Catching bugs

The Federal Select Agent Program and lessons for AI regulation

Institute for AI Policy and Strategy (IAPS)

December 15, 2023

AUTHORS

Bill Anderson-Samways - Research Analyst

Ashwin Acharya - Researcher, former

Executive summary

Governments have yet to decide how artificial intelligence (AI) should be regulated, although meaningful progress has been made in the last six months (e.g., White House, 2023, paragraph 1). **Several features of the AI risk landscape present difficulties for any regulator hoping to grapple with the technology:**

- There is a relative lack of AI expertise in government.
- Risks from AI are difficult to forecast and are characterized by high uncertainty.
- Some risks may arise during the R&D (i.e., training) phase, not just during deployment (Anderljung et al., 2023, p. 20).
- AI progress is currently extremely rapid; if it is to be effective, regulation must be designed to keep pace.

Existing regulations in areas that share some of these attributes could help us determine how to best regulate AI. Dual-use biological and life-sciences research is one such area, sharing each of these attributes to some extent.

This paper therefore examines the US Federal Select Agent Program (FSAP), which regulates entities/individuals that work with a list of dangerous pathogens and toxins (known as "Biological Select Agents and Toxins," or BSATs) (CDC, 2023: see list). Such entities/individuals must:

1. Register with (i.e., obtain a license from) the Centers for Disease Control and Prevention (CDC) or the Animal and Plant Health Inspection Service (APHIS), and
2. Submit to regulations/inspections regarding biosecurity and biosafety.

Overall, FSAP seems like it has had a positive impact. For example, since the early 2000s, FSAP has suspended/revoked at least 10 entities' licenses to use dangerous agents, suggesting that risky behaviors would occur by default and have been made less likely by FSAP.

The remainder of the report explores several more specific dimensions of FSAP that seem particularly relevant to AI regulation.

Firstly, **FSAP covers the earliest stages of the technology cycle, including R&D.** (read more)

- This has probably reduced risk: simply handling select agents (even without "deploying" them) is highly risky.

Secondly, **FSAP does not seem to delegate any core functions to third-party experts;**

most of its expertise resides in-house. (read more)

- Third-parties could usefully verify or advise on FSAP inspections. But it's unclear whether they should replace federal inspectors, who mostly have adequate expertise.

Thirdly, **FSAP appears relatively poor at handling uncertainty/tail-risks.** (read more)

- FSAP regulates a list of pathogens *known* to be dangerous at present. We argue that checklist-style systems have some generic advantages, but cannot capture the biggest biological tail-risks – namely risks from enhanced potential pandemic pathogens (ePPPs). Such risks are too uncertain for a checklist-style system to verify.
- A more "risk-based" approach (explained below) could deal with this shortcoming. Other countries (notably Canada) have introduced risk-based measures that deal with ePPPs.
- On the other hand, attempts to reform FSAP suggest that *very* high uncertainty can make formal risk-assessment difficult.

Finally, **FSAP seems not to engage in much "anticipatory" regulation.** (read more)

- FSAP does not attempt to anticipate novel risks (from ePPPs). Indeed, it was set up largely reactively, as a response to terrorist incidents involving existing pathogens.

We suggest that **all of the above holds several lessons for AI regulation.** (read more)

- FSAP and its Canadian equivalent provide a good precedent for an AI licensing regime: they require a license for the R&D phase, but only for high-risk activities.
- FSAP suggests that federal inspectors, if well-trained and advised by third parties, can do a good job of regulating complex high-risk technical domains – although replicating that approach may be difficult for AI, as governments have struggled to hire AI talent (e.g., see Berglind, Fadia, and Isherwood, 2022, paragraph 4).
- Similar to ePPPs, AI entails highly uncertain tail-risks. Thus, AI regulations should almost certainly include some strong "risk-based" component.
- However, given that risk assessment in highly uncertain domains is very difficult, it may be necessary to complement risk assessments with slowing/pausing AI development or investing in safety, until more reliable risk assessment measures are available. Using multiple complementary risk management frameworks can also help ensure coverage of tail risks (Ee et. al., 2023, 8-9).
- Checklist-style systems *are* good at addressing known threats with clear control measures. Checklists should thus be combined with risk-based systems addressing less certain threats. The Canadian equivalent to FSAP seems like a good example of a system that combines checklists and risk-based measures.

Table of contents

Editorial note

This report was produced by IAPS in December 2023. It contains a mixture of information about FSAP's operations, evaluative claims about FSAP's success, and our tentative normative claims about lessons for AI. We distinguish the descriptive information and evaluative claims about FSAP in the text where possible.

The evaluative claims about FSAP are our own judgements, based on all the evidence that we could find in the time available for this research project. **The nature of that evidence varies from point to point, but broadly speaking, it is always one or both of the following:**

1. Data that provides some evidence of FSAP's effectiveness, whether or not the authors of said data intended it as evidence of FSAP's effectiveness (e.g., the number of times FSAP has suspended or revoked an entity's registration with the program).

2. Expert judgements of FSAP's effectiveness, which we gathered from:
    a. The secondary literature.
    b. Interviews with three former CDC officials with years or decades of experience working with the Select Agent Program as inspectors and with training as microbiologists.
        i. One of these interviewees also had experience as a senior official in a high-containment laboratory subject to inspection by FSAP.
        ii. (Two of the experts also reviewed this report to verify its accuracy).

It's worth noting that some of the "lessons for AI" detailed in this report are not supported with reference to evidence from FSAP itself:

- Instead, our "lessons" sometimes simply consist of proposals for AI regulation which AI experts already had decently strong reason to believe might be useful, but for which FSAP nonetheless provides an interesting model (e.g., a licensing regime).

- And sometimes they are more abstract theoretical lessons about regulation *in general* that we came across when conducting this case-study, but which incidentally seem relevant to both FSAP and AI regulation (e.g., our section on the generic benefits of checklist-style systems).

Detailed summary

Summary of FSAP and overall evaluation

The first section of this report briefly describes how FSAP works, and evaluates its overall effectiveness. (read more)

Description:

FSAP regulates most entities and individuals that possess, use, or transfer "select agents" – a list of dangerous pathogens and toxins (Anderljung et al., 2023, p. 21). [1] FSAP is jointly administered by the Centers for Disease Control and Prevention (CDC) and the Animal and Plant Health Inspection Service (APHIS). By law, entities and individuals working with select agents must register with CDC or APHIS (to work with the *specific* select agents they wish to work with). This requires background checks by the FBI and an initial inspection by CDC or APHIS to ensure compliance with the Select Agent Regulations (Brooks, 2018, p. 4).

The regulations themselves require, among other things:

- A biosecurity plan to guard against theft, loss, or unauthorized access (Brooks, 2018, p. 6).
- A biosafety plan to guard against infection of personnel (Brooks, 2018, p. 6). Entities must also adhere to checklists of biosafety standards based on the Biosafety in Microbiological and Biomedical Laboratories (BMBL) handbook (Kirkpatrick et al., 2018, p. 12).
- An incident response plan to deal with biosafety or biosecurity incidents in the event that they do occur. (That information was drawn from our interviews).

Further inspections take place to:

- Approve/disapprove renewal of registration every three years;
- Approve/disapprove updating the details of the registration, e.g., to allow entities to work with additional agents for which they are not currently registered (GAO, 2017, p. 24);
- Spot-check compliance with the regulations (often via unannounced inspections);
- Investigate serious breaches, e.g., loss/release.

[1] There are certain exemptions, detailed in Appendix 1.

Evaluation:

Overall, FSAP seems to have had a positive impact.

Prior to FSAP, there were no nationwide US regulations on entities using select agents, which are inherently dangerous pathogens. Moreover, **as of 2016, FSAP had suspended or revoked ~10 entities' ability to use select agents after they violated FSAP regulations** (GAO, 2017, p. 36). For context, there were only ~300 entities registered with FSAP in 2016 (GAO, 2017, p. 11). Given the significant proportion of labs that have had their registrations suspended, revoked, or even denied in the first place,[2] **our interviewees believed that FSAP's registration (licensing) process specifically is very effective.**

FSAP's performance on our specific dimensions of interest

The remainder of the report then examines several more specific aspects of FSAP that seem particularly relevant to possible AI regulations.

Regulation of development (as well as deployment)

Firstly, we discuss how FSAP regulations – unlike many other federal regulations – apply to the development (R&D) stage of technological development, as well as the "deployment" stage. (read more)

Description:

FSAP regulations essentially cover all work involving the possession or use of select agents, including R&D (Anderljung et al., 2023, p. 21). FSAP also prohibits certain specific research *experiments* done with select agents, e.g., those that enhance resistance to drugs (unless approval is granted by NIH as well as either the head of the Department of Health and Human Services[3] or the head of the US Department of Agriculture[4]) (Kirkpatrick et al., 2018, p. 12).

Evaluation:

It seems positive that the regulations cover development as well as deployment.

[2] Apparently some labs never pass the initial registration process, according to our interviewees.
[3] Or a designee of theirs.
[4] Or a designee of theirs.

Two of our interviewees noted that regulating R&D in this case is obviously necessary; simply handling select agents (not just deploying them) involves inherent risks – both risks from spillover events, and risks from biosecurity breaches (e.g., exfiltration). It is worth noting that interest in using select agents during R&D has only risen since the 2000s, when FSAP was set up (Finlay, 2010, p. 3).

Moreover, as of late 2021, FSAP had also prevented ~10 experiments that fell under its potentially hazardous experiments category (Smith et al., 2023, p. 1), which provides some evidence of FSAP's counterfactual impact in that specific area.

It's plausible that FSAP regulations on bioengineering R&D have hindered effective biodefense (Berger, 2011: see abstract), but there is only limited and mixed evidence on that claim (Morse, 2015, p. 7).

Level of delegation to third-party auditors and experts

Secondly, we note that FSAP does not appear to delegate any analysis to third-party auditors/experts. (read more)

Description:

FSAP does not appear to delegate any of its inspections of biolabs, or analysis of select agents, to third-party auditors or experts.

Evaluation:

There is some evidence that bringing in third parties could play a useful role in giving scientific advice to inspectors and/or testing select agent material to check whether it is dangerous.

For example, the Department of Health and Human Services (HHS) recommended considering the viability of third-party testing of select agent material (HHS, 2015, p. 7, our emphasis), though their reasoning for that recommendation was somewhat unclear.

Moreover, the US Government Accountability Office (GAO) suggested that FSAP could learn from countries that provide more external expert advice to their inspectors, e.g., the UK (GAO, 2017, p. 48). One of our interviewees agreed, on the basis that inspectors often lack the technical or practical expertise to keep up with the very latest developments in research utilizing select agents.

However, it is unclear whether third-parties should *replace* federal inspectors. That is because federal inspectors themselves usually have pretty strong technical knowledge, although more continuous training should be provided to federal inspectors so that they can keep up with scientific developments (GAO, 2017, p. 23).

Handling of uncertainty and tail-risk assessment

The third dimension we examine is FSAP's handling of uncertainty and tail-risks. (read more)

Description:

FSAP takes a tiered approach to risks from select agents – namely separating out especially dangerous "tier 1" select agents from the rest of the agents on the list (GAO, 2017, p. 13-14). **Nonetheless, regulations *within* each tier take the form of standardized "checklists," rather than adopting a "risk-based" approach** (i.e., assessing risk in each specific situation at hand) (Burnett et al., 2016, p. 39).

Evaluation:

There are some generic advantages to checklist-type systems – notably their ability to address hazards with well-understood threat levels and containment systems. They also appear to possess advantages such as clarity, less requirement for technical expertise among government auditors (Burnett et al., 2016, p. 53), an ability to handle complex tasks where even well-trained experts may miss key steps (Powell, Jain, & Juneja, 2019: see "Comment" section), and *possibly* less opportunity for regulated entities to "game" inspections (although in fact the evidence is very ambiguous here).[5]

However, overall, FSAP's checklist-style system seems to make it relatively poor at handling uncertainty/tail risks.

In particular, FSAP seems to be poor at dealing with uncertain biological tail risks (Global Catastrophic Biological Risks or "GCBRs"), which seem especially similar to AI risk. FSAP

[5] On the one hand, checklists seem less gameable than non-checklist systems, as they provide a standardized and clear set of guidelines that arguably leaves less ambiguity for companies to exploit.

On the other hand, checklists may be *more* gameable than non-checklist systems, if the outcome of having such checklists is simply that regulated entities "know what's coming" in terms of audits and can hence paper over safety faults to address the checklist. Ways of mitigating that possibility include keeping the checklist non-public, updating the checklist frequently, and making heavy use of unannounced inspections.

standards cover only a checklist of pathogens that are already known to be dangerous, but not pathogens that could be engineered to become more dangerous (Lewis et al, 2019, p. 979). Such engineered novel pathogens pose the biggest GCBRs (Inglesby, 2018).[6] Tier- and checklist-style standards work best when risk categories can be easily distinguished based on standardized criteria. By contrast, they are poorly suited to mitigate GCBRs, which are *high-uncertainty* – even informed experts will disagree about the level of risk that would be created by a particular pathogen that could be bioengineered but does not yet exist (Epstein, 2023).[7]

There have been various proposals to update FSAP to make it more "risk-based," in some cases explicitly in order to deal with tail-risks from bioengineered agents. However, these proposals have not been put into effect. The reason for that remains very unclear (but could *plausibly* have to do with, e.g., a lack of requisite expertise to conduct risk-assessments among CDC inspectors, and/or CDC fearing that inadequate risk assessments would leave them open to being held liable in the event of a failure).

Other countries, such as the UK and Canada, appear to have superior systems for dealing with uncertainty/tail-risks (GAO, 2017, p. 47). The Canadian system seems particularly good (although it is still likely far from perfect) (Pannu et al. 2022, paragraph 4).

- It requires all entities conducting potentially dangerous pathogen research – including (unlike FSAP) research involving novel bioengineered pathogens – to obtain a license from the government (PHA Canada, 2018).[8]
- To obtain such a license, entities must conduct a risk assessment and come up with a plan for controlling any risks identified, inclusive of the planning, implementation, and use/dissemination phases of the research (Government of Canada, 2023a, paragraph 1; PHA Canada, 2018: see section 3.1. entitled "Identify research with dual-use potential").
- The Canadian system also does not sacrifice the benefits of checklist-style systems – it does include a list of pathogens *known* to be dangerous, which are subject to particularly stringent regulation. But Canadian regulations are not restricted to such pathogens.

It's worth noting that the United States does have *some* policies regulating risks from novel bioengineered agents, notably its DURC/PC3O policy, alongside NIH guidelines on rDNA research. However, these policies are limited in scope; they only apply to entities which receive federal funding, while FSAP applies to all entities using select agents.

[6] See the section after Wiblin's question "So what kind of levers are actually available to reduce the chance that biotechnology goes on to cause harm?"
[7] See the section "Utilizing existing legislation." Our emphasis.
[8] See section 2.1.1. entitled "Pathogen and toxin regulation in Canada"

Level of anticipatory regulation

Finally, we examine whether FSAP has engaged in "anticipatory" regulation, and how that might have impacted its effectiveness. (read more)

Description:

We could find no evidence that FSAP engages in anticipatory regulation-setting. Instead, FSAP regulations seem largely *reactive*, set up in response to terrorist incidents, and responding to threats from known biological agents (Kirkpatrick et al., 2018, p. 25).

Evaluation:

Our overall impression is that **the lack of anticipatory regulation is a major failing of FSAP, as its more "reactive" model seems unsuited for an area subject to such rapid technological change (**Kirkpatrick et al., 2018, p. 25**).** AI is one other such area, implying that an "anticipatory" approach will also be needed there.

Concrete lessons from FSAP for the AI case

Positive lessons

FSAP potentially provides a good precedent for an AI licensing regime

FSAP (and the Canadian biosafety regime) provide a good precedent for an AI licensing regime in some respects. They require a license (and adherence to high regulatory standards) for the R&D phase as well as the deployment phase, a process that seems quite effective at containing risk. They also attempt to ensure that those regulations do not burden entities undertaking lower-risk activities.

Federal inspectors can have the expertise to audit frontier technology companies

FSAP suggests that the federal government is capable of hiring inspectors with relevant expertise to audit frontier technology companies, as long as those inspectors receive good

ongoing training and receive regular third-party input into/auditing of their inspections.[9] However, hiring experts might be much harder for AI than for biological research, as governments have struggled to hire top talent in the AI domain (e.g., see Berglind, Fadia, and Isherwood, 2022, paragraph 4).

Checklist-style systems have their uses

Checklist-style regulations have *some* advantages (although see the major drawbacks noted below), meaning that **an ideal AI regulatory regime could combine *some* well-calibrated checklist-style standards** (e.g., evaluating models on the Machiavelli benchmark for antisocial behavior) **with more risk-based standards.**

Most notably, some hazards are very clearly severe – select agents and toxins, as previously noted, represent *known* biohazards. Moreover, they have well-known control measures. In such cases, it seems obviously reasonable to mandate such control measures through checklists. Similarly, well-made checklists enable auditors to automatically filter out many known *lower*-risk systems. That increases the efficiency and therefore the effectiveness of the auditors.

Furthermore, a regime purely based on risk-based standards (without any kind of checklist system) could rely excessively on inspectors having world-class levels of technical training (which seems by no means guaranteed). And there is evidence demonstrating that checklists can be useful even in highly-skilled professions, as some tasks are so complicated that even the most experienced expert will miss key steps.

Finally, a purely risk-based system *might* be easier for regulated entities to "game" (although in fact the evidence is ambiguous here – see above). To mitigate the risk of regulated entities "gaming" such checklists, **an ideal regulatory regime could:**

(a) **Keep the checklist private from regulated entities,**[10] or
(b) **Update the checklist frequently,**[11] so that the regulated entities do not "know what's coming" from the last inspection.

[9] Although, on the other hand, perhaps it's easier for the US government to hire experts in the life sciences than to hire AI experts, given that the life sciences are closely tied to public funding, the government runs some important life science institutions, and such institutions are probably closer in terms of average salary to most regulatory jobs. It would be interesting to investigate this further.

[10] If the checklist is intended for use by external auditors (as with a checklist of standardized model evaluations), rather than by the companies themselves.

[11] (As with the above footnote).

(c) (Possibly) rely somewhat heavily on unannounced inspections (although this may be less relevant to the AI case, as the regulated entity's ability to convincingly "teach [models] to the test" may be orthogonal to inspection timing).
(d) Always combine checklists with risk-based regulations.

Negative lessons

AI regulation should be (at least partially) risk-based

Checklist-style standards may be particularly bad at handling risks that are *extreme, uncertain, and rapidly advancing* [12] – all of which apply in the AI case. **Thus, AI regulations should almost certainly include some strong risk-based component** that allows regulators to assess and mitigate risk based on their own judgment of the specific circumstances at hand in a given case.

But risk-based regulation can be difficult where there is high uncertainty

On the other hand, a close examination of past proposed changes to FSAP reveals that developing good risk-based approval systems for domains subject to very high uncertainty (like AI risk and extreme biorisk) is difficult. Similarly, it may be very difficult to predict AI model capabilities with the degree of certainty necessary for risk-based regulations to work. (As an extreme example, if the distribution of expert opinion over the likelihood of a particular capability manifesting itself was near-Knightian, regulations could hardly be calibrated to a given risk level.)

In that case, the right approach may be to combine risk-based regulation with other policies – such as:

- Checklist-style standards (see above point on checklists as a complement to risk-based standards),
- Slowing/pausing AI development (if there is insufficient evidence as to whether the risk is at an acceptable level),
- Adopting "defense-in-depth" approaches, which assemble multiple *independent* layers of safety and security measures, so that no single measure is exclusively relied upon. Defense-in-depth strategies mitigate highly uncertain risks by providing "buffer room" and reducing the likelihood of unanticipated and correlated failures kicking out all defenses at once (Ee et. al., 2023, p. 8-9).

[12] Because in that case, the checklist may become outdated.

- Investing more heavily in AI safety (although, as in biodefense, some such research is dual-use and could thus speed up dangerous capabilities (see Guest et al., 2023, p. 2-3)).

Overview of FSAP and its effectiveness

Broad description of FSAP

The Federal Select Agent Program (FSAP) regulates the activities of individuals and entities that possess, use, or transfer certain biological agents known to be highly risky (Anderljung et al., 2023, p. 21). [13] (That is, unless the individual or entity in question has been lawfully exempted – see Appendix 1 for more info). The agents on the list are termed "Select Agents"; there are more than 80 of them (Brooks, 2018, p. 4).

Agencies involved in FSAP include:

- The Centers for Disease Control and Prevention (CDC) in the Department of Health and Human Services (HHS), which has responsibility for regulating select agents that pose a direct risk to humans (Brooks, 2018, p. 4).
- The Animal and Plant Health Inspection Service (APHIS) in the US Department of Agriculture (USDA), which has responsibility for regulating select agents that pose a risk to agriculture (Brooks, 2018, p. 4).
- (To a lesser extent) The Federal Bureau of Investigation (FBI) in the Department of Justice (DOJ), which undertakes background checks on the individuals/entities working with select agents, as well as the Responsible Officials (ROs) in each entity (see below in this section) (Brooks, 2018, p. 4).

Individuals and entities working with select agents must register with the relevant government entity – either CDC or APHIS (Kirkpatrick et al., 2018, p. 12). Unless the entity/individual has received a certificate of registration from CDC/APHIS, they may not possess, use, or transfer any select agent (Code of Federal Regulations, 2023).[14] Entities/individuals register to work with specific select agents; if they later wish to work with additional select agents (for which they are not currently registered), they must obtain registration for that too (Code of Federal Regulations, 2023).[15]

[13] It's worth noting that FSAP typically does not undertake DOD laboratory inspections; those are handled by the Department of Army Inspector General and Army MEDCOM (thanks to Emma Williamson for that point). Indeed, "The Department of Defense has imposed a more stringent layer of regulation called biological surety (biosurety) on top of the requirements of 42 CFR 73 [the select agent regulations]" (Pastel et al., 2006: see "Summary" section).

[14] See the section "Registration and related security risk assessments"
[15] See the section "Registration and related security risk assessments"

During the registration process, the entity must designate a "Responsible Official" – i.e., someone within the entity responsible for ensuring FSAP compliance (Brooks, 2018, p. 5). They must also submit to a registration inspection by CDC/APHIS and a "Security Risk Assessment" by the FBI (see below) (Brooks, 2018, p. 4), and they must possess rigorous training specific to the research they are undertaking (Brooks, 2018, p. 6).

Registration with FSAP requires individuals and entities to comply with several sets of rules – for example:

- *Developing a security plan* to guard against theft of, loss of, or access by unauthorized persons to, select agents (Anderljung et al., 2023, p. 21).
- *Developing a biosafety plan* to guard against accidental release (Brooks, 2018, p. 6).
- *Adhering to FSAP biosafety checklists* based on the Biosafety in Microbiological and Biomedical Laboratories (BMBL) standards (Kirkpatrick et al., 2018: 12). BMBL classifies laboratories into different biosafety levels, e.g., BSL-4, BSL-3, etc., and specifies codes of practice for each level (Kirkpatrick et al., 2018, p. 12). CDC and APHIS have the power to check for compliance with their BMBL-based checklists and levy penalties for noncompliance (Kirkpatrick et al., 2018, p. 12).
- *Developing incident response plans* to deal with biosafety/biosecurity incidents in the event that they do occur (Brooks, 2018, p. 6).
- *Undertaking inventory checks* on (e.g., vials of) select agents, and conducting inventory audits under certain circumstances (Brooks, 2018, p. 6).
- *Complying with reporting requirements* in the event of theft, loss, or release (Kirkpatrick et al., 2018, p. 12).
- *Getting permission to undertake certain experiments* with select agents, from the National Institutes of Health (NIH) as well as the secretary of HHS or USDA (see the section "Regulation of development (as well as deployment)" for more info) (Kirkpatrick et al., 2018: 12).

Registered entities must also perform Security Risk Assessments (SRAs) to guard against misuse (Anderljung et al., 2023, p. 21). Each entity and individual working with select agents, as well as the Responsible Official (RO) within each entity, must submit to an SRA by the FBI (Brooks, 2018, p. 4). The SRA is basically a thorough background check. However, it's worth noting that "federal, state, or local governmental agenc[ies] or accredited public academic institutions… are not required to undergo an SRA because they are presumed to have similar security measures already" (Brooks, 2018, p. 5).

Finally, registered entities must submit to CDC/APHIS inspections that ensure compliance with the above regulations (Anderljung et al., 2023, p. 21).

- The most common and comprehensive form of inspection is the *new registration/three-year renewal inspection*. In order to initially obtain registration,

entities must submit to an inspection verifying compliance with the above regulations. Similar inspections take place every three years to determine whether to renew the entity's registration (GAO, 2017, p. 24).

- *Verification inspections* occur at least once between each renewal inspection. Such inspections are often unannounced and aim to spot-check entities' compliance with the above regulations (GAO, 2017: 24).
- When entities want to significantly amend their select agent registration, e.g., working with additional select agents with which they are not yet registered to work, they must submit to a *registration amendment inspection* (GAO, 2017, p. 24).
- *Other inspections* must take place following loss/release, occupational exposure to a select agent, or serious previously noted deficiencies that entities have not addressed. These can be announced or unannounced, and serious violations can result in referral to HHS/APHIS enforcement services or the FBI (GAO, 2017, p. 24).

In general, to ensure compliance with FSAP regulations, inspections review the "physical safety and security components of the facility, examine the documentation available, and interview laboratory personnel to collect information used to complete the checklists" (Brooks, 2018, p. 6) If an entity violates the regulations, they could have their registration revoked or suspended, have their applications to conduct research denied, or face civil or criminal penalties (Brooks, 2018, p. 6).

Evaluation of overall effectiveness

Overall, FSAP's existence seems net-positive. A counterfactual world in which FSAP did not exist would probably be a less safe world.

Before FSAP, there were no national regulations on using these inherently dangerous agents. Our interviewees believed that that had been a major problem. For example, two of our interviewees had the impression that, although there were various ad-hoc and piecemeal restrictions on *using* certain agents, there were no regulations on the *transfer* of these agents – for example, there was no system of permissions to prevent a "bad actor" from acquiring them.[16]

In particular, FSAP's registration process seems quite effective. That seems relevant to AI regulation, as many experts believe that a licensing regime may be necessary for AI. Such a regime would require frontier AI developers to possess a license for said development, which can be revoked if the developer fails to ensure safety (Anderljung et al., 2023, p. 20).

[16] One of our interviewees recounted an anecdote she had heard about a pathogen that we would now designate a select agent being carried on a commercial airline! We are uncertain how accurate that anecdote is though.

FSAP's registration process is essentially a form of licensing. CDC/APHIS grant (or decide not to grant) a registration certificate to a regulated entity once they are satisfied that the entity meets FSAP's safety criteria; and they can revoke this registration at any point, if their inspections (or other evidence) indicate that the entity has violated FSAP's safety criteria (Code of Federal Regulations, 2023[17]; Code of Federal Regulations, 2023[18]). FSAP regulations state that "Upon revocation or suspension of a certificate of registration, the individual or entity must… Immediately stop all use of each select agent or toxin covered by the revocation or suspension order"(Code of Federal Regulations, 2023[19]).

From 2003 through 2016, FSAP suspended or revoked 10 entities' registrations in response to serious violations of the select agent regulations (GAO, 2017, p. 36**).** There were only ~300 entities registered with FSAP in 2016 (GAO, 2017, p. 11), so that seems a reasonably high proportion. Given that (as noted above) there was no national regulation of select agents before FSAP, the counterfactual impact of FSAP regulation (and specifically its registration process) therefore seems fairly good.

When we asked interviewees whether they thought that the registration/deregistration process specifically was effective, all three responded that it was effective, although one noted that it could be a long/arduous process and that that could hinder biodefense research.

Our first interviewee, a former FSAP inspector, noted that there are not many entities that *want* to conduct research with select agents, and even then FSAP is not just "giving out" registration to anyone who wants to work with such agents. In some cases, entities never get approved for registration. She thought that the amount of registration certificates (i.e., licenses to operate) given to entities was about right to ensure biosafety and biosecurity.

Our second interviewee, also a former FSAP inspector, agreed. She noted that where she had been aware of serious issues at maximum containment entities, these had all been shut down (i.e., had their registrations revoked). She thought that the registration process was very comprehensive.

Our third interviewee, a former CDC official currently in a senior position at a high-containment laboratory (itself subject to FSAP inspections), said that she thought the registration process was both effective and necessary. However, she noted that the length of the registration process could be quite arduous – one of her current principal investigators (researchers) was currently applying to work with select agents, and was looking at a year's wait before getting registered. She noted that that could hinder biodefense efforts necessary to deal with pathogens such as COVID-19. (See below for our thoughts on that).

[17] See the section "Registration and related security risk assessments"
[18] See the section "Denial, revocation, or suspension of registration"
[19] See the section "Denial, revocation, or suspension of registration"

The fact that FSAP's registration process is effective implies that an AI licensing regime could indeed work, and could even be modeled on FSAP's regime in certain respects. (read more)

Regulation of development (as well as deployment)

Brief description of relevant FSAP procedures

Anderljung et al. (2023) specifically identify FSAP as an (unusual) example of regulations that apply to the development (R&D) phase as well as the deployment phase (Anderljung et al., 2023, p. 21)**. Such regulations seem especially relevant to AI (**Anderljung et al., 2023, p. 21). AI models with dangerous capabilities can be stolen or leaked during the R&D stage (Anderljung et al., 2023, p. 20). More speculatively, risks from misaligned AI systems may arise during the training process (Ngo, Chan and Mindermann, 2023, p. 8-11).

As previously mentioned, FSAP regulations cover all work involving select agents – *including* R&D (except where an entity has explicitly gained exemption from such regulations – see Appendix 1) (Code of Federal Regulations, 2023[20]) **FSAP also prohibits certain experiments done with select agents:** "An individual or entity may not conduct, or possess products resulting from, the following experiments unless approved by and conducted in accordance with the conditions prescribed by the HHS Secretary:

> "(1) Experiments that involve the deliberate transfer of, or selection for, a drug resistance trait to select agents that are not known to acquire the trait naturally, if such acquisition could compromise the control of disease agents in humans, veterinary medicine, or agriculture.
>
> "(2) Experiments involving the deliberate formation of synthetic or recombinant DNA containing genes for the biosynthesis of select toxins lethal for vertebrates at an LD[50] <100 ng/kg body weight.
>
> "(3) Experiments that involve the creation of SARS–CoV/SARS–CoV–2 chimeric viruses resulting from any deliberate manipulation of SARS–CoV–2 to incorporate nucleic acids coding for SARS–CoV virulence factors or vice versa" (Code of Federal Regulations, 2023[21]).

Evaluation of effectiveness

Overall, it seems like such R&D phase regulations have been fairly effective.

[20] See the section "Registration and related security risk assessments"
[21] See the section "Restricted experiments"

FSAP's broad restrictions on the R&D process seem to be effective at containing risk

The counterfactual impact of FSAP's general restrictions on R&D with select agents seems strong.

The 2000s saw growing interest in R&D using select agents (Finlay, 2010, p. 3). And all three of the experts that we interviewed believed that select agents needed to be regulated during the R&D stage and that FSAP regulation was effective at doing so – including our third interviewee, herself a senior official at a high-containment laboratory *subject* to inspection.

Our first interviewee, a former FSAP inspector, noted that **handling select agents is inherently dangerous, and thus *all* parts of the process should be regulated, which seems intuitive.** Our third interviewee (the senior official) made a similar point. In most other domains subject to federal regulation – for example automobiles or financial markets – risks come from the *product* of the research (e.g., cars crashing, faulty financial products). In this case however, it is a *component* of the product, namely the select agent itself, which entails risk. So the R&D phase, which necessarily employs those components, needs to be regulated.

FSAP restrictions on certain kinds of experimentation also seem counterfactually effective

The counterfactual impact of specific restrictions on *experimentation* also seems quite good. For example, "[f]rom January 2014 to December 2021, DSAT received 113 requests to conduct potential restricted experiments [with select agents]... Of the 20 requests that met the definition of a restricted experiment, 8 were denied because the experiments had the potential to compromise disease control in humans" (Smith et al., 2023, p. 1). That is quite a high rate of denial, suggesting reasonable counterfactual effectiveness.

On the other hand, the small number of total requests does suggest "that restricted experimental research with select agents and toxins is [only] being proposed or conducted by a very small subset (n=9, 3%) of the total 276 of entities that are registered with the select agent program" (Smith et al., 2023, p. 6).

It seems plausible that FSAP hinders effective biodefense R&D, but the evidence is contradictory

Berger (2011) speculates that **"restrictions on access to select agents… [could] potentially could hinder *defensive* research and development"** (Berger, 2011: see abstract). That seems at least *plausible* based on data we have seen elsewhere – for example, "over 20% of select agent researchers surveyed in 2004 and 2005 noted that the regulations were affecting their ability to collaborate domestically and internationally, and about 40% claimed that they had to use research funding to make required security upgrades" (Morse, 2015, p. 7).

Two of our interviewees themselves expressed concerns that FSAP makes it harder for scientists to conduct biodefense research. One, a former FSAP inspector, believed that most researchers shy away from pursuing research with select agents due to the regulatory burden. As an anecdote, she noted that at her current facility one professor of infectious diseases had told her that he did not want to work with select agents for that reason.

Our third interviewee, herself a senior official at a high-containment laboratory, believed that red tape from CDC regulations (albeit not FSAP) had made it harder to conduct necessary biodefense research against COVID-19 during the pandemic. She argued that, for that reason, COVID-19 should not be placed on the FSAP list.

However, Morse (2015), in a comprehensive literature review, notes that the evidence on whether FSAP has stifled innovation is contradictory, and that further analysis is needed in order to verify such claims (Morse, 2015, p. 7).

(It's also worth noting that FSAP regulations do not apply to research and development into *novel* pathogens, as outlined below).

Level of delegation to third-party auditors and experts

Brief description of relevant FSAP procedures

FSAP does not delegate any of its inspection functions or analysis to third-party auditors or experts. We could find no evidence suggesting that it does any such delegation. Moreover, that was the impression of all three of our expert interviewees. And in 2015 the US Department of Health and Human Services (HHS) recommended that "Federal inspectors' ability to conduct effective and meaningful oversight *should be strengthened by… considering the viability of independent third-party testing of select agent or toxin material*," implying that third party *testing* at least was at least not in place at the time of writing (HHS, 2015, p. 7, our emphasis).

It is worth noting that the Responsible Official's (RO's) role is to "ensure the entity is compliant with all [Select Agent regulations]" in the absence of federal inspectors (CDC 2020a: see bullet-point list). Unique responsibilities of the RO in this regard include:

- Writing and implementing biosecurity plans and biosafety plans for the laboratory.
- Writing and implementing an incident response plan for the laboratory.
- Conducting drills or other exercises to test the effectiveness of the above plans, at least once a year.
- Reviewing (and revising where necessary) said plans at least once a year.
- Providing training on biosecurity, biosafety, and incident response (CDC 2020b, p. 6).

Thus, FSAP does not delegate any of its work to third-parties, but it does "delegate" some responsibility for ensuring compliance to ROs in the regulated organizations.

Evaluation of effectiveness

More advice and verification from third-party experts could improve FSAP

Overall, there is *some* indication that bringing in third parties could play a useful role in giving scientific advice to inspectors and/or testing select agent material to verify whether or not it is hazardous.

The above-mentioned HHS quote implies that HHS thought that independent testing by third-party experts would be a good thing. We could find no evidence as to *why* HHS made that recommendation. However, the only actual analysis in HHS's piece is limited to a discussion of well-known failures of FSAP regulations, namely the US Army facility Dugway's shipping of anthrax and botulinum toxins at above-threshold concentrations (HHS, 2015, p. 4-5). In the anthrax case, Dugway's internal testing showed that above-threshold concentrations were still present despite attempts to bring concentrations below the threshold. Yet Dugway ignored the results of the internal tests. Perhaps, therefore, HHS's view was that Dugway would have felt less able to ignore the results of a third-party test than its own internal tests.

Other countries draw more on external expertise for *advice*, and the US Government Accountability Office (GAO) seems to think FSAP could learn from that. For example, the GAO notes that, in the UK, "regulators at the Health and Safety Executive have access to external expert advisory committees to advise on issues related to new or emerging pathogens, diseases, or other scientific issues that inspectors may encounter during inspections or when developing policy" (GAO, 2017, p. 48). (Although one of our interviewees, a former FSAP inspector, noted that CDC DSAT (Division of Select Agents and Toxins) officials do periodically hold meetings to review the regulations – for example, to add or remove select agents from the list – and was under the impression that DSAT does consult external experts during that process).

Our third interviewee (a senior official at a high-containment laboratory) agreed that third-party advice to inspectors/verification of their findings could be useful. Having worked as both an inspector and a practitioner subject to inspection, she said that "once you join CDC, you lose the hands-on component – becoming more of an administrator, and less of a researcher." She noted that CDC inspectors typically have relevant degrees (something we elaborate further below), but are sometimes criticized for lacking a deep enough research background, and themselves want support in keeping up-to-date with cutting-edge technology. It therefore seems that expert advisors could provide useful input on novel research areas in which inspectors might lack expertise. To some degree, Responsible Officials (ROs) play this role; our interviewee noted with approval that DSAT recommends that the RO should be someone who has knowledge of the day-to-day operations of the entity, and ideally is someone doing active research at the entity.

It is unclear whether third parties should *replace* federal inspectors, who typically have a reasonable amount of technical expertise

It remains somewhat unclear whether third-parties should replace federal inspectors**.** It seems like the main takeaway is instead that **third-parties should play a verification/advisory function to *support* existing federal inspections.**

Our reasoning here is that it seems like **FSAP inspectors themselves have decent technical knowledge, although there is room for improvement in that respect.**

GAO interviewed various laboratory personnel. Most said that FSAP inspectors generally have appropriate expertise (GAO, 2017, p. 31). Moreover, FSAP inspectors overwhelmingly have relevant advanced degrees, including many with PhDs in microbiology or related fields (GAO, 2017, p. 31).

Nonetheless, CDC and APHIS internal reviews (from 2015 and 2016, respectively) did identify some shortcomings. For example, inspectors' skill levels and approaches to inspection were quite variable, and many inspectors reported that they themselves were in need of additional training opportunities to keep up with advancements in the field (GAO, 2017: 31).

One of our interviewees expressed similar concerns. She believed that some inspectors simply know a lot about one specific area of bioscience, and focus too much on that during the inspections. She argued that that may be a disadvantage if FSAP were ever to adopt the more risk-based approach that we recommend below, which in her view requires more holistic knowledge. She also noted that her experience with the nationwide training facility for FSAP inspectors was "not the best." In her words, "Some inspectors come straight out of university,[22] get five days of training, and become inspectors." She argued that there should be more careful selection of the inspectors, as well as training that is more consistent with the tasks they need to undertake. **That suggests that any AI regulator that utilizes federal inspections should provide generous funding for continuous training for inspectors.**

[22] Albeit, we would note, with a relevant degree, if the above-cited information from the GAO report is to be believed.

Handling of uncertainty and tail-risk assessment

Brief description of relevant FSAP procedures

FSAP takes a tiered approach to risks from biological select agents and toxins (BSATs). In 2010, a presidential executive order directed HHS and USDA to tier the agents on the list by their degree of riskiness (GAO, 2017, p. 13-14). As of December 2016, ~50% of FSAP-registered entities were registered to work with tier 1 agents (the riskiest class of agents) (GAO, 2017, p. 13-14).

Within each tier, FSAP follows a "checklist" approach, recommending a standard set of procedures that each regulated entity should follow. This contrasts with "risk-based" approaches, where regulators undertake risk-assessments and then set requirements based on the specific risks found in a given case (Burnett et al., 2016, p. 39). For example, the UK Health and Safety Executive (HSE) "prioritizes which laboratories to inspect during the year by assessing the level of risk a specific laboratory or program may have on worker or public health and safety or the environment" (GAO, 2017, p. 44). The FSAP tiers do not really do that – they simply create standardized checklists for entities handling different groups of select agents (Burnett et al., 2016, p. 39).[23]

Evaluation of effectiveness

Overall, it seems that FSAP is quite poor at handling uncertainty/tail risks:

- Its reliance on a predetermined "checklist" of pathogens and corresponding checklist of containment actions does seem useful for addressing threats with known risk levels and control measures. It also seems useful in terms of clarity, accessibility, and dealing with complex multi-step tasks, and *possibly* is harder to "game" than a more flexible risk-based system (although the evidence is ambiguous here).

[23] So for example, FSAP's checklist for "incident response" states that "Entities with Tier 1 select agents and toxins must have the following additional incident response policies or procedures…

- "The incident response plan must fully describe the entity's response procedures for failure of intrusion detection or alarm system…
- "The incident response plan must describe procedures for how the entity will notify the appropriate Federal, State, or local law enforcement agencies of suspicious activity that may be criminal in nature and related to the entity, its personnel, or its select agents or toxins." (CDC, 2023, p. 3)

- Overall however,the focus on a list of pathogens currently ***known*** to be dangerous means that FSAP does not address the largest biological tail risks, i.e., risks from enhanced Potential Pandemic Pathogens (ePPPs).
- The "checklist" approach seems particularly poorly suited to dealing with risks from ePPPs as they are subject to high uncertainty – even-well informed experts might disagree on whether a given pathogen has the potential to become a high-risk ePPP. In such cases, using a binary "yes-no" criterion does not make much sense.
- Proposals to update FSAP to make it more "risk-based" and less "checklist-based" have largely not borne fruit.

FSAP's system does possess some advantages

Addressing known threats

Perhaps most obviously, FSAP's system seems well-adapted to its present purpose – regulating a list of pathogens and toxins where the risks are *known* to be reasonably high, and where known control measures can be turned into a corresponding checklist of actions. Intuitively, it seems both inefficient and potentially harmful to leave decisions about whether and how to regulate such pathogens to the whims of individual risk assessors.

On a similar note, if an entity is known to present a *low* threat, auditors can use a checklist to automatically "filter out" such systems from their analyses, increasing the efficiency of their work. FSAP inspectors do not have to worry about the (presumably overwhelming majority of) pathogens known to be lower-risk, such as the common cold.

Clarity

Secondly, it seems generically true (with respect to FSAP but also to all other regulations) **that checklists set "clear requirements and expectations"** (Burnett et al., 2016, p. 53).

To illustrate this point with reference to FSAP, a 2009 survey of US biodefense researchers found that "41% of respondents preferr[ed] the clear regulatory guidance" of FSAP to the Biosafety in Microbiological and Biomedical Laboratories (BMBL) handbook (Sutton, 2009, p. 225), which adopts a more risk-based approach (Stepanlos, 2023[24]). (Although FSAP *does* draw on BMBL, it does so in order to produce "checklist"-type requirements). According to the survey, "respondents indicated that more specificity is needed for training requirements in BSL [biosafety levels] (88.4%), regulatory compliance (76.3%), and emergency response (61.1%) training, rather than merely what is 'appropriate'" for a given case (which is BMBL's approach) (Sutton, 2009, p. 225).

[24] See the section "Key principles and procedures"

However, we would note that this does *not* seem like strong evidence against adopting a more risk-based approach.

- Firstly, the evidence presented in the survey is a little one-sided – it does not say whether the remaining 59% of respondents all preferred BMBL to FSAP, or whether there was a large proportion of "don't know" responses, etc. We could not find the original data underpinning the survey.
- Secondly, and more importantly, BMBL is an advisory *standard*, not a regulatory document. As a result, as far as we are aware, there are no government auditors who undertake risk assessments of laboratories according to BMBL and mandate safety measures if they deem laboratories to be falling short.
- So laboratory researchers' worries about BMBL could be based on the fact that they themselves (the researchers) have been left to interpret how they should be complying with BMBL, rather than a government body giving them instructions based on a risk-assessment.

Moreover, one of our interviewees, a former FSAP inspector, noted that even FSAP checklists sometimes do not provide adequate clarity to judge whether a situation is or is not in breach of the checklist – in that case, judgment calls have to be made, often collectively by the inspection team.

Accessibility

A third, generic (rather than FSAP-specific) advantage of checklist-style systems is that "[i]nspectors and regulated entities require less technical depth" in order to implement the regulations (Burnett et al., 2016, p. 53).

(That actually seems *less* relevant in FSAP's case, as inspectors seem to have decent technical knowledge (see above)).

Ability to handle complex tasks

Atul Gawande's "The Checklist Manifesto" (2009) claims that **checklists also seem generically useful when handling complex tasks involving multiple difficult steps, which even an experienced expert may fail to remember to execute in full** (for example, complicated surgical tasks). Since then, several empirical studies have supported that

claim.[25] (We are unsure whether that claim applies to FSAP, but it seems like a generally useful point to bear in mind).

Harder to "game"?

Fifthly, it seems at least *plausible* that checklists may be harder to "game" – in the sense of regulated entities exploiting loopholes in existing regulation, which enable them to adhere to the letter of the law without actually improving safety in any meaningful sense.

For instance, it seems intuitive that checklists could prevent bias from determining whether an FSAP-regulated entity clears its annual inspection. We can imagine a biosecurity regulator captured by industry would find it easier to give an entity an easy pass if it did not have clear criteria against which it had to check entity safety performance.

On the other hand, it is plausible that checklists would be easier to game if the checklist is publicly available, as regulated entities can just paper over safety faults to address the checklist.

FSAP provides some evidence that checklists can have that undesirable feature. Bjork and Sosin (2017) undertake a survey of ~2,250 individual violations of FSAP standards detected by FSAP inspectors, and find that unannounced inspections (i.e., inspections about which the inspected parties had no prior knowledge) reveal a greater number of high-risk violations than announced (i.e., scheduled) inspections (Bjork and Sosin, 2017, p. 1). And two of our interviewees confirmed that, noting that with announced inspections, regulated entities can fudge data and "clean up" in advance of an inspection to make sure that the entity looks good. If there was no publicly available checklist of desiderata, however, that would presumably not be possible.

All-in-all, that suggests that checklists should ideally be:

(a) **kept non-public**,
(b) **updated frequently** (so that regulated entities cannot just learn what inspectors are looking for from the last inspection), and/or
(c) **enforced by unannounced (rather than announced) inspections**,
(d) **combined with more "risk-based" approaches.**

[25] For example, Powell, Jain, & Juneja (2019: see "Comment" section) cite studies demonstrating that "A WHO surgical checklist has been shown to consistently reduce perioperative mortality in a variety of settings, and this strategy has now permeated other areas of healthcare delivery with some success."

However, on balance, FSAP's checklist-based approach seems poor at handling uncertainty/tail-risks

FSAP's use of risk "tiers" does seem like an improvement on the previous regime, which treated all agents on the FSAP list of select agents as equally risky (GAO, 2017, p. 13-14). **However, overall *FSAP seems to have performed poorly when handling similar risks to extreme AI risk* in the bioengineering domain, i.e., tail risks from bioengineered agents that are subject to high uncertainty.** [26] All three of our interviewees agreed with that claim.

Poor handling of tail risks

As noted above, FSAP standards cover only a checklist of pathogens that are currently known to be dangerous. Many experts think that that approach fails to adequately address GCBRs, as it does not address risks from agents that could be *engineered* to be highly dangerous.

For example, biorisk expert Tom Inglesby (2018), being interviewed by Rob Wiblin on the *80,000 Hours* podcast, notes that FSAP "does not yet really formally touch on GCBR related issues. It's very concretely based on pathogens that exist that are already known to be high consequence" (Inglesby, 2018[27]).

As a result of such concerns, Lewis et al. (2019) note that "Although [the] select agent [list]… can be [a] heuristic… for enhanced caution, we urge stakeholders to be mindful of the prospect for misuse of work that does not fall into [that category]" (Lewis et al, 2019: 979).

Poor handling of uncertainty

One reason that FSAP finds it difficult to deal with tail risks from potential bioengineered agents is that experts disagree about which agents could be bioengineered to become highly dangerous. As Epstein (2023) notes, *"Even informed experts might disagree on whether a particular pathogen is an enhanced potential pandemic pathogen*, or whether a proposed experiment is likely to generate one" (Epstein, 2023[28]). Simple checklist-style systems,

[26] FSAP does regulate *some* (highly specific) bioengineering experiments with *existing* select agents – see the above section on "Regulation of development (as well as deployment)" – but this is only a very small subset of potential experiments, and does not cover bioengineering experiments using pathogens not currently on the select agents list.

[27] See the section after Wiblin's question "So what kind of levers are actually available to reduce the chance that biotechnology goes on to cause harm?"
[28] See the section "Utilizing existing legislation". Our emphasis.

which assume that a pathogen clearly does or does not fall into a given binary category, are clearly inadequate here.[29]

Other authors have raised more prosaic concerns around FSAP's ability to handle uncertainty. For example, it can even be hard to interpret whether a particular pathogen falls under the *existing* list of select agents, because "Microorganisms that are added to the select agents list are codified by standards of taxonomy, which is a problematic approach given the uncertainty surrounding what constitutes a microbial species. Consequently, the boundaries between a select agent pathogen and a similar sequence from a related species are often equivocal and unclear" (Kalra and Parker, 2022, p. 64). As a result, by 2016 there were thirteen documented cases where such ambiguities led to microbial collections being destroyed, potentially "impeding scientific advances meant to promote public health" (Kalra and Parker, 2022, p. 65).

Proposals to amend FSAP to deal with these problems have not yet come to fruition, and show no sign of doing so soon

Since FSAP's inception, there have been several proposals to at least consider updating the program to make it more risk-based. Some of those proposals (particularly the *earlier* ones) explicitly aimed to better deal with novel risks from bioengineered agents. **However, as far as we can tell, those proposals have not really come to fruition.**

Lessons from the 2010 National Research Council report

Overview of the NRC report

The first such proposal – and probably the most interesting – was a 2010 National Research Council report by a committee of bioscientists charged with investigating whether "predicted features and properties encoded by nucleic acids, such as virulence or pathogenicity," could be "used in lieu of the current finite list of specific agents and taxonomic definitions" (National Research Council, 2010: vii, our emphasis):

That idea arose directly from the concerns that we raise above – that existing pathogens that are *not* select agents could be bioengineered to become more dangerous, or (more speculatively) that synthetic biology may allow the development of entirely novel pathogens (National Research Council, 2010, p. 1).

[29] On the other hand, one reviewer noted that high uncertainty should give us some doubt as to whether any given expert's judgment will perform better than chance. That could be an argument in favor of employing highly *specialized* experts who spend a lot of time on thinking about ePPP risks and are therefore likely to perform better than chance.

Such a system would be relatively similar to what (we think) many envisage in a regulatory system for AI:

- It is *prospective* – i.e., it tries to *predict* what dangerous capabilities could emerge in the future.
- It is a system based on *risk factors,* rather than a list-based system – it stipulates "If we predict that, in any given case, virus X will exhibit risk factor Y [e.g., "more virulent"/"more pathogenic"], we will regulate it."

Below, we outline several lessons from the NRC report which seem particularly relevant to AI risk regulation.

Lesson 1: Risk assessment in highly uncertain environments is hard

One interesting lesson from the NRC report is that risk analyses of emergent dangers may be extremely difficult, since the properties of organisms (as with future AI models) are not easily predictable. The committee concluded that features such as pathogenicity and virulence could not "*plausibly be predicted with the degree of certainty required for regulatory purposes*" using sequence-based technologies for decades, if not more than a century (National Research Council, 2010: 2, our emphasis). Such prediction would "require an extraordinarily detailed understanding of host, pathogen, and environment interactions integrated at the systems, organism, population, and ecosystem levels" (National Research Council, 2010, p. 2). Moreover, many properties that make an agent dangerous – e.g., the presence or absence of available countermeasures for the disease it causes – are socio-technical, not biological, phenomena (National Research Council, 2010, p. 2, 110).

Similarly, because many unanticipated model capabilities emerge suddenly beyond a given scaling threshold (see, e.g., Wei et al., 2022: abstract), predicting the capabilities of a model in a manner suitable for formal risk-assessment may be extremely hard. That means that society must (a) invest in finding out how evaluators might be able to predict such capabilities, but also (b) find a way of pausing progress until researchers have figured out how to predict such capabilities.

Lesson 2: Analysis of proxy variables could help overcome that difficulty, but only imperfectly

One response to such an uncertain risk-landscape is developing *proxies* for danger, such as using data and expert knowledge to catalog "early warning sign" behaviors, or by identifying broad development processes that seem to produce harmful results.

In the case of FSAP, the 2010 committee did note that while it was not possible to *predict* features such as virulence or pathogenicity from genome sequences, it could be possible to classify *existing* sequences that "potentially could be used to produce a threat" as "sequences of concern." Such sequences could be added to a database and marked with a "yellow flag" (National Research Council, 2010: 3-4).

One analogy with AI here could be: a model demonstrates a dangerous capability, e.g., creating a novel pathogen; evaluators identify previous behaviors in the model that could have been "early warning signs" of it having said dangerous capabilities; evaluators add those behaviors to a database and tell researchers/evaluators to look out for them.

Instead of early warning sign behaviors, evaluators could also try to identify potentially risky *development processes* – for example, checking whether a dangerous-seeming model was of unusual/unprecedented size, was trained on different types of data than usual, was given different kinds of fine-tuning than normal, etc. If evaluators determined that those variables were risk-factors, they could add them to a database and/or maybe require higher precautions/greater presumption of risk when building a model that shares those features.

Lesson 3: Risk-assessment techniques can be dual-use

Moreover, the NRC report noted that "Developing the ability to predict [e.g. …] pathogenicity from genome sequence raises serious dual-use concerns, because prediction and design go hand in hand" (National Research Council, 2010, p. 6). That is possibly a risk with various AI risk-assessment methods as well – for example, developing scaling laws that help us predict when dangerous behaviors might emerge, or checklists of types of training datasets or precursor behaviors that could be correlated with the emergence of such dangerous behaviors.

The downside seems worse when:

- The dangerous property being tested for is appealing to develop, e.g., because it bestows economic or military advantages.
- The risk assessment provides a "recipe" for building the dangerous capability. For example, public benchmarks evaluating dangerous behavior, like the Machiavelli benchmark, can be trained against to produce a model that maximizes the risk in question.

Thus, evaluation processes testing for sufficiently dangerous behaviors in AI models should probably not be made public, at least not public enough to provide easy learning for malevolent actors.

On the other hand, the genomic sequencing example seems like a fairly extreme example of "providing a recipe," compared to most examples of AI risk evaluations. But they seem sufficiently analogous that this consideration should not be ignored.

Other recommendations that FSAP adopt a more risk-based approach

Aside from the NRC report, a second proposal to update FSAP – also directly addressing the concerns about bioengineering that we outline above – arose from an investigation by the Fast Track Action Committee on the Select Agent Regulations (FTAC-SAR), set up by the National Science and Technology Council in the wake of a series of biosafety incidents in 2014 (FTAC-SAR, 2015, p. iii).

FTAC-SAR made a number of concrete recommendations, none of which are particularly relevant to this case-study. However, in its "Issues for further analysis" section – which stops short of making concrete recommendations for regulatory change, and instead outlines proposals that should be investigated further first – it determined that CDC should "explore the feasibility of adopting a "risk-based" approach to managing the safety and security oversight of biological agents and toxins" (FTAC-SAR, 2015, p. v).

The committee made that recommendation precisely because it recognized that FSAP only regulates pathogens currently known to be dangerous, failing to deal with the possibility that bad actors or poorly-informed researchers could manipulate pathogens to make them more dangerous (FTAC-SAR, 2015, p. 20). It noted that a risk-based approach may be superior to a list-based approach here, but that the feasibility of implementing such an approach should be investigated first (FTAC-SAR, 2015, p. 20).

A 2015 CDC internal review and a 2017 Sandia National Laboratories independent review also recommended that FSAP adopt a more "risk-based" approach. However, those examples seem less relevant, as they focused largely on getting FSAP inspectors to conduct more risk assessments relating to activities that *already fell within the scope of the program*, rather than wholly revising FSAP so that it could identify and deal with the biggest tail risks (from bioengineered agents).

FSAP's own attempts at adopting a more risk-based approach

There is some evidence that FSAP itself has tried to adopt a more risk-based approach. FSAP's strategic plan from FY18-FY21[30] states that one of the program's key goals is to "Leverage… *risk-based* approaches to guide FSAP operations" (CDC, 2018, p. 4). However, once again, the "risk-based" approach proposed here seems to be about assessing which activities in entities regulated under the *current scope* of the program (i.e., those handling

[30] Its latest published strategic plan, as far as we can tell from this link and a google keyword search.

BSATs) posed the highest risks, so FSAP could better allocate its resources (CDC, 2018, p. 9). There is nothing in the strategic plan mentioning risks from novel bioengineered agents.

Even FSAP's more limited efforts to develop a more risk-based approach seem to have not yet been fulfilled. Our third interviewee, a senior official at a high-containment laboratory subject to FSAP inspections, believed that CDC DSAT started to become interested in such an approach some time between 2015 and 2020, which would be consistent with the above evidence. However, she believed that COVID-19 had "thrown that [the risk-based approach] out the window." We would guess that she meant that the pandemic had distracted CDC officials from reforming FSAP – i.e., that officials during the past ~3 years had been largely reacting to the challenges posed by the pandemic, rather than trying to make FSAP more anticipatory/risk-based. But that seems uncertain.

In general, our interviewee's impression was that leadership at CDC and at the regulated entities tend to be afraid of the idea of "risk-assessment" for two reasons:

1. Concerns (both among laboratory personnel and inspectors) that they lack the proper expertise to conduct adequate risk-assessments.
2. Concerns (among CDC officials) about liability – with a checklist-based system, CDC cannot be held liable for incidents at FSAP-regulated entities; instead, they can just claim that the entity should have followed the checklist. With a risk-based system, CDC could be held liable if its inspectors fail to adequately assess a given risk.

It seems plausible that liability concerns could present an issue for AI risk assessors as well; further research could look into that.

Other countries appear to deal better with uncertain tail risks

Our impression is that other countries – in particular Canada and the United Kingdom – deal better with uncertain tail risks from bioengineered agents.

A 2017 GAO report noted that Canada and the UK provide particularly instructive examples of more "risk-based" biological regulation (GAO, 2017, p. 43). This section therefore focuses on those two regimes.

Canada

In Canada, laboratories that conduct "controlled activities" with human pathogens or toxins must obtain a license from the government (PHA Canada, 2018[31]). Those licenses cover a broad range of activities, including possessing, handling, using, producing, storing,

[31] See section 2.1.1. entitled "Pathogen and toxin regulation in Canada"

transferring, importing, exporting, releasing, abandoning, disposing, and permitting any person access (Government of Canada, 2023b[32]).

If they are conducting scientific research,[33] **laboratories must develop a Plan for Administrative Oversight (PAO) describing how they will control biosafety/biosecurity risks, including "risks from research with dual-use potential," such as risky bioengineering research (**PHA Canada, 2018[34]). Our understanding is that, if the laboratories fail to submit such a plan, they will not get issued a license (Government of Canada, 2023d: paragraph 1). A given laboratory must also report "any activities that could result in the creation of a human pathogen with increased virulence, pathogenicity, or communicability" to the laboratory's license-holder and its Biological Safety Officer (Government of Canada, 2023e: see the section "Licenses").

In practice, the above regulations cover all potentially dangerous bioengineering research, including all "gain-of-function" research (Government of Canada, 2023d[35]). In that respect, they are superior to the UK regulations (see below).

The Canadian government also provides non-mandatory "Guidelines for Dual-Use Research," where it states (among other things) that "Every research project should be reviewed for dual-use potential during the planning stages, throughout the course of the project… and prior to the use or dissemination of the results" (PHA Canada, 2018[36]).

Our impression is therefore that the Canadian system is a relatively positive example of risk-based biosecurity/biosafety regulations. Pannu et al. (2022), in a *Science* paper, also highlight the Canadian system as "address[ing] key elements" of the problem of dual-use research (Pannu et al. 2022: paragraph 4). And our third interviewee also seemed to think that the Canadian system was a good example of a "risk-based" approach, although it is worth noting that that was only a weak impression and she had not worked in Canada.

The Canadian system also does not sacrifice the benefits of list-based systems as outlined above - it includes a list of pathogens *known* to be dangerous, known as Security Sensitive

[32] See the section "Applying for a Human Pathogens and Toxins Act licence: what you need to know."
[33] Rather than, say, diagnostics or standardized private industry production (Government of Canada, 2023c: see section 6.0: "Issue Analysis").
[34] See Chapter 3, entitled "Risk assessment of research with dual-use potential"
[35] See the section entitled "Elements to be included in Plans for Administrative Oversight for Pathogens and Toxins in a Research Setting"
[36] See section 3.1. entitled "Identify research with dual-use potential"

Biological Agents (SSBAs), which are subject to particularly stringent regulation.[37] However, our point is that Canadian biolaboratory regulations do not *only* cover such pathogens – indeed such pathogens "represent only 0.2% of all regulated [biolaboratory] work in Canada" (Pomerleau-Normandin, Heisz and Tanguay, 2018, p. 298).

The Canadian system is probably not completely comprehensive, though – Pannu et al. only state that the Canadian system addresses "key elements," not that it is comprehensive (Pannu et al. 2022: paragraph 4). It seems uncertain what such a comprehensive system would look like.

United Kingdom

The Chemical, Explosives and Microbiological Hazards Division within the UK Health and Safety Executive (HSE) "regulates sectors that have the potential for low-probability, high-consequence incidents," including biological laboratories (GAO, 2017, p. 43).

The most relevant piece of regulation under the Division's authority seems to be the Genetically Modified Organisms (Contained Use) Regulations 2014. **The regulations stipulate that *all* research involving genetic modification of pathogens must be preceded by a risk-assessment that is submitted to the regulatory authorities (**UK HSE, 2023a: See the section "What is 'contained use?'"). If a risk-assessment has not been conducted, the research cannot be carried out (UK HSE, 2023b, p. 6). After the risk assessment has been completed, the research is classified under a given "containment level", which determines the control measures to be implemented (UK HSE, 2023b, p. 34).

The regulations explicitly attempt to "take account of advances in technology, for example synthetic biology is largely encompassed by the definitions in the regulations and is likely to remain so within the foreseeable future" (UK HSE, 2023c: see the section "Why have the regulations changed?"). **And they explicitly mention tail risks from bioengineering**, noting that "There are some types of work where particular caution must be exercised. These are generally cases where the pathogenicity or host-range of a pathogen has been enhanced or altered" (UK HSE, 2023b, p. 7).

---

[37] One reviewer noted the following:

- The Human Pathogens and Toxins Act itself specifies five different lists – one for toxins, and four for pathogens ordered by different risk-tiers (groups 2-4). (Government of Canada, 2023f: see Schedules 1-5 at the end of the document).
- One of the pathogens lists is for pathogens that are completely illegal to work with – currently this just includes variola virus (Smallpox) (Government of Canada, 2023f: see Schedule 5, "Prohibited Human Pathogens and Toxins").
- The government additionally maintains a list of SSBAs. This is the intersection of the pathogen risk groups 3-4 mentioned above, and the Australia Group export list.

We could not find much evidence on the effectiveness of the regulations. The UK HSE does note, however, in its page on the GMO Contained Use regulations, that "[t]he safety record in this industry is extremely good" (UK HSE, 2023a: see the section "What is 'contained use?'"). Nonetheless, in a review of different countries' policies on Dual-Use Research of Concern (DURC) in the biosciences – which covers the risks from bioengineered agents that we outline above – Piers Millett (2017) interviewed two experts from the UK who noted that **"apart from existing Health & Safety and GM regulations, and dual-use export control legislation, there are no laws or regulations specific to DURC" (**Millett, 2017, p. 11). That suggests to us that forms of dangerous bioengineering research that are not covered by the GMO Contained Use Regulations, for example, gain-of-function research that involves repeatedly passing a pathogen through an organism, are simply not covered by existing UK regulations.

Level of anticipatory regulation

Brief description of relevant FSAP procedures

After a reasonably extensive keyword search on databases such as Google Scholar, **we could find no evidence that FSAP engages in anticipatory regulation-setting.**

Instead, FSAP regulations seem more to be *reactions* to existing threats/events. The original program, to govern the transfer of select agents, was set up in response to the Larry Wayne Harris vaccine hoax (Kirkpatrick et al., 2018, p. 25). FSAP was then expanded to cover possession, in direct response to the 2001 World Trade Center attacks and anthrax attacks (Kirkpatrick et al., 2018, p. 25). (And as previously noted, FSAP's focus is on a list of select agents that are known to be highly pathogenic *at present*).

Evaluation of effectiveness

Ill-adapted to rapid technological change

In general the "reactive" nature of FSAP seems somewhat negative, as it seems to make FSAP ill-adapted for dealing with the rapidly-changing worlds of biosecurity and biosafety.

Kirkpatrick et al. (2018) note that "[most] policies… remain static or undergo only incremental changes until an exogenous event creates the conditions necessary for a dramatic changes", something they call the "reactive model of policy-making" (Kirkpatrick et al., 2018, p. 25). They note that it "is widely viewed as being unsuited for an era of rapid technological change" (Kirkpatrick et al., 2018, p. 25). The authors argue that US biosecurity policy (of which FSAP is a part) fits that model particularly well. As noted above, "The major U.S. biosecurity policies were enacted in direct response to specific incidents associated with biological terrorism" (Kirkpatrick et al., 2018, p. 25). In fact, Kirkpatrick et al. claim that, "[p]erhaps in part because biosecurity policy is enshrined in legislation and implemented through regulations, policy in this area [i.e. the biosecurity area] is the slowest to change [of all life-sciences regulation]" (Kirkpatrick et al., 2018, p. 25).

Moreover, the World Health Organization (WHO) has noted (with respect to FSAP specifically) that "Due to the speed of [technical] advancements, lists can quickly become outdated and create holes in the biorisk management system as new technologies and their associated risks are not listed" (WHO, 2022, p. 51). As noted above, our impression is that

FSAP has not been particularly effective at anticipating novel risks from bioengineered agents. Gronvall (2008), quoting from a National Science Advisory Board for Biosecurity (NSABB) report, notes that "advances in synthetic genomics and synthetic biology could lead to new pathogens that will not easily fit into [FSAP's] oversight framework" (Gronvall, 2008: see bullet-point list).

Given that AI is a similarly rapidly-advancing domain, we might infer that reactive systems of regulation are similarly ill-suited to AI.

Relevance to the AI case

Positive lessons for the AI case

FSAP and parallel biorisk regimes in other countries provide a good precedent for an AI licensing regime in some respects

As Anderljung et al. (2023) note, most licensing regimes require companies to obtain a license to widely "deploy" (e.g., use/sell) some product once it is developed, to follow rules about safety, and to regularly demonstrate that they are following said rules (Anderljung et al., 2023, p. 21).

However, it is rare to require licenses for the R&D phase (Anderljung et al., 2023, p. 21). **FSAP is a good example of an exception – it requires a license ("registration") for any activity involving select agents, including R&D.** And there is some counterfactual evidence in favor of that licensing regime being fairly effective at reducing risk – as previously noted, a number of entities have had their licenses revoked for serious breaches of FSAP regulations.

Other countries' licensing regimes for biological research seem still better tuned to reducing the largest risks from bioengineering R&D. For example, Canada requires all laboratories that conduct "controlled activities" with human pathogens or toxins to obtain a license, and one condition for obtaining that license is that laboratories must conduct risk assessments and come up with risk mitigation measures for novel/unanticipated risks (e.g., from bioengineered agents). It stipulates that said risk mitigation should apply to the research design, implementation, and use/deployment phases.

That suggests that development/R&D-phase regulations could be set up for AI, and could be relatively stringent/effective. On the other hand, FSAP regulations were only put in place as a reaction to terrorist attacks, so perhaps that implies that similarly intensive regulations on frontier AI R&D will not occur until after some kind of similar "warning shot" event.

Moreover, both FSAP and the Canadian regime also attempt to ensure that said R&D-phase regulations are proportionate to the risks. FSAP only regulates research involving the most dangerous existing pathogens. Although we have argued above that that approach does not adequately deal with GCBRs from bioengineered pathogens, it at least ensures that stringent R&D-phase regulations do not apply to laboratories conducting innocuous research. The Canadian system, meanwhile, is a risk-based system – by nature it requires more stringent safety measures for activities judged to be higher-risk (and vice-versa).

Similarly, R&D-phase licensing requirements for AI should only be applied to the highest-risk models, to avoid stifling innovation and cementing the market power of the leading laboratories. Again, FSAP/Canada's biosafety regime provide a precedent here. (On the other hand, as previously noted, some have suggested that FSAP regulations stifle biodefense innovation, although the evidence here is quite mixed, and further analysis is probably needed to confirm/falsify such claims).

With appropriate training, federal agencies can do a good job of conducting inspections

FSAP is a good example of a *federal* program managing to employ inspectors with a decent amount of expertise (although as noted above, more continuous training should be provided, and third-parties could be brought in to advise/audit inspectors).

That suggests that **a well-funded federal regulatory body that provided ongoing training for its inspectors and included regular external input/verification *could* do a good job of regulating frontier AI systems**; it would not have to rely entirely on third parties.

On the other hand, it seems at least plausible that an AI regulator would have a more difficult time attracting AI experts of a similar caliber, for example because private-sector salaries are relatively speaking much higher in the AI sector. Further research is needed here.

Checklist-style systems have some advantages, if combined with risk-based regulation

There are some generic advantages to "checklist"-style regulations, most notably that some hazards are known to be severe with a strong degree of certainty (e.g., the plague or ebola have known high fatality rates) and have well-known control measures. In that case, having a checklist of stringent controls around the hazard seems sensible. Relatedly, it seems that checklists could help auditors to automatically remove from their consideration large numbers of entities known to be lower-risk (e.g., weak viruses that have not undergone viral engineering, or smaller "narrow" AI models in low-stakes applications).

Checklist-style systems also seem clearer, to entail less requirements for technical expertise on the part of auditors, to provide useful guidance on complex tasks involving multiple easy-to-forget steps, and to *maybe* offer less opportunity for regulated entities to "game" inspections (although the evidence here is somewhat contradictory).

That suggests that an ideal AI auditing regime *might* include a *combination* of checklist-style regulations and risk-based regulations.

- The Machiavelli benchmark could be a good example of a checklist-style system for AI: it employs a standardized set of prompts and then evaluates how harmful the AI's response is to each prompt.

On the other hand, if such a checklist was made public, laboratories could "teach to the test," giving a false impression of model safety (for example, by including the checklist prompts in the model's training data). **So we think that our recommendation relies on such checklists:**

(a) **being kept private from labs**,

(b) **being updated frequently**, so that labs do not "know what's coming" – which has the additional advantage of ensuring the list keeps up with technical developments,

(c) (possibly) being enforced by unannounced inspections (although it seems possible that that suggestion is less relevant for the AI case, as whether laboratories can "teach models to the test" is perhaps orthogonal to inspection timing),

(d) **always being combined with risk-based regulations** (see advantages of said regulations for the AI case below). As outlined above, the Canadian biosafety regime seems to do a good job of combining more list-based regulations with risk-based regulations (it has a similar list of "*known*-to-be-dangerous agents," but said agents only account for a small proportion of pathogens covered by the regulations, which regulate all *potentially* hazardous pathogens).

Negative lessons for the AI case

Checklists seem inadequate for dealing with extreme risks; "risk-based" regulation is probably needed

FSAP's "checklist" approach to regulation seems to have prevented it from adequately handling uncertainty/tail risks, and from anticipating future risks.

Checklists appear to be particularly bad at handling risks that are:

- ***Extreme*** - because such checklists tend to lump risks into either one category or several broad categories (e.g., "high," "medium," "low"), which fails to single out risks that may be *particularly* extreme/likely.

- ***Uncertain*** - because in that case, experts will disagree about whether the criteria on the checklist are being fulfilled.
- ***Rapidly advancing*** - because such checklists can quickly become outdated, and because phenomena that currently appear harmless (according to the checklist) could become more dangerous in the future due to technical advancements.

In particular, those considerations suggest that a more "risk-based" approach to regulation, that gives regulators the discretion to assess and mitigate risk based on the specific situation confronting them in any given instant, would be more appropriate for dealing with extreme AI risks.

An AI-specific intuition in this direction is that (as noted above) laboratories could overcome "checklist"-style regulations by simply teaching their models to *look* safe according to the checklist (when in fact they perhaps are not). Moreover, even if AI laboratories do not/cannot do that, checklist-style regulations applied to the training process could simply fail to capture risks that might arise when the model is deployed. That is because a fixed set of standardized prompts cannot possibly capture all the factors that would arise outside of the testing environment (i.e., when the model is made public) to make the model more dangerous, e.g., a much wider distribution of possible prompts.

It may be hard to develop good risk-based regulations if capabilities are hard to predict

An important proposed component for frontier AI regulation involves conducting risk-assessments informed by evaluations ("evals") of dangerous model capabilities (see, for example, Anderljung et al., 2023: abstract, 3-4, 24). Such evaluations would aim to detect model behaviors that could be predictive of extreme risks from that model.

The 2010 National Research Council report on FSAP, detailed above, evaluated a somewhat analogous proposal – the idea that dangerous pathogen "capabilities," such as virulence or transmissibility, might be predicted using genome sequences. That would allow FSAP to go beyond its "list-based" approach and instead focus on emergent pathogens.

However, the NRC report found that predicting such traits was a task of so much complexity that doing so with the degree of certainty required for regulatory purposes would not be possible for many decades. **Similarly, it may be incredibly hard to anticipate AI model capabilities, which often emerge unexpectedly, with the degree of certainty required for formal risk-assessment.** In that case, the correct way forward may be a precautionary approach which combines investment in predictive abilities with a system for pausing development/progress until our ability to anticipate/detect risks has improved.

One alternative system suggested by the NRC report was to identify *existing* "sequences of concern" which have the potential to create novel biothreats, add them to a database, and mark them with a "yellow flag."

- An analogy with the AI case might be: Evaluators cannot predict many dangerous capabilities before they emerge; but they can look at dangerous behaviors exhibited by existing models (e.g., developing novel pathogens/toxins), identify prior behaviors/training parameters that retrospectively seem correlated with said dangerous behaviors, and add the latter to a database with a yellow flag.
- Instead of early warning sign *behaviors*, evaluators could also try to identify early warning sign *development processes* – for example, checking whether a model is of unusual/unprecedented size, has been trained on different types of data than usual, has been given kinds of different fine-tuning than normal, etc. If evaluators determined that those variables were risk-factors, they could add them to a database and/or maybe require higher precautions/greater presumption of risk when building a model that shares those features.

Moreover, if evaluating/predicting novel capabilities is extremely difficult, it *may* make more sense to invest heavily in safety research. On the other hand, AI safety research may *itself* inadvertently advance capabilities (making such research more akin to, e.g., biodefense research using gain-of-function techniques).

Other possible ways of making up for the difficulty of assessing risks from AI systems include:

- **Combining risk-based approaches with checklists** that catch at least *some* dangerous behaviors (see above).
- **Adopting "defense-in-depth" approaches** such as hardening cyber or phishing defenses at key institutions.

Finally, even if a dangerous capability *was* easy to predict, actually predicting it in a public risk-assessment could be very dangerous, as 'prediction' and 'development' often go hand-in-hand. The NRC report recognized this, and consequently advised against the development of a sequence-based prediction system for dangerous pathogens. Similar concerns exist with respect to AI capabilities evaluations. For example, public benchmarks evaluating dangerous behavior, like the Machiavelli benchmark, can be trained against to produce a model that maximizes the risk in question.

Thus, evaluation processes testing for sufficiently dangerous behaviors in AIs should probably not be made public, at least not public enough to provide easy learning for malevolent actors.

Acknowledgements

We are grateful to the following people for discussion and input: Michael Aird, Marie Davidsen Buhl, Oliver Guest, Rose Hadshar, Patrick Levermore, Cullen O'Keefe, Joe O'Brien, and Zoe Williams. Mistakes and opinions are our own. We are also grateful to Adam Papineau for copyediting.

Appendix 1: Exemptions to FSAP regulations

Some of the exemptions to FSAP's general rules seem reasonable – for example, if the entity in question has destroyed or transferred all of its select agent material within a given time window and according to strict procedures, it becomes exempt from the regulations (Code of Federal Regulations, 2023: see the section "Exemptions for HHS select agents and toxins").

Others seem less reasonable – for example, according to Brooks (2018), "clinical or diagnostic labs that receive "specimens" for diagnosis or verification do not have to register with CDC or APHIS as long as they follow certain protocols after a reception… This means that a non-CDC-registered—and therefore a possibly non-compliant—lab could be housing perhaps even a Tier 1 select agent without proper protocols in place to ensure its containment from both an infectious disease standpoint and security standpoint" (Brooks, 2018, p. 6).

Moreover, "entities that possess, use, or transfer attenuated strains of select agents and toxins are also exempt from registration." Such an exemption is relatively difficult to obtain, but Brooks argues that it could still pose a risk to biosecurity: "If an entity wishing to possess such an attenuated strain is exempt from registration, such an entity could grant access to individuals that have not undergone the [vetting…] process" (Brooks, 2018, p. 7).

References

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (arXiv:2307.03718). arXiv. https://doi.org/10.48550/arXiv.2307.03718

Berger, K. (2011). Select Agent Rules. In R. F. Pilch & R. A. Zilinskas (Eds.), *Encyclopedia of Bioterrorism Defense* (1st ed., pp. 1–7). Wiley. https://doi.org/10.1002/0471686786.ebd0191

Berglind, N., Fadia, A., & Isherwood, T. (2022). *AI in Government: Capturing the Potential Value*. https://perma.cc/TR84-FUFS

Bjork, A., & Sosin, D. M. (2017). Characterization of Departures from Regulatory Requirements Identified During Inspections Conducted by the US Federal Select Agent Program, 2014-15. *Health Security*, *15*(6), 587–598. https://doi.org/10.1089/hs.2017.0054

Brooks, C. (2018). Classifying and Regulating Biological Agents in the United States: Problems Posed to Global Biosecurity. *Journal of Biosecurity, Biosafety, and Biodefense Law*, *9*(1), 20180001. https://doi.org/10.1515/jbbbl-2018-0001

Burnett, L. C., & Brodsky, B. H. (2016). *Biological Select Agents and Toxins: Risk-Based Assessment Management and Oversight* (SAND-2016-12706). Sandia National Lab. (SNL-NM), Albuquerque, NM (United States). https://doi.org/10.2172/1432263

Centers for Disease Control and Prevention, & Animal and Plant Health Inspection Service. (2020b). *Responsible Official Resource Manual*. https://perma.cc/5JCX-GEY4

Centers for Disease Control and Prevention (CDC). (2018). *FSAP Strategic Plan, FY18-FY21*. https://perma.cc/CK45-CLXP

Centers for Disease Control and Prevention (CDC). (2020a). *Responsible Official Resource Manual: Fundamental Responsibilities of the Responsible Official*. https://perma.cc/MFX9-8TRF

Centers for Disease Control and Prevention (CDC). (2023). *Inspection Checklist for Incident Response (7 CFR 331, 9 CFR 121, 42 CFR 73)*. https://perma.cc/KU82-6CPF

Code of Federal Regulations. (2023). *42 CFR Part 73—Select Agents and Toxins*. https://perma.cc/UJ9D-SY4E

Ee, S., O'Brien, J., Williams, Z., El-Dakhakhni, A., Aird, M., & Lintz, A. (2023). *Adapting Cybersecurity Frameworks to Manage Frontier AI Risks: A Defense-in-Depth Approach*. Institute for AI Policy and Strategy. https://perma.cc/9YCV-KMTZ

Epstein, G. D. (2023, February 16). Private-Sector Research Could Pose a Pandemic Risk. Here's What to Do About it. *Bulletin of the Atomic Scientists*. https://perma.cc/SKF8-Y8JS

Fast Track Action Committee on the Select Agent Regulations (FTAC-SAR). (2015). *Fast Track Action Committee Report: Recommendations on the Select Agent Regulations Based on Broad Stakeholder Engagement*. Science and Technology Policy Office. https://perma.cc/C6E6-HMAS

Federal Select Agent Program (FSAP). (2023, August 1). *Select Agents and Toxins List.* https://perma.cc/2CFY-FBWC

Finlay, B. D. (2010). *Pharmaceutical Terror: Getting Health Care Reform Right*. https://perma.cc/D6SB-Y6VM

Gawande, A. (2009). *The Checklist Manifesto*. Profile. https://perma.cc/L7NY-X74E

Government of Canada. (2023b). *Licensing Program* [Service initiation]. https://www.canada.ca/en/public-health/services/laboratory-biosafety-biosecurity/licensing-program.html

Government of Canada. (2023c). *Scientific Research Policy for Human Pathogens and Toxins* [Guidance;education and awareness]. https://www.canada.ca/en/public-health/services/laboratory-biosafety-biosecurity/scientific-research-policy-human-pathogens-toxins.html

Government of Canada. (2023d). *Plan for Administrative Oversight for Pathogens and Toxins in a Research Setting—Required Elements and Guidance* [Policies]. https://www.canada.ca/en/public-health/services/laboratory-biosafety-biosecurity/licensing-program/plan-administrative-oversight-pathogens-toxins-a-research-setting-required-elements-guidance.html

Government of Canada. (2023e). *Consolidated Federal Laws of Canada: Human Pathogens and Toxins Regulations*. https://perma.cc/6K8M-FZTR

Government of Canada. (2023f). *Consolidated Federal Laws of Canada: Human Pathogens and Toxins Act*. https://perma.cc/MV7P-S2RZ

Government of Canada (2023a). *Plan for Administrative Oversight for Pathogens and Toxins in a Research Setting—Required Elements and Guidance* [Policies]. https://www.canada.ca/en/public-health/services/laboratory-biosafety-biosecurity/licensing-program/plan-administrative-oversight-pathogens-toxins-a-research-setting-required-elements-guidance.html

Gronvall, G. K. (2008, October 29). Improving the Select Agent Program. *Bulletin of the Atomic Scientists*. https://perma.cc/45YQ-W4LL

Guest, O. (2023). *Safeguarding the Safeguards: How Best to Promote Alignment in the Public Interest*. Institute for AI Policy and Strategy. https://perma.cc/QA2W-B4HV

Inglesby, T. (2018). *How to Prevent Global Catastrophic Biological Risks*. 80,000 Hours. https://80000hours.org/podcast/episodes/tom-inglesby-health-security/

Kalra, S., & Parker, M. (2022). The Select Agent Regulations: Structure and Stricture. *Georgetown Scientific Research Journal*, *2*(2), 60–69. https://doi.org/10.48091/gsr.v2i2.46

Kirkpatrick, J., Koblentz, G. D., Palmer, M. J., Denton, S. W., & Tiu, B. (2018). *Biotechnology Risk Assessment: Landscape and Options* [Working Paper]. George Mason University. https://doi.org/10.13021/ad32-4542

Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A., & Gronvall, G. (2019). Information Hazards in Biotechnology. *Risk Analysis*, *39*(5), 975–981. https://doi.org/10.1111/risa.13235

Millett, P. D. (2017). *Gaps in the International Governance of Dual-Use Research of Concern*. https://perma.cc/4EV3-Y9QU

Morse, S. A. (2015). Pathogen Security-Help or Hindrance? *Frontiers in Bioengineering and Biotechnology*, *2*. https://doi.org/10.3389/fbioe.2014.00083

National Research Council. (2010). *Sequence-Based Classification of Select Agents: A Brighter Line*. National Academies Press. https://doi.org/10.17226/12970

Ngo, R., Chan, L., & Mindermann, S. (2023). *The Alignment Problem from a Deep Learning Perspective* (arXiv:2209.00626). arXiv. https://doi.org/10.48550/arXiv.2209.00626

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., & Hendrycks, D. (2023). *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark* (arXiv:2304.03279). arXiv. https://doi.org/10.48550/arXiv.2304.03279

Pannu, J., Palmer, M. J., Cicero, A., Relman, D. A., Lipsitch, M., & Inglesby, T. (2022). Strengthen Oversight of Risky Research on Pathogens. *Science*, *378*(6625), 1170–1172. https://doi.org/10.1126/science.adf6020

Powell, J., Jain, R., & Juneja, A. (2019). The Checklist Manifesto Revisited. *The National Medical Journal of India*, *32*(4), 232. https://doi.org/10.4103/0970-258X.291308

Public Health Agency (PHA) Canada. (2018, June 29). *Canadian Biosafety Guideline—Dual-Use in Life Science Research* [Consultations]. https://www.canada.ca/en/public-health/programs/consultation-biosafety-guideline-dual-use-life-science-research/document.html

Smith, J., Gangadharan, D., Hemphill, M., & Edwin, S. (2023). Review of Restricted Experiment Requests, Division of Select Agents and Toxins, US Centers for Disease Control and Prevention, 2014-2021. *Health Security*, *21*(3), 207–213. https://doi.org/10.1089/hs.2022.0155

Stepanlos. (2023). Biosafety Regulations (BMBL) and their Relevance for AI. *Effective Altruism Forum*. https://forum.effectivealtruism.org/posts/g38CkMbFzKBtdzFXY/biosafety-regulations-bmbl-and-their-relevance-for-ai

Sutton, V. (2009). Survey Finds Biodefense Researcher Anxiety—Over Inadvertently Violating Regulations. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, *7*(2), 225–226. https://doi.org/10.1089/bsp.2009.0015

UK Health and Safety Executive (HSE). (2023a). *What are GMOs?* Retrieved December 14, 2023, from https://perma.cc/HG48-ZNCQ

UK Health and Safety Executive (HSE). (2023b). *The SACGM Compendium of Guidance. Part 2: Risk Assessment of Genetically Modified Microorganisms (Other Than Those Associated with Plants).* https://perma.cc/LB95-FPX2

US Department of Health and Human Services (HHS). (2015, July 28). *Continuing Concerns with the Federal Select Agent Program: Department of Defense Shipments of Live Anthrax*. https://perma.cc/ZA96-GRYL

US Government Accountability Office (GAO). (2017). *High-Containment Laboratories: Coordinated Actions Needed to Enhance the Select Agent Program's Oversight of Hazardous Pathogens*. https://perma.cc/KY5A-ACXU

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

White House. (2023, October 30). *FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*. https://perma.cc/R9UK-TXPB