do a fair amount of thinking through prioritization arguments of the sorts outlined in the 80,000 career profile for ourselves based mostly on historical examples of bioweapons programs and gain-of-function research, without needing to do a lot of brainstorming new hazards. Based on our experience, we'd wager that it's rare for a prioritization decision to hinge on looking at the mechanistic details of specific speculative future threats.

Won't someone else—maybe someone less responsible than me—blab about specific scary things if I don't?
Probably, at some point. But they might not do so for a very long time, and that time could be valuable. Some ideas are, in retrospect, remarkably slow to emerge – for example, Darwin's discovery of natural selection as the mechanism of evolution preceded Wallace's by about 20 years, and, upon first hearing about it, contemporary Thomas Henry Huxley exclaimed, *How extremely stupid not to have thought of that!* The bicycle may be another example of an "after-its-time" idea.

Also, you sharing the scary thing won't necessarily stop a hypothetical less responsible person from blabbing about it, and the risk of the information being misused could be cumulative.

All of this seems awful for epistemics. How can we hope to meet biosecurity challenges if our mental models are stunted by norms against exploring possibilities and openly sharing information?
Agreed, it's awful. Caution about info hazards can limit our ability to coordinate, constrain innovative thinking, and make it harder for new people to figure out what's going on. We can and should strive to get continuously better at making wise openness-secrecy tradeoffs. But given how gnarly these situations are, it's likely we'd find even theoretically optimal tradeoffs frustrating.

One thing to note is that the target audience for this guidance is people relatively new to biosecurity, in contexts like reading groups. The case for engaging with info hazards changes when somebody is in a strong position to use that information effectively and responsibly to reduce biological risk (although as written above, we think that most of the important work in biosecurity actually doesn't require much, if any, engagement with info hazards). These are tough judgment calls, though, and we think the best default norm is to refrain from exploring and discussing info hazards, especially when starting out.

I think what you've written here is paranoid, dumb, and bad for the world, and I want to tell you what I think.
Chris and Tessa would be grateful for the feedback, harsh or gentle, major or minor. In addition, the BlueDot Impact team would appreciate feedback more generally about how we handle info hazards on our courses. This is an active area where we are thinking about what norms are best, and we're not confident that we're getting things right. In fact, we're pretty sure some seasoned colleagues in the field would reasonably disagree with some of what we've said here. This doc is a work-in-progress, and feedback will be necessary for future improvement.

Several colleagues have provided very helpful feedback in writing this document, though opinions and mistakes remain our own. We'd like to gratefully acknowledge feedback from Janvi Ahuja, Dewi Erwan, Gregory Lewis, David Manheim, Nadia Mir-Montazeri, Bria Persaud, Jonas Sandbrink, Janika Schmitt, Andrew Snyder-Beattie, and James Wagstaff, among others.