NATIONAL ACADEMIES *Sciences Engineering Medicine*

Background Briefing Packet

Charting a Responsible Future in AI and Biosecurity: a two-part webinar

November 8 & 15, 2023

Forum on Microbial Threats | Board on Global Health | Health and Medicine Division

Contents

[1] Note: Executive summary only. For full statement, see: https://www.helenabiosecurity.org/
[2] Topic-relevant excerpts only (introduction, chapters 4-5). For full publication, see: https://www.legalpriorities.org/research.html

The National Academies' Statement on Preventing Discrimination, Harassment, And Bullying: Policy for Participants in NASEM Activities (Updated December 2, 2021)

The National Academies of Sciences, Engineering, and Medicine (NASEM) are committed to the principles of diversity, inclusion, integrity, civility, and respect in all of our activities. We look to you to be a partner in this commitment by helping us to maintain a professional and cordial environment. **All forms of discrimination, harassment, and bullying are prohibited in any NASEM activity.** This policy applies to all participants in all settings and locations in which NASEM work and activities are conducted, including committee meetings, workshops, conferences, and other work and social functions where employees, volunteers, sponsors, vendors, or guests are present.

Discrimination is prejudicial treatment of individuals or groups of people based on their race, ethnicity, color, national origin, sex, sexual orientation, gender identity, age, religion, disability, veteran status, or any other characteristic protected by applicable laws.

Sexual harassment is unwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature that creates an intimidating, hostile, or offensive environment.

Other types of harassment include any verbal or physical conduct directed at individuals or groups of people because of their race, ethnicity, color, national origin, sex, sexual orientation, gender identity, age, religion, disability, veteran status, or any other characteristic protected by applicable laws, that creates an intimidating, hostile, or offensive environment.

Bullying is unwelcome, aggressive behavior involving the use of influence, threat, intimidation, or coercion to dominate others in the professional environment.

REPORTING AND RESOLUTION
Any violation of this policy should be reported. If you experience or witness discrimination, harassment, or bullying, you are encouraged to make your unease or disapproval known to the individual at the time the incident occurs, if you are comfortable doing so. You are also urged to report any incident by:

- Filing a complaint with the Office of Human Resources at 202-334-3400 or hrservicecenter@nas.edu, or
- Reporting the incident to an employee involved in the activity in which the member or volunteer is participating, who will then file a complaint with the Office of Human Resources.

Complaints should be filed as soon as possible after an incident. To ensure the prompt and thorough investigation of the complaint, the complainant should provide as much information as is possible, such as names, dates, locations, and steps taken. The Office of Human Resources will investigate the alleged violation in consultation with the Office of the General Counsel.

If an investigation results in a finding that an individual has committed a violation, NASEM will take the actions necessary to protect those involved in its activities from any future discrimination, harassment, or bullying, including in appropriate circumstances **the removal of an individual from current NASEM activities and a ban on participation in future activities.**

CONFIDENTIALITY
Information contained in a complaint is kept confidential, and information is revealed only on a need-to-know basis. NASEM will not retaliate or tolerate retaliation against anyone who makes a good faith report of discrimination, harassment, or bullying.

The National Academies' Statement on Diversity and Inclusion

The National Academies of Sciences, Engineering, and Medicine value diversity in our members, volunteers, and staff and strive for a culture of inclusion in our workplace and activities. Convening a diverse community to exchange ideas and perspectives enhances the quality of our work and increases our relevance as advisers to the nation about the most complex issues facing the nation and the world.

To promote diversity and inclusion in the sciences, engineering, and medicine, we are committed to increasing the diversity of the National Academies' staff, members, and volunteers to reflect the populations we serve. We pledge to cultivate an environment and culture that promotes inclusion and values respectful participation of all individuals who help advance the mission of the institution.

IN THE PIPELINE IN SILICO

Making Up Proteins

30 JAN 2023 • 12:00 AM ET • BY DEREK LOWE • 8 MIN READ • COMMENTS

One of the biggest themes of the last few decades of discovery in biology and chemistry is the constant effort to extract knowledge from the billions of years of evolutionary optimization that we find ourselves surrounded by. It's not easy, because there are no annotations in the code and no documentation left lying around. We're left with the output from untold millennia of "Hey, whatever works". So although the answers to the "Why" and "How" questions come hard, we at least know that when we study living systems in detail we are seeing Things That Have Been Proven To Work.

Protein structure and function (and the underlying RNA and DNA sequences) is a perfect landscape to explore these ideas. Examined closely, we can start to reconstruct how we ended up with the proteins we have (there are several mechanisms at work), and we're busy working out why they have the folds and shapes that they do. As we find it, that collection is very large, but it's a lot smaller than it (theoretically) could be. That's what allows the current protein-structure prediction programs to work as well as they do: proteins end up using this large-but-finite list of motifs over and over again, and they're associated with particular amino acid sequences.

What software like AlphaFold is doing is very much (in human terms) like looking over a protein sequence and saying "OK, I've seen those six or eight amino acids in that order before. . .yeah, that generally makes a turn like this thing here - and when that happens, it can go a couple of different ways. If you get these residues coming up next, it's generally a short spacer to make room for one of those hinky-looking flattish sheet things coming in at an angle, but if it's the other ones, the ones with a proline in the middle, then it's a sign that it's gonna bend around like so instead, and when it does that it means that there's usually another sort-of-matching bend later in the sequence that's going to come around and fit in with it, so I should check for that, too. . ." So just imagine yourself having learned all of those little motifs you could from the existing protein structures and what tends to lead to what and match up with what, and using that knowledge relentlessly and thoroughly at completely inhuman speed and efficiency. And there you are.

But as mentioned above, the number of protein shapes that we have is still nothing compared to the number that could be. So how come we have what we have? Wouldn't you figure that there could be other folds and loops that could also work, but that for some reason evolution just never got around to? Or perhaps there aren't? Maybe protein sequence/structure/function relationships have constraints on them that we don't yet understand? And thus if you ran the whole evolution-of-live thing over and over you'd always wind up with something recognizably like we have now? Obviously, no one knows, and we don't quite have the power as a species to run those experiments, nor the time (nor the funding, come to think of it). But what we can do is try to explore unusual protein geometries and see if they seem to be useful, and shed some light on the problem from that direction.

That's where this new paper comes in. The thing is, making those new protein sequences is a matter for some thought as well. If you just wander out there starting at random and looking for function, you can expect to take a rather long time to discover things. I mean, let's say that you improve on Nature by a thousandfold in speed of experiments: that means that you should have some interesting results in only a few hundred thousand years. So ruling that out, you can start from known functional proteins and start mutating them. But the problem there is that your starting point is already optimized in some direction, and the number of changes you need to make to find new activities might take you through some "activity deserts" where the intermediates are nonfunctional. And that comes down to how you're assaying function, as well. In a living cell, a protein that mutates and loses its initial function is surrounded by huge numbers of possibilities to fit in somewhere else, and occasionally one manages to. But you're not going to be picking that up in a few targeted *in vitro* assays, are you? Another option is to try to

compute your way through and predict new functions *de novo*, but let's be honest. Press releases aside, we really don't have the knowledge to do that yet. In either case, you might well find that most of your hits are things that aren't very far from where you started.

The paper linked above tries to crack this problem using similar technology as used in computational approaches to human language. If you feed vast amount of meaningful text into such a system, it will assemble gargantuan lists of correlations. A sentence starting with "I watered the. . ." is far more likely to end with "lawn" or "houseplants"than it is to end with "leopard" or "harpoon". This is how the predictive-text features on a smartphone messaging platform are working. With a bit more context, these things become even more powerful. The phrase "Cut the sodium. . ." will have a different ending if the context is dietary advice than it will if it's a procedure for a Birch reduction. And this is one part of how the systems like ChatGPT work. If the sample size is large enough, it will have seen human-generated text that has branched off in several directions from that start, and will look for more context to decide whether to go down the ". . .for cardiovascular health" pathway as opposed to the less-common but equally valid ". . .under a layer of inert solvent" one.

The idea of using such language models on protein sequences is certainly not a new one, and it's been applied in several different ways before. But the current paper is trying to see if these algorithms are now robust enough to generate plausibly functional proteins, without knowing what function you have in mind. If you have a deep and varied knowledge of "what amino acid tends to come next" in a huge number of situations, you can presumably generate things that kind of look like they should or could be real proteins, but aren't.

And thus, ProGen, a 1.2-billion parameter neural network trained from a database of 280 million protein sequences, all tagged with some of that extra-context information (protein family, mechanistic and biological functions, etc.) The team tried this out with lysozyme proteins, which are certainly a class that we know a lot about in both structure and function. They generated one million lyzozymish sequences, and then selected 100 of them based on how well the model seemed to generate them and how different they were to known sequences. Their average length (93 to 179 residues) was certainly comparable to known lyzozymes, but they included specific pairwise interactions between amino acids that have never been seen in any natural lysozymes. These were compared to a positive control group, 100 selected from about fifty-six thousand curated lysozymes from the real world.

72% of those natural lysozymes expressed well in cell-free protein synthesis, and 72% of the artificial ones did, too. They then turned these loose on a standard assay of fluorescein-labeled bacterial cell walls, which are engineered to be flourescence-quenched until the structures are broken up by enzymatic action. 59% of the natural lyzozyme proteins met the cutoff for functionality, while 73% of the artificial ones did. Some of those had rather low sequence homology to the natural enzymes, but worked as efficiently as the "real" ones. The different residues are evenly spread across the sequences as well, so it's not like they clustered into "differences that make so difference" regions (i.e., some of the mutations are in the active sites and in other regions that are known to influence high-level conformational changes). Even going back and deliberately picking a new set of sequences that were deliberately chosen for 40% or lower sequence identity to any known lysozymes still produced some active enzymes.

Now, sequence identity is one thing, but that takes us back to something earlier in this post. Perhaps you can get similar overall structures from very different sequences - and that turned out to be the case here. Using AlphaFold to predict the structure of the new artificial sequences showed that they roughly matched known lyzozymes in three dimensions, and that was the case for the low-sequence-identity ones as well. In this case, then, we see that there are far more ways than are known in nature to arrive at more or less the same place, structurally (and functionally).

You're very likely not going to be able to use these techniques, then, to arrive at totally new protein folds doing totally new things. But you can expand what's known about the pathways that evolution didn't take. It'll be interesting to see if some protein classes are more constrained than the lysozymes, for example, and some of them surely are. As an extreme example, consider the photosynthesis protein RuBisCO, which by enzymatic standards just barely seems to work at all and has proven spectacularly difficult to improve by mutation or computational design (but is nonetheless the keystone for most of the life on the surface of the earth). I would not expect to generate a big ol' list of alternate RuBisCOs, because it seems to be wedged into a pretty tight slot already.

This commentary at *Nature* calls the technique "hallucinating functional protein sequences", and that's pretty accurate. I particularly like the use of a phrase from Frances Arnold in her Nobel lecture, that "today we can for all practical purposes read, write, and edit any sequence of DNA, but we cannot compose it". At both the DNA and the protein level, we can sequence

like crazy, so “read” is indeed pretty well taken care of. And thanks to CRISPR and many other editing techniques, we can write fairly well, too. AlphaFold (and RosettaFold, etc.) are doing a good job at turning those letter into structures. But what to write? Having a keyboard in front of you is not sufficient to produce a poem, a novel, (or, I should add, a blog post).

Jorge Luis Borges gave us the vision of the Library of Babel, the huge (but not infinite!) set of all the same-sized books that could be produced from a given set of letters. Everything that can or could be written down is there, every secret and every truth about everything, along with every version with every minor typographical error, every pernicious error and mistake that could be written down about all of them, and every possible commentary on them as well. The perfectly phrase right instructions for finding and learning anything, the ill-phrased ones, the garbled ones, the utterly wrong ones. All there. None of us can write anything down that isn’t in that collection. And there is a Library of Babel of protein sequences, too, where all of that applies in exactly the same way. *(Edit: I was very happy to think of this literary comparison while writing this post, but Frances Arnold got there well before!)*

But to stick with the language metaphor via Borges, we can recognize the letters any of the books in the library we might pick up and read them off in order. We can write down letter combinations and hit “print”. AlphaFold (and RosettaFold, etc.) will take those sequences of letters and recognize the similarities to known laundry lists, airport thrillers, holy texts, tax forms, or sonnets and fit them into the structural categories they know about as well as they can. And what this new ProGen technique will do is to take a bunch of known recipes for (say) pasta sauce or bread rolls and produce a bunch of new ones that might look a bit weird at first inspection, but turn out to produce edible and acceptable pasta sauces or bread when you actually try them out, because there are in the end a lot of ways to get there, far more than chefs have ever actually tried.

But using software to create our own amazing dishes (make me something great with scallops in it that no restaurant has ever served anything close to!), our own rousing anthems (find me a melody line that doesn’t make me think of any other song I’ve ever heard!), or our own proteins (build me an enzyme that does this reaction, although it’s never seen in any living system!) . . .that, we’re still working on. It’s really, really, hard. But it might not be impossible, either.

ABOUT THE AUTHOR

Derek Lowe

Derek Lowe, an Arkansan by birth, got his BA from Hendrix College and his PhD in organic chemistry from Duke before spending time in Germany on a Humboldt Fellowship on his post-doc. He’s worked for several major pharmaceutical companies since 1989 on drug discovery projects against schizophrenia, Alzheimer’s, diabetes, osteoporosis and other diseases.

IN THE PIPELINE

Derek Lowe’s commentary on drug discovery and the pharma industry. An editorially independent blog, all content is Derek’s own, and he does not in any way speak for his employer.

YOU MAY ALSO LIKE

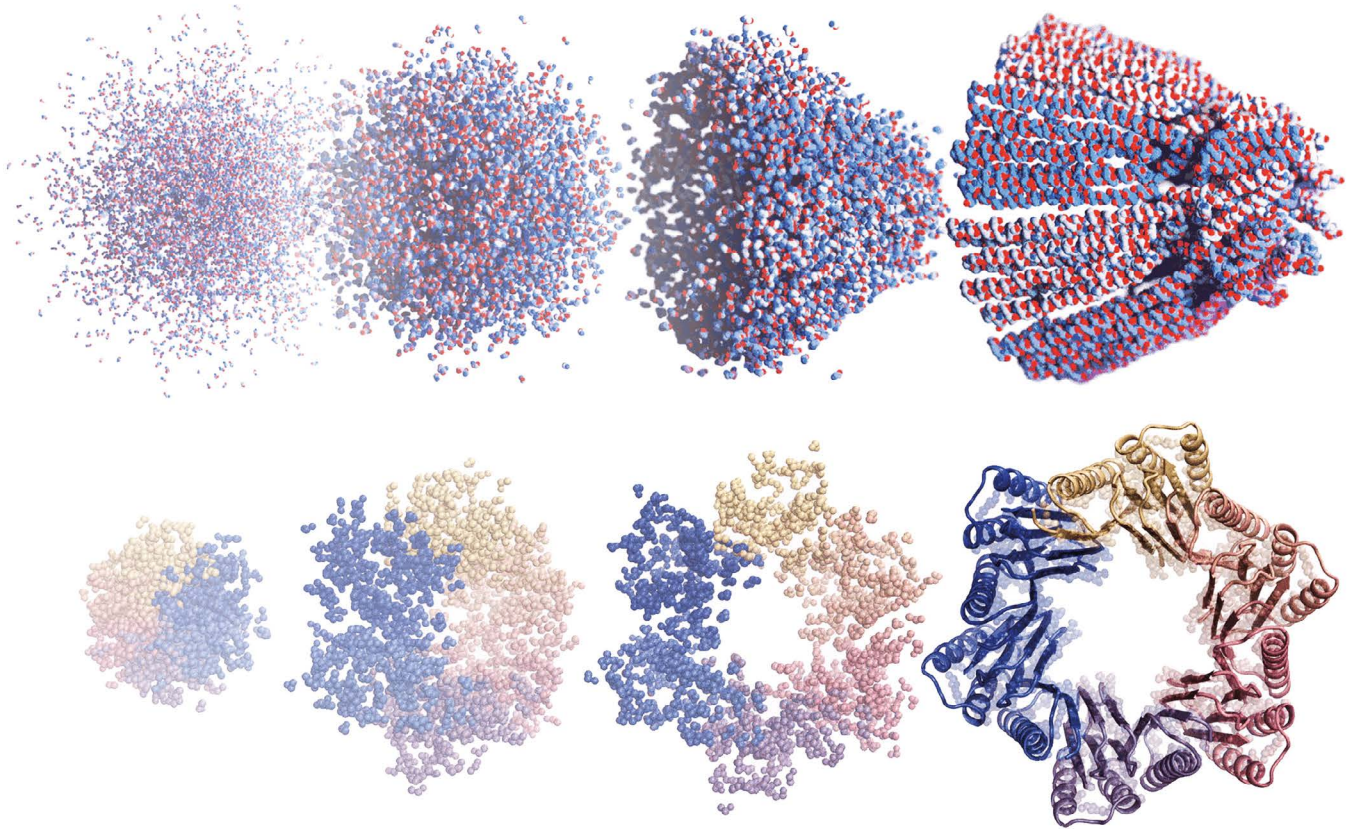
30 MAR 2022 | BY DEREK LOWE
The Uselessness of Phenylephrine

21 SEP 2023 | BY DEREK LOWE
Target Based Drug Discovery - A Waste of Time?

23 FEB 2010 | BY DEREK LOWE
Things I Won't Work With: Dioxygen Difluoride

26 FEB 2008 | BY DEREK LOWE
Sand Won't Save You This Time

VIEW MORE ›



IAN C. HAYDON/UW INSTITUTE FOR PROTEIN DESIGN

Two protein assemblies (right) were developed using an artificial-intelligence tool called RFdiffusion.

'TRANSFORMATIVE' AI DESIGNS CUSTOM PROTEINS ON DEMAND

Computer-devised biomolecules could form the basis of new vaccines or medicines. **By Ewen Callaway**

"OK. Here we go." David Juergens, a computational chemist at the University of Washington (UW) in Seattle, is about to design a protein that, in 3-billion-plus years of tinkering, evolution has never produced.

On a video call, Juergens opens a cloud-based version of an artificial intelligence (AI) tool he helped to develop, called RFdiffusion. This neural network, and others like it, are helping to bring the creation of custom proteins — until recently a highly technical and often unsuccessful pursuit — to mainstream science.

These proteins could form the basis for vaccines, therapeutics and biomaterials. "It's been a completely transformative moment," says Gevorg Grigoryan, the co-founder and chief technical officer of Generate Biomedicines in Somerville, Massachusetts, a biotechnology company applying protein design to drug development.

The tools are inspired by AI software that synthesizes realistic images, such as the Midjourney software that, this year, was famously used to produce a viral image of Pope Francis wearing a designer white puffer jacket. A similar conceptual approach, researchers have found, can churn out realistic protein shapes to criteria that designers specify — meaning, for instance, that it's possible to speedily draw up new proteins that should bind tightly to another biomolecule. And early experiments show that when researchers manufacture these proteins, a useful fraction do perform as the software suggests.

The tools have revolutionized the process of designing proteins in the past year, researchers say. "It is an explosion in capabilities," says Mohammed AlQuraishi, a computational biologist at Columbia University in New York City, whose team has developed one such tool for protein design. "You can now create designs that have sought-after qualities."

"You're building a protein structure customized for a problem," says David Baker, a computational biophysicist at UW whose group, which includes Juergens, developed RFdiffusion. The team released the software in March 2023, and a paper describing the neural network appears this week in *Nature*[1]. (A preprint version was released in late 2022, at around the same time that several other

teams, including AlQuraishi's and Grigoryan's, reported similar neural networks[2,3]).

For the first time, protein designers now have the kinds of reproducible and robust tools around which a new industry can be created, Grigoryan adds. "The next challenge becomes, what do you do with it?"

Grand designs

Juergens inputs a few specifications for the protein he wants into a web form resembling an online tax calculator. It must be 100 amino acids long and form a symmetrical two-protein complex called a homodimer. Many cell receptors adopt this configuration, and a new homodimer could be a synthetic cell-signalling molecule, chimes in Joe Watson, a UW computational biochemist who co-developed RFdiffusion, and is also on the video call. But this morning's design isn't meant to do anything except resemble a realistic protein.

Researchers have struggled for decades to build new proteins. At first, they tried to cobble together useful parts of existing proteins, such as a pocket of an enzyme in which a chemical reaction is catalysed. This approach relied on understanding how proteins fold up and work, as well as intuition and a lot of trial and error. Scientists sometimes screened thousands of designs to identify one that worked as hoped.

A light-bulb moment came with AlphaFold (developed by the London-based AI firm DeepMind, now Google DeepMind) and other AI-based models that could accurately predict protein structures from amino-acid sequences, says Baker. Designers realized that these neural networks, trained on real protein sequences and structures, could also help to create proteins from scratch.

In the past few years, Baker's team and others in the field have released a slew of AI-based protein-design tools (*Nature* **609**, 661–662; 2022). One approach these tools use, called hallucination, involves creating a random string of amino acids that is then optimized by AlphaFold, or a similar tool called RoseTTAFold, until it resembles something that the neural network suggests is likely to fold into a specific structure. Another, called inpainting, takes a specified snippet of a protein sequence or structure and builds the rest of the molecule around it using RoseTTAFold.

But these tools are far from perfect. Experiments tended to show that structures designed by hallucination methods didn't always form well-folded proteins when they were made in the laboratory, and ended up as gunk at the bottom of a test tube, for instance. Hallucination methods also struggled to make anything but small proteins (although other researchers showed, in a February preprint, how the technique could be used to design longer molecules[4]). Inpainting also did a poor job of forming proteins when given shorter snippets. Even when the approach did produce a theoretical protein structure, it wasn't able to come up with diverse solutions to a problem that would increase the odds of success.

That is where RFdiffusion and similar protein-designing AIs, released in recent months, come in. They are based on the same principles as neural networks that generate realistic images, such as Stable Diffusion, DALL-E and Midjourney. These 'diffusion' networks are trained on data, be they images or protein structures, which are then made progressively noisier, eventually bearing no resemblance to the starting image or structure. The network then learns to 'denoise' the data, performing the task in reverse.

Networks such as RFdiffusion are trained on tens of thousands of real protein structures stored in a repository called the Protein Data Bank (PDB). When the network makes a new protein, it begins with total noise: a random assortment of amino acids. "You're asking what is the protein that gave rise to the noise," explains Watson. After rounds of denoising, it produces something resembling a real – but new – protein.

When Baker's team tested RFdiffusion without providing any guidance except the length of the protein, the network generated diverse, realistic-looking proteins, different from anything it had been trained on in the PDB.

"The design process is almost unrecognizable compared to a year ago."

But the researchers are also able to direct the program to make proteins according to specific design constraints during the denoising process, a process called conditioning.

For instance, Baker's team conditioned RFdiffusion to make proteins that include a specific fold, or that can nestle against the surface of another molecule (an interaction that underlies binding). Grigoryan's team even developed a diffusion network called Chroma and then conditioned it to make proteins shaped to resemble the 26 capital letters used in English, as well the Arabic numerals[3].

Signal from noise

Juergens' computer screen initially shows noise, the random assortment of amino acids that the AI system starts with. They are represented as red, smudgy squiggles that resemble a toddler's fingerpainting. They morph, frame by frame, into ever-more-complex shapes, with protein-like features such as tight spirals known as α-helices and ribbony shapes that double back on themselves, called β-sheets. "It's a nice mixed alpha–beta topology," says Juergens, smiling as he admires a creation that took only a few minutes to make. "This is looking good."

The tool has gained widespread use in Baker's laboratory. "The design process is almost unrecognizable compared to a year ago," he says. The neural network has excelled in design challenges that have been inefficient, difficult or impossible using other approaches.

In one analysis reported in their study[1], the researchers started with a snippet from another protein, such as a portion of a viral protein recognized by immune cells, and tasked AI-based tools with churning out 100 different new proteins, to see how many would incorporate the desired motif. The team carried out this challenge for 25 different initial shapes. The results didn't always incorporate the starting snippet, but RFdiffusion produced at least one protein that did for 23 of the motifs, compared with 15 for hallucination and 12 for inpainting.

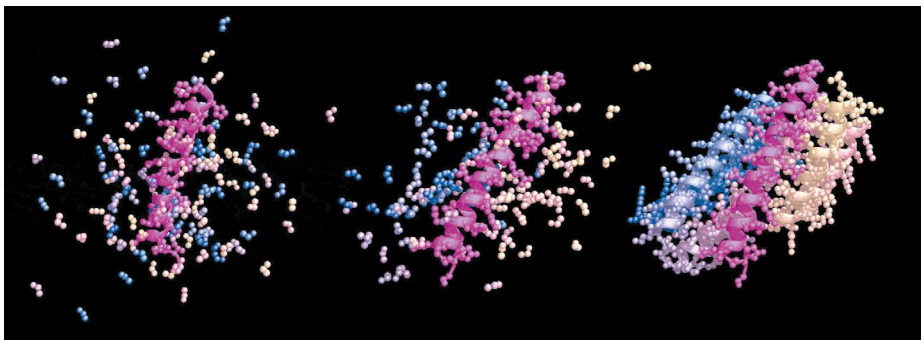
RFdiffusion has also proved adept at making proteins that self-assemble into complex nanoparticles that might be able to deliver drugs or vaccine components. Previous AI approaches[5] can also make these kinds of protein, but Watson says RFdiffusion's designs are much more sophisticated.

Neural networks such as RFdiffusion seem to really shine when tasked with designing proteins that can stick to another specified protein. Baker's team has used the network to create proteins that bind strongly to proteins implicated in cancers, autoimmune diseases and other conditions. One as-yet unpublished success, he says, was to design strong binders for a hard-to-target immune-signalling molecule called the tumour necrosis factor receptor – the target for antibody drugs that generate billions of dollars in revenue each year. "It is broadening the space of proteins we can make binders to and make meaningful therapies" for, Watson says.

Real-world testing

Baker's team is cranking out so many designs that testing whether they work as intended has become a serious bottleneck. "One machine-learning person can generate enough designs to keep 100 biologists busy for months," says Kevin Yang, a biomedical machine-learning researcher at Microsoft Research in Cambridge, Massachusetts whose team has developed its own diffusion-based protein design tool[6].

But early signs suggest that RFdiffusion's creations are the real deal. In another challenge described in their study, Baker's team tasked the tool with designing proteins containing a key stretch of p53, a signalling molecule that is overactive in many cancers (and a sought-after drug target). When the researchers made 95 of the software's designs (by engineering bacteria to express the proteins), more than half maintained p53's ability to bind to its natural target, MDM2. The best designs did so around 1,000 times more strongly



IAN C. HAYDON/UW INSTITUTE FOR PROTEIN DESIGN

RFdiffusion generated a protein that binds to the parathyroid hormone, shown in pink.

than did natural p53. When the researchers attempted this task with hallucination, the designs — although predicted to work — did not pan out in the test tube, says Watson.

Overall, Baker says his team has found that 10–20% of RFdiffusion's designs bind to their intended target strongly enough to be useful, compared with less than 1% for earlier, pre-AI methods. (Previous machine-learning approaches were not able to reliably design binders, Watson says). Biochemist Matthias Gloegl, a colleague at UW, says that lately he has been hitting success rates approaching 50%, which means it can take just a week or two to come up with working designs, as opposed to months. "It's really insane," he says.

The cloud-based version of RFdiffusion had around 100 users each day by late June, according to Sergey Ovchinnikov, an evolutionary biologist at Harvard University in Cambridge, Massachusetts. Joel Mackay, a biochemist at the University of Sydney in Australia, has been dabbling with RFdiffusion to design proteins capable of binding to other proteins that his lab studies, which include molecules called transcription factors that control gene activity in cells. He found the design process simple, and used computer modelling to validate that, in theory, the proteins should bind to the transcription factors.

Mackay is now testing whether the proteins can alter gene expression as intended when they are produced in cells. He has his fingers crossed, because such a finding would amount to a simple way to switch specific transcription factors on and off within cells, instead of using drugs that can take years to identify, if they can be discovered at all. "If this method works reliably for our types of proteins, it would be a total game-changer," he says.

Future improvements

The latest models such as RFdiffusion are a "step change" says Charlotte Deane, an immune informatician at the University of Oxford, UK. But key challenges remain. "What it will do is inspire people to see how far we can push these diffusion methods," she says.

One application that she and other scientists and biotechnology companies are particularly interested in is designing more complex binding proteins such as antibodies, or the protein receptors used by T cells (a type of immune cell). These proteins have flexible loops that interlock with their targets, as opposed to the sandwich-like, flat interfaces that RFdiffusion has excelled at so far. Baker says they are making progress with antibodies.

Ovchinnikov and others say it's challenging, in general, to design biomolecules whose function depends on floppy regions that give them the ability to adopt many different shapes. These are features that have proved difficult to model using AI. "If the problem is, can we bind to something else and inhibit it," says Ovchinnikov, "I think that problem is going to be solved with these methods. But in order to do something more complex, more like what nature does, you need to introduce some flexibility."

Tanja Kortemme, a computational biologist at the University of California, San Francisco, is using RFdiffusion to design proteins that can be used as sensors or as switches to control cells. She says that if a protein's active site depends on the placement of a few amino acids, the AI network does well, but it struggles to design proteins with more-complex active sites, requiring many more key amino acids to be in place — a challenge she and her colleagues are trying to tackle.

"You can really imagine we will be able to write descriptions of a protein and have them synthesized."

Another limitation of the latest diffusion methods is their inability to create proteins that are vastly different from natural proteins, says Yang. That is because the AI systems have been trained only on existing proteins that scientists have characterized, he says, and tend to create proteins that resemble those. Generating more-alien-looking proteins might require a better understanding of the physics that imbues proteins with their function.

That could make it easier to design proteins to carry out tasks no natural protein has ever evolved to do. "There's still a lot of room to grow," Yang says.

The latest protein-design tools have proved to be extremely powerful at creating proteins that can do a particular task — so long as that function can be described in terms of a shape, such as the surface of a protein to bind to, says AlQuraishi. But, he adds, tools such as RFdiffusion aren't yet able to handle other kinds of specifications, such as making a protein that can carry out a particular reaction regardless of its shape — when "you know what you want but you don't know what the geometry is".

Future protein-design tools will also need the capacity to churn out proteins to numerous different criteria, says Grigoryan. A potential therapeutic protein must not only bind to its target, but also not bind to others and should possess properties that make it easy to mass-produce.

One direction that researchers are exploring is whether proteins could be designed using plain language text descriptions, similar to the prompts fed to image-generation tools such as Midjourney. "You can really imagine we will be able to write descriptions of a protein and have them synthesized and tested," says Watson.

Grigoryan and his colleagues have taken a step towards this goal. In their December 2022 preprint[3], they trained Chroma to attach descriptions to its designs and spit out designs to text-based specifications, including 'protein with a CHAD domain' (a protein shape incorporating multiple helices) or 'crystal structure of aminotransferases' (enzymes involved in making and breaking down proteins).

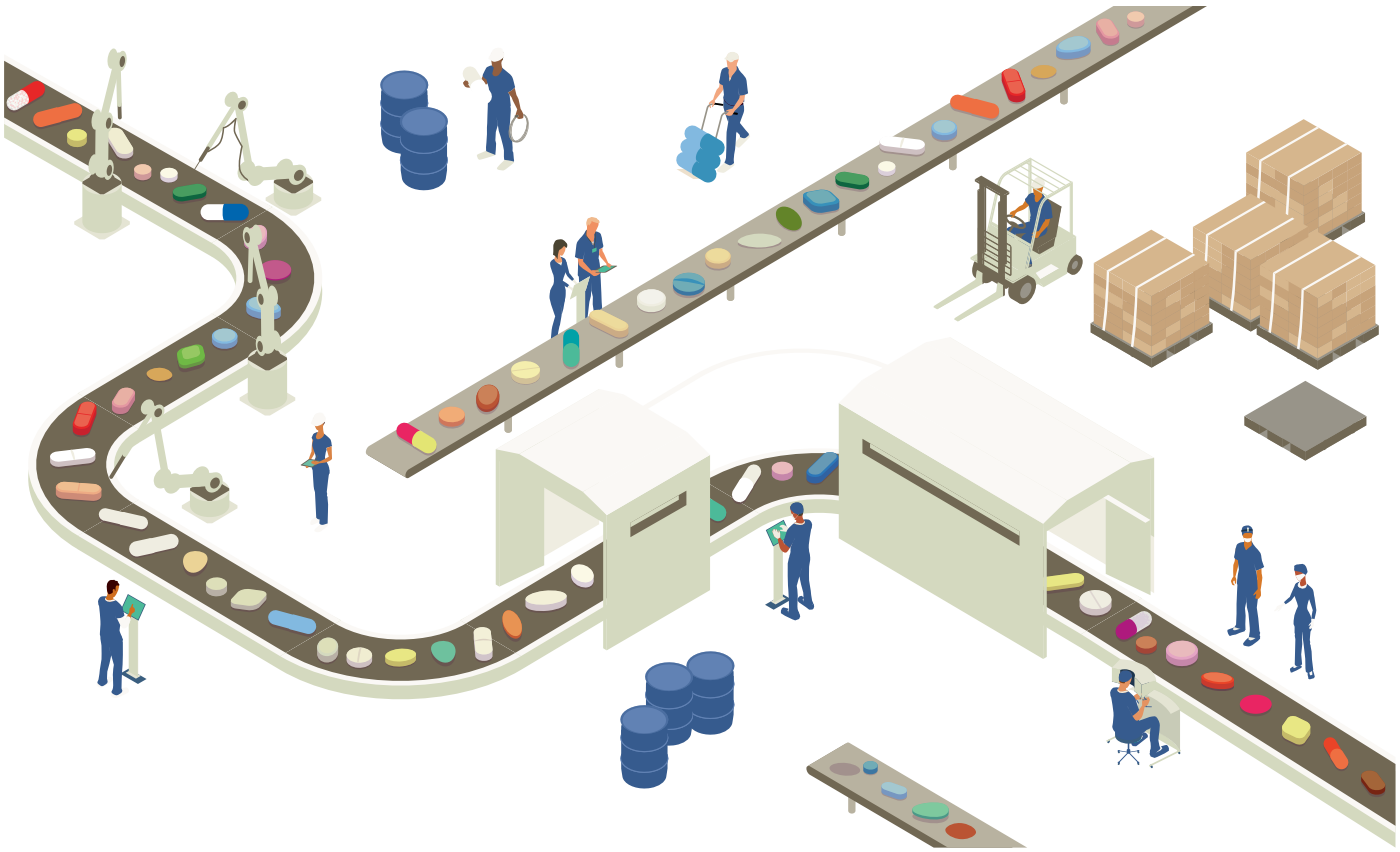
The protein Juergens created in a few minutes this morning is only a model of a protein's 3D structure. Juergens then uses another AI tool to come up with sequences of amino acids that should fold up into that structure. As a final check, he plugs the sequences into AlphaFold to see whether the software predicts folded structures that match the design. They're spot on, with the AlphaFold predictions differing from the design by an average of just 1 ångström (the width of a hydrogen atom).

"This is at the accuracy that we would class as a design success," says Watson. The only thing left to do, he says, is to see how the protein performs in real life.

Ewen Callaway is a senior reporter for *Nature* in Bristol, UK.

1. Watson, J. L. *et al. Nature* https://doi.org/10.1038/s41586-023-06415-8 (2023).
2. Lin, Y. & AlQuraishi, M. Preprint at https://arxiv.org/abs/2301.12485 (2023).
3. Ingraham, J. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2022.12.01.518682 (2022).
4. Frank, C. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2023.02.24.529906 (2023).
5. Wicky, B. I. M. *et al*. *Science* **378**, 56–61 (2022).
6. Wu, K. E. Preprint at https://arxiv.org/abs/2209.15611 (2022).

https://doi.org/10.1038/s41591-023-02361-0



INSIDE THE NASCENT INDUSTRY OF AI-DESIGNED DRUGS

Artificial intelligence tools are beginning to upend the drug discovery pipeline, with several new compounds entering clinical trials. **By Carrie Arnold**

Drug discovery is expensive, inefficient, and fraught with failure. An estimated 86% of drug candidates developed between 2000 and 2015 did not meet their stated endpoints.

Despite this challenge, the use of artificial intelligence (AI) and machine learning to understand drug targets better and synthesize chemical compounds to interact with them has not been easy to sell. Alex Zhavoronkov would know. When the CEO and founder of Insilico Medicine, with offices in Hong Kong and New York, first started trying to raise funding nearly a decade ago, he struggled to find others who shared his vision.

"It was such a grand goal, but every time I went to a venture capitalist, they never gave me money," says Zhavoronkov.

Even as recently as 5 years ago, his presentations had to explain to pharma collaborators why AI was so promising. Not anymore. Now he is at the forefront of drug discovery's AI nascent revolution.

"We've managed to get here in three years, and we didn't fail. And we did it multiple times," Zhavoronkov says.

The persistence of Zhavoronkov and a small cadre of other startup founders, including Exscientia's Andrew Hopkins and BenevolentAI's Bryn Williams-Jones, means that not only are some of the biggest players in pharma already convinced of the utility of AI in drug development, but also some of these drugs are beginning their ultimate test in clinical trials (Table 1).

"In the last couple of years, AI has gone from being hypothetically interesting to real programs moving towards the clinic," says Williams-Jones. "There's no shortcuts to drug discovery. We can have better informed ideas, but you still have to go through the rest of the [development] process."

These trials are still in their early days, says Hopkins, so it is not yet clear which compound will cross the finish line first. But he is

Table 1 | Selected AI-designed drugs in or entering clinical trials

Treatment	Organization	Description	Phase	Lead indication
REC-2282	Recursion	Small molecule pan-HDAC inhibitor	2/3	Neurofibromatosis type 2
REC-994	Recursion	Small molecule superoxide scavenger	2	Cerebral cavernous malformation
REC-4881	Recursion	Small molecule inhibitor of MEK1 and MEK2	2	Familial adenomatous polyposis
INS018_055	InSilico Medicine	Small molecule inhibitor	2	Idiopathic pulmonary fibrosis
BEN-2293	BenevolentAI	Topical pan-tyrosine kinase inhibitor	2a	Atopic dermatitis
EXS-21546	Exscientia and Evotec	A_{2A} receptor antagonist	1b/2	Solid tumors carrying high adenosine signatures.
RLY-4008	Relay Therapeutics	Inhibitor of FGFR2	1/2	FGFR2-altered cholangiocarcinoma
EXS-4318	Exscientia	PKC-θ inhibitor	1/2	Inflammatory and autoimmune conditions
BEN-8744	BenevolentAI	Small molecule PDE10 inhibitor	1	Ulcerative colitis
Undisclosed	Recursion	Small molecular inhibitor of RBM39, a CDK12-associated protein	Pre-clinical	HRD-negative ovarian cancer

confident that the use of AI is leaving an indelible mark on drug development and promises to make the process better, faster, and cheaper, as well as enabling the development of more first-in-class compounds.

"We expect this year to see some major advances in the number of molecules and approved drugs produced by generative AI methods that are moving forward," Hopkins says.

Entering trials

As AI-designed drugs enter clinical trials, pharma companies can see how their new compounds are paying off. The preliminary read-outs look promising. In June 2022, Exscientia announced preliminary results from a phase 1 trial of EXS-21546, a highly selective A_{2A} receptor antagonist developed with Germany's Hamburg-based Evotec. The small molecule has subsequently entered phase 1b/2 trials for patients with solid tumors carrying high adenosine signatures.

Exscientia's next AI-developed candidate, a small molecule called EXS4318, is not far behind. A selective protein kinase C-theta (PKC-θ) inhibitor, designed for inflammatory and autoimmune conditions, EXS4318 has been licensed to Bristol Myers Squibb in a partnership worth up to US$1.2 billion, according to a company press release. The company has 16 other AI-designed drugs in its pipeline, including drugs for COVID-19, tuberculosis, malaria, and hypophosphatasia – a rare, inherited disorder that affects bones and teeth.

"It's not just about using generative AI to help us to precision design an exact molecule," Hopkins says, "but also actually helping us precision design which patients are responders and non-responders."

What this would look like in practice, Hopkins says, is performing deep, multi-omics (single-cell proteomics, transcriptomics, and genomics) analyses of participants before the trial starts to identify multi-gene signature biomarkers. This will help the researchers to determine which participants are most likely to respond – and why. At the end of the trial, Exscientia will be able to go to regulators with a drug that consistently works well in a very defined patient population.

"This is where AI is going to lead as well. It's not just about using AI to make drug discovery better, but about how we can create better drugs overall," Hopkins says.

In January 2023, Insilico Medicine announced an encouraging topline readout of its phase 1 safety and pharmacokinetics trial of INS018_055, designed by AI for idiopathic pulmonary fibrosis, a progressive disease that causes scarring of the lungs. Their proprietary AI platforms identified a new target (which Zhavoronkov would identify only as 'target X') and a small molecule inhibitor, which was granted breakthrough status by the Food and Drug Administration (FDA) in February.

"It's the first time anyone in our industry has developed a novel target of a molecule, and completed phase one trials, all the way with AI," Zhavoronkov says. He expects phase two readouts in the first half of 2023. It is part of Insilico's growing pipeline targeting diseases associated with aging. What makes Insilico's work more impressive, according to Zhavoronkov, is that the company only began development on INS018_055 in February 2021.

"We have 31 therapeutic programs. In 2020, we had zero," says Zhavoronkov.

AI for analysis

Recursion, a biopharma startup based in Salt Lake City, Utah, uses AI not to design molecules but to analyze data from millions of experiments and billions of microscopy images that their lab is gathering with the help of robots.

"Just like Google has all these cars driving around taking pictures that they turn into really useful maps for all of us, we've done the same thing with biology," says Chris Gibson, Recursion's co-founder and CEO.

Recursion is also working to develop a therapeutic agent for ovarian cancer that targets a gene that their AI systems indicated was part of the same pathway as CDK12, an existing target that has proved challenging to inhibit directly. In preclinical studies that target the CDK12-associated protein, 40% of mice showed a complete response. When the compound was paired with a PARP inhibitor, tumors were eliminated in four out of five mice. The company also has three other compounds in clinical trials for oncology and rare diseases: familial adenomatous polyposis, cerebral cavernous malformation, and neurofibromatosis type 2.

"Biology and chemistry are so broad and complex. Your goal isn't to find everything. Your goal is to find something really good and advance it," Gibson says.

Relay Therapeutics has developed an oral, small molecule inhibitor of FGFR2, a receptor tyrosine kinase that is overactive in certain cancers, such as intrahepatic cholangiocarcinoma. Existing FGFR inhibitors are not very selective, but the company is testing RLY-4008, which is only active against FGFR2. At the end of 2022, BenevolentAI completed a phase 2a trial for BEN2293, a topical

ointment for the treatment of atopic dermatitis (eczema). The treatment was found to be safe but did not meet its secondary endpoint of reducing itch and inflammation, according to a company press release in April 2023.

BenevolentAI has also filed a clinical trial application with the UK Medicines and Healthcare Products Regulatory Agency (MHRA) for BEN-8744, a small molecule phosphodiesterase 10 (PDE10) inhibitor designed to treat ulcerative colitis. If approved, Williams-Jones says BenevolentAI plans on beginning a phase 1 trial in the first half of 2023. But for BenevolentAI, as for everyone else, he points out this is still early days.

"Biology is hard, and we don't know very much in real terms," says Williams-Jones. Every time scientists think that they have made a big step forward in simplifying the drug development process, he says, they stumble across two or three other issues that they did not expect.

The protein folding problem

Much of AI-driven drug discovery builds on protein folding. By the latter half of the twentieth century, biochemists had decoded some of the basics of protein structure tenets that now fill biology textbooks. A string of amino acids, proteins fold into complex, three-dimensional structures based on the atomic interactions between the backbone and amino acid side chains. This structure determines the protein's function. As crystallography and electron microscopy began to crack open the atomic-level structures of proteins, biochemists began to wonder whether it might be possible to predict the final structure of a protein complex using only its amino acid sequence. The discovery of α-helices and β-sheets in the 1960s made the promise seem almost tractable.

Then reality began to sink in. Twenty simple amino acid building blocks could give rise to a dizzying array of proteins – greater than the number of stars in the universe, Baker says. Methods such as multiple sequence alignment (MSA) enabled structural bioinformatics experts to compare the amino acid sequences of numerous protein homologues to determine domains, disordered regions, and other elements of local secondary structure. But even the most advanced MSA methods could not reveal allosteric interactions, or how different α-helix regions were arranged next to each other.

AI and machine learning took a completely different approach. "Machine learning is based on the results you attain rather than a statistical model that describes the population," Deane says. "It's about finding predictive patterns in the data."

Instead of applying the laws of physics to every single atom or bond, what if scientists began to look for similarities between proteins? If they could assemble a reasonably broad base of protein structures (gathered the old-fashioned way, through painstaking crystallography, X-ray diffraction, and electron microscopy techniques), then perhaps scientists could try to figure out the similarities between proteins and use that to predict a protein's structure.

"With deep learning, you don't really try and simulate the actual folding process. You're not trying to find the lowest energy state. It's more about pattern recognition," Baker says.

The intellectual leap to this way of thinking was profoundly important, says Alan Lipkus, senior data analyst at Chemical Abstracts Service in Columbus, Ohio.

Weird molecules

By the early 2010s, computer scientists and computational chemists had developed the prototypes of groundbreaking AI systems such as RoseTTAFold and DeepMind's AlphaFold. Most modern machine learning algorithms devoted to predicting protein structure contain four different modules: an input module that contains the amino acid sequence and structures from homologous proteins; a sophisticated neural network that uses pattern recognition algorithms to transform the amino acid sequence into spatial information of the protein; an output module that converts the spatial information into a preliminary three-dimensional structure; and a refinement process that enables fine-tuning. Using these algorithms, AlphaFold2 can predict single protein domain structures down to 2.1 Å, essentially solving the protein structure problem. It is a staggering accomplishment, Baker says, but he wants to move beyond it.

"By just predicting protein structure, you're stuck with whatever exists in nature. You can't make anything new. But now we can make all these brand-new proteins for cancer therapeutics and clinical trials. You can make all kinds of different things with protein design," Baker says.

Beyond the basic science accomplishment, these advances have also given a huge leg up to pharma. Determining a protein's structure was a major hurdle in designing the right molecule to alter its function. Determining the structure of a small molecule was simple compared to a protein. Even biologics designed by AI were a possibility, antibodies just being one specific type of protein. This progress did not remove the need for experiments and tinkering – no computer algorithm is yet that good – but it narrowed down the number of possibilities to help scientists prioritize molecules that were far more likely to have the desired effect without causing undue toxicity.

The molecules these AI systems helped to design, however, looked very different from compounds designed by medicinal chemists. When InSilico's Zhavoronkov began pitching his AI therapeutic design service to pharma companies, he included examples of several molecules his system had built. Their novelty immediately grabbed the attention of potential pharma partners, some of whom helped provide series A and B funding rounds.

"They said to me: Alex, these molecules look weird. Tell us how you did it," Zhavaoronkov says. "We did something in chemistry that humans could not do."

And it is this weirdness that just might be AI's biggest strength in pharmacology. Although the total number of possible chemicals in the universe – what some scientists refer to as chemical space – is vast, humans have only explored tiny slivers of this space. Synthetic chemists develop expertise working with certain types of compound or performing specific reactions, says Lipkus, leading to a few small areas of chemical space that are well mapped out. Most of chemical space remains terra incognita.

Many clinical trials test tweaks of existing drugs, which may give a slightly improved safety or efficacy. However, a much bigger prize is a first-in-class drug against an entirely new target, which AI-designed drugs are well-positioned for.

Lipkus and his colleague Todd Wills (now a senior vice president at Cass Information Systems) analyzed the novelty and creativity of pharmaceutical molecules using the chemical abstract service database of thousands of molecules, which "is probably the best representation of they known chemical universe", Lipkus says. They compared the uniqueness of a molecule's scaffold and shape, which they defined as the atom-to-atom connectivity that prunes back all but the most basic information about a compound's structure. 'Me too' drugs, they pointed out, tend to consist of small alterations to a drug's chemical side chains rather than large-scale shifts in molecular structure. A growing number of pharmaceutical compounds, they pointed out in a 2019 paper in the *Journal of Organic Chemistry*, are showing signs of creativity, with more unique

structures and scaffolds. AI, Lipkus says, will only accelerate this trend.

"It's one more piece of evidence that there's value in looking for novel structures," Lipkus says. "Talking to people in the drug industry, they want to break away from these scaffolds that have been used so heavily."

AI tools also enable drug developers to explore the chemical world much more quickly.

"It allows us to explore a much broader slot or chemical space than we'd be able to using experimental methods on their own," says Don Bergstrom, president of research and development at Relay Therapeutics.

Neglected diseases

AI-designed drugs are not just being developed for potential blockbuster status. In Geneva, Switzerland, the Drugs for Neglected Diseases Institute (DNDi) is using machine learning to create better drugs for conditions that predominantly affect the world's poor, such as Chagas disease and dengue fever. Charles Mowbray, discovery director at DNDi, says the institute is also turning to AI strategies to guide its drug repurposing pipeline as part of its global efforts to develop therapies for neglected diseases. For such diseases, speed is critical; AI can help scientists generate hypotheses and test them more quickly.

"These tools don't replace a scientist, they complement them," Mowbray says. "[AI] enables them to have all the information at their fingertips, to ask the good questions, to refine their queries, and to iterate until they can figure out what they're really after." This synergy is true for machine learning across drug development, he adds.

Even as the impacts of AI in drug design are beginning to emerge in clinical trials, these strategies are joining other AI tools in clinical trial design, manufacturing, and more. There is no doubt that machine learning is profoundly reshaping the pharmaceutical industry, Lipkus says. As for how the effects of AI-developed drugs will play out, he is more circumspect, saying that is still up in the air.

"Nothing guarantees anything. Drug discovery is really difficult. I don't know if people expect AI to just pop out the design of a molecule that's your next blockbuster, says Lipkus. "It's all kind of a crapshoot."

Carrie Arnold
Science writer, Richmond, VA, USA.

Published online: 1 June 2023

HHS Public Access

Author manuscript

Nat Mach Intell. Author manuscript; available in PMC 2023 March 07.

Published in final edited form as:

Nat Mach Intell. 2022 March ; 4(3): 189–191. doi:10.1038/s42256-022-00465-9.

Dual Use of Artificial Intelligence-powered Drug Discovery

Fabio Urbina[1], **Filippa Lentzos**[2], **Cédric Invernizzi**[3], **Sean Ekins**[1]

[1]Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC 27606, USA.

[2]Department of Global Health & Social Medicine, King's College London, United Kingdom.

[3]Spiez Laboratory, Federal Department of Defence, Civil Protection and Sports, Switzerland.

Abstract

An international security conference explored how artificial intelligence (AI) technologies for drug discovery could be misused for de novo design of biochemical weapons. A thought experiment evolved into a computational proof.

The Swiss Federal Institute for NBC-Protection—Spiez Laboratory—is part of the 'convergence' conference series[1] set up by the Swiss government to identify developments in chemistry, biology and enabling technologies, which may have implications for the Chemical and Biological Weapons Conventions. Meeting every two years, the conference brings together an international group of scientific and disarmament experts to explore the current state of the art in the chemical and biological fields and their trajectories, to think through potential security implications, and to consider how these implications can most effectively be managed internationally. The meeting convenes for three days of discussion on the possibilities of harm, should the intent be there, from cutting edge chemical and biological technologies. Our drug discovery company received an invitation to contribute a presentation on how AI technologies for drug discovery could be potentially misused.

Risk of misuse

The thought had never struck us. We were vaguely aware of security concerns around work with pathogens or toxic chemicals, but that did not relate to us; we primarily operate in a virtual setting. Our work is rooted in building machine learning models for therapeutic and toxic targets to better assist in the design of new molecules for drug discovery. We have spent decades using computers and AI to improve human health—not to degrade it. We were naïve in thinking about the potential misuse of our trade, as our aim had always been to avoid molecular features that could interfere with the many different classes of proteins essential to human life. Even our projects on Ebola and neurotoxins, which could have sparked thoughts about the potential negative implications of our machine learning models, had not set our alarm bells ringing.

Competing interests

F.U. and S.E. work for Collaborations Pharmaceuticals, Inc. F.L. and C.I. have no conflicts of interest.

Our company—Collaborations Pharmaceuticals, Inc—had recently published computational machine learning models for toxicity prediction in different areas, and, in developing our presentation to the Spiez meeting, we opted to explore how AI could be used to design toxic molecules. It was a thought exercise we had not considered before that ultimately evolved into a computational proof of concept for making biochemical weapons.

Generation of new toxic molecules

We had previously designed a commercial *de novo* molecule generator which we called MegaSyn[2] which is guided by machine learning model predictions of bioactivity for the purpose of finding new therapeutic inhibitors of targets for human diseases. This generative model normally penalizes predicted toxicity and rewards predicted target activity. We simply proposed to invert this logic using the same approach to design molecules *de novo*, but now guiding the model to reward both toxicity and bioactivity instead. We trained the AI with molecules from a public database using a collection of primarily drug-like molecules (that are synthesizable and likely to be absorbed) and their bioactivities. We opted to score the designed molecules with an organism-specific lethal dose (LD_{50}) model[3], and a specific model using data from the same public database which would ordinarily be used to help derive compounds for treatment of neurological diseases (details of the approach are withheld but were available during the review process). The underlying generative software is built on and similar to other open-source software that is readily available[4]. To narrow the universe of molecules we chose to drive the generative model towards compounds like the nerve agent VX, one of the most toxic chemical warfare agents developed during the 20^{th} century—a few salt-sized grains of VX, (6–10 mg)[5], is sufficient to kill a person. Nerve agents such as Novichoks have also been in the headlines recently[6].

In less than 6 hours after starting on our in-house server, our model generated forty thousand molecules that scored within our desired threshold. In the process, the AI designed not only VX, but many other known chemical warfare agents that we identified through visual confirmation with structures in public chemistry databases. Many new molecules were also designed that looked equally plausible. These new molecules were predicted to be more toxic based on the predicted LD_{50} in comparison to publicly known chemical warfare agents (Figure 1). This was unexpected as the datasets we used for training the AI did not include these nerve agents. The virtual molecules even occupied a region of molecular property space that was entirely separate to the many thousands of molecules in the organism-specific LD_{50} model, which is mainly made up of pesticides, environmental toxins, and drugs (Figure 1). By inverting the use of our machine learning models, we had transformed our innocuous generative model from a helpful tool of medicine to a generator of likely deadly molecules.

Our toxicity models were originally created for use in avoiding toxicity, enabling us to better virtually screen molecules (for pharmaceutical and consumer product applications) before ultimately confirming their toxicity through *in vitro* testing. The inverse, however, has always been true: the better we can predict toxicity, the better we can steer our generative model to design new molecules in a region of chemical space populated by predominantly lethal molecules. We did not assess the virtual molecules for synthesizability or explore how

to make them with retrosynthesis software. Both of these processes have readily available commercial and open-source software, which can be easily plugged into the *de novo* design process of new molecules[7]. We also did not physically synthesize any of the molecules either, but with a global array of hundreds of commercial companies offering chemical synthesis, it is not necessarily too big of a step, which is poorly regulated with few if any checks to prevent synthesis of new extremely toxic agents that could potentially be used as chemical weapons. Importantly, we had a human-in-the-loop with a firm moral and ethical 'don't-go-there' voice to intervene. But what if the human was removed or replaced with a bad actor? With current breakthroughs and research into autonomous synthesis[8], a complete design-make-test cycle applicable to making not only drugs, but toxins, is within reach. Our proof-of-concept highlights how a non-human autonomous creator of a deadly chemical weapon is entirely feasible.

A wake-up call

Without being overly alarmist, this should serve as a wake-up call for our colleagues in the 'AI in drug discovery' community. While some domain expertise in chemistry or toxicology is still required to generate toxic substances or biological agents that can cause significant harm, when these fields intersect with machine learning models, where all you need is the ability to code and to understand the output of the models themselves, they dramatically lower technical thresholds. Open source machine learning software is the primary route for learning and creating new models like ours, and toxicity datasets[10] that provide a baseline model for predictions for a range of targets related to human health are readily available.

Our proof of concept was focused on VX-like compounds, but it is equally applicable to other toxic small molecules with similar or different mechanisms with minimal adjustments to our protocol. Retrosynthesis software tools are also improving in parallel, allowing new synthesis routes to be investigated for known and unknown molecules. It is therefore entirely possible that novel routes can be predicted for chemical warfare agents, circumventing national and international lists of watched or controlled precursor chemicals for known synthesis routes.

The reality is that this is not science fiction. We are but one very small company in a universe of many hundreds of companies using AI software for drug discovery and *de novo* design. How many of them have even considered repurposing, or misuse, possibilities? Most will work on small molecules and many of the companies are very well funded and likely using the global chemistry network to make their AI designed molecules. How many people are familiar with the know-how to find the pockets of chemical space that can be filled with molecules predicted to be orders of magnitude more toxic than VX? We do not currently have answers to these questions. There has not previously been significant discussion in the scientific community about this dual use concern of AI used for *de novo* molecule design, at least not publicly. Discussion of societal impact of AI has principally focused on aspects like safety, privacy, discrimination and potential criminal misuse[10], but not national and international security. When we think of drug discovery, we normally do not consider technology misuse potential. We are not trained to consider it, and it is not even required for machine learning research, but we can now share our experience with other

companies and individuals. AI generative machine learning tools are equally applicable to larger molecules (peptides, macrolactones etc.) and to other industries like consumer products and agrochemicals that also have interests in designing and making new molecules with specific physicochemical and biological properties. This greatly increases the breadth of the potential audience that should be paying attention to these concerns.

For us, the genie is out of the medicine bottle when it comes to repurposing our machine learning. We must now ask: what are the implications? Our own commercial tools as well as open-source software tools and many datasets that populate public databases are available with no oversight. If the threat of harm, or actual harm, occurs with ties back to machine learning, what impact will this have on how this technology is perceived? Will hype in the press on AI-designed drugs suddenly flip to AI-designed toxins, public shaming, and decreased investment in these technologies? As a field, we should open a conversation on this topic. The reputational risk is substantial; it only takes one bad apple that takes what we have vaguely described to the next logical step, or an adversarial state looking for a technological edge. How do we prevent this? Can we lock away all the tools and throw away the key? Do we monitor software downloads or restrict sales to certain groups? We could follow the example of machine-learning models like GPT-3[11] which was initially waitlist restricted to prevent abuse and has an API for public usage. Even today, without a waitlist, GPT-3 has safeguards in place to prevent abuse, Content Guidelines, a free content filter and monitoring of applications that use GPT-3 for abuse. We know of no recent toxicity or target model publications that discuss these concerns of dual use, similarly. As responsible scientists, we need to ensure that misuse of AI is prevented, and that the tools and models we develop are only used for good.

By going as close as we dared, we have still crossed a grey moral boundary, demonstrating that designing virtual potential toxic molecules is possible without much effort, time or computational resources. We can easily erase the thousands of molecules we created, but we cannot delete the knowledge of how to recreate them.

The broader impacts on society

There is a need for discussions across traditional boundaries and multiple disciplines to allow for a fresh look at AI for *de novo* design and related technologies from different perspectives and with a wide variety of mindsets. Here, we give some recommendations which we believe will reduce potential dual-use concerns for AI in drug discovery. Scientific conferences, like the Society of Toxicology and American Chemical Society, for example should actively foster a dialogue among experts from industry, academia and policy making on the implications of our computational tools. There has been recent discussion in this journal regarding requirements for broader impact statements from authors submitting to conferences, institutional review boards and funding bodies as well as addressing potential challenges[12]. Making increased visibility a continuous effort and a key priority would greatly assist in raising awareness about potential dual use aspects of cutting-edge technologies and would generate the outreach necessary to have everyone active in our field engage in responsible science. We can take inspiration from examples such as The Hague Ethical Guidelines[13], which promote a culture of responsible conduct in the chemical

sciences and guard against the misuse of chemistry, in order to have AI-focused drug discovery, pharmaceutical, and possibly other companies agree to a code of conduct to train employees, secure their technology and prevent access and potential misuse. The use of a public-facing API for models with code and data available upon request would greatly enhance the security and control over how published models are utilized without adding much hindrance to accessibility. While MegaSyn is a commercial product and thus we have control over who has access to it, going forward, we will implement restrictions or an API for any forward-facing models. A reporting structure or hotline to authorities, if there is a lapse or if we become aware of anyone working on developing toxic molecules for non-therapeutic uses, may also be valuable. Finally, universities should redouble their efforts in ethical training of science students and broaden the scope to other disciplines, particularly computing students, so that they are aware of the potential misuse of AI from an early stage of their career as well as understand the potential for broader impact[12]. We hope that by raising awareness of this technology we will have gone some way to demonstrating that while AI can have important applications for healthcare and other industries, we should also remain diligent against the potential for dual use, in the same way that we would with physical resources such as molecules or biologics.

Acknowledgments

We are grateful to the organizers and participants of the Spiez Convergence conference 2021 for their feedback and questions.

Cédric Invernizzi contributed to this article in his personal capacity. The views expressed in this article are those of the authors only and do not necessarily represent the position or opinion of Spiez Laboratory or the Swiss Government.

Funding

We kindly acknowledge NIH funding from R44GM122196-02A1 from NIGMS and 1R43ES031038-01 and 1R43ES033855-01 from NIEHS for our machine learning software development and applications. “Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R43ES031038 and 1R43ES033855-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.”

References

1. <https://www.spiezlab.admin.ch/en/home/meta/refconvergence.html> (2021).
2. Urbina F, Lowden CT, Culberson JC & Ekins S <10.33774/chemrxiv-2021-nlwvs> (2021).
3. Mansouri K et al. Environ Health Perspect 129, 79001, (2021). [PubMed: 34242083]
4. Blaschke T et al. J Chem Inf Model 60, 5918–5922, (2020). [PubMed: 33118816]
5. Anon <https://www.ncbi.nlm.nih.gov/books/NBK233724/> (1997).
6. Aroniadou-Anderjaska V, Apland JP, Figueiredo TH, De Araujo Furtado M & Braga MF Neuropharmacology 181, 108298, (2020). [PubMed: 32898558]
7. Genheden S et al. J Cheminform 12, 70, (2020). [PubMed: 33292482]
8. Coley CW et al. Science 365, eaax1566, (2019). [PubMed: 31395756]
9. Dix DJ et al. Toxicol Sci 95, 5–12, (2007). [PubMed: 16963515]
10. Hutson M <https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai> (2021).
11. Brown TB et al. <https://arxiv.org/abs/2005.14165> (2020).
12. Prunkl CEA et al. Nature Machine Intelligence 3, 104–110, (2021)
13. .13. <https://www.opcw.org/hague-ethical-guidelines> (2021).

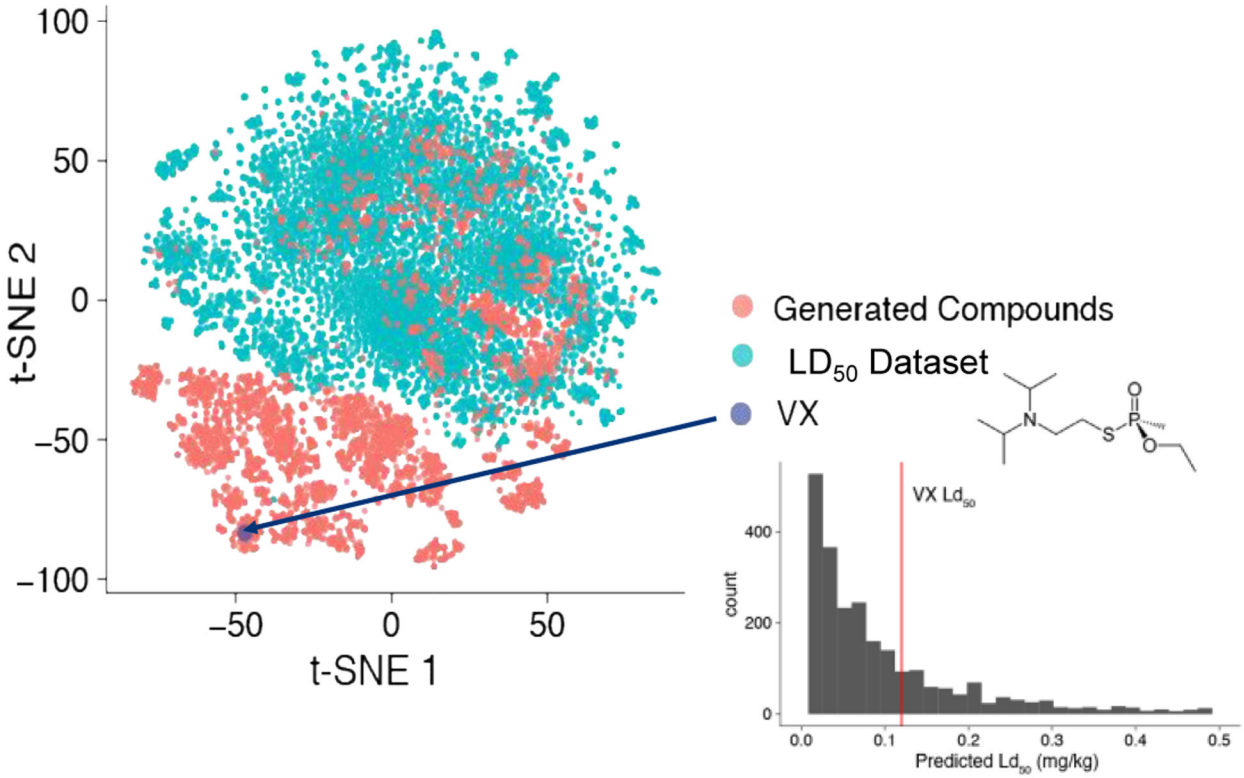
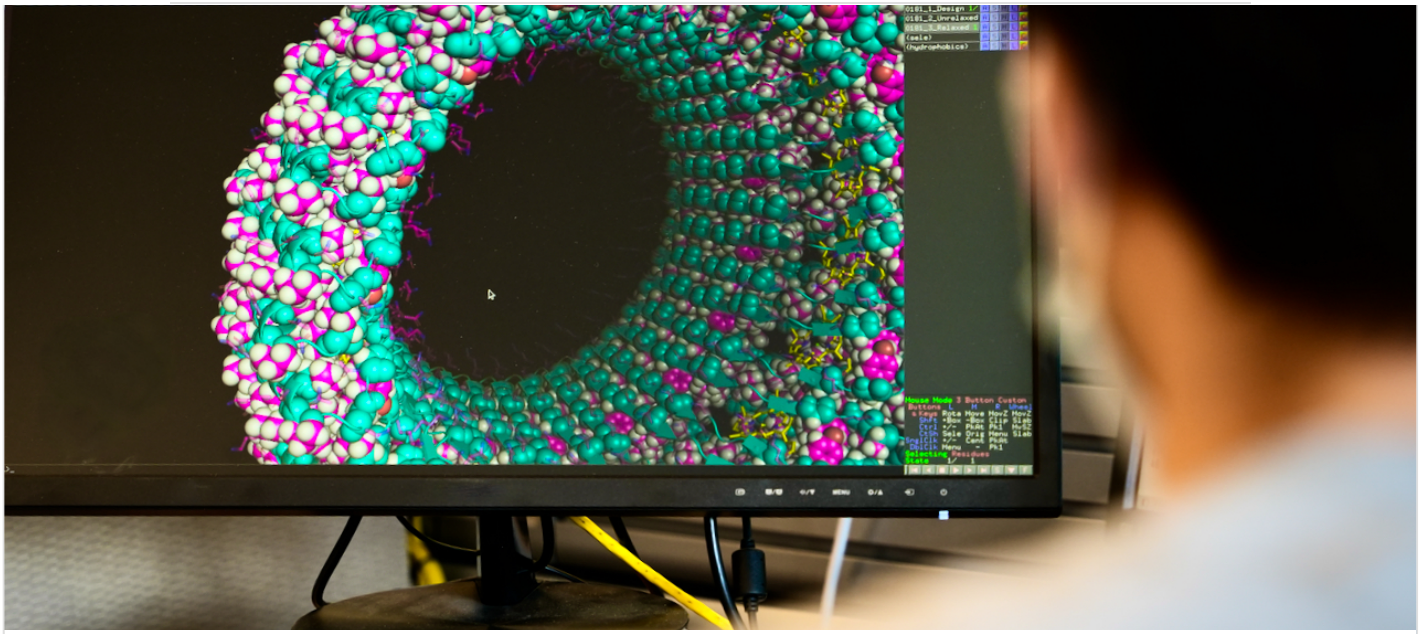


Figure 1.

A t-SNE plot visualization of the LD_{50} dataset (cyan) and top 2000 MegaSyn AI-generated and predicted toxic molecules (salmon) illustrating VX (purple and 2D structure). Many of the molecules generated are predicted as more toxic *in vivo* in the animal model than VX (histogram showing cutoff for VX LD_{50}).



A scientist designing a new protein using computer software

October 19, 2023 09:37 AM EDT
Updated 11:05 AM
AI, In Focus

ENDPOINTS *in* FOCUS

After years of hype, the first AI-designed drugs fall short in the clinic

Andrew Dunn

Earlier this month, UK-based Exscientia slipped into a pipeline update that a Phase I/II study of its cancer drug candidate EXS-21546 was winding down. That cut followed a decision last year by its partner Sumitomo Pharma to abandon another of its AI-designed drugs. In April, a test of BenevolentAI's dermatitis drug fell short as well. And Recursion Pharmaceuticals — the third of AI's early generation — hasn't recorded a trial failure but has had a handful of clinical setbacks that don't necessarily bode well.

Patrick Malone

There's no shortage of AI naysayers, and the 0-for-3 start suggests that AI hype has set unrealistic expectations. Clinical wins are rarities in biotech, where an estimated 5% or 10% of drugs that head into human testing actually get approved.

"If you take the hype and PR at face value over the last 10 years, you would think it goes from 5% to 90%," Patrick Malone, a principal at KdT Ventures, said of AI. "But if you know how these models work, it goes from 5% to maybe 6% or 7%."

These three companies have been at this for roughly a decade, combining to rack up an accumulated deficit of over $1.5 billion.

Ivan Griffin

The reality check of the clinic, paired with a dour biotech market, has beaten up these first-generation biotechs that went public in 2021 or 2022. Their stock prices are all down at least 75%, underperforming the biotech market, even as new AI startups have continued to raise substantial sums of money. Generate:Biomedicines, Inceptive, Iambic and Genesis, for instance, have combined to raise $673 million over the past few months.

Executives at these first-generation companies say it's too early for a verdict on

His biotech announced the first AI-designed drug had entered the clinic. Its partner Sumitomo led the Phase I study in obsessive-compulsive disorder, with the *Financial Times* calling the trial's start a "critical milestone for the role of machine learning in medicine."

About two years later, in January 2022, Sumitomo disclosed they had abandoned the drug, which failed to meet the study's criteria. In an interview, Hopkins said Exscientia's job was to just design the molecule, with Sumitomo making the clinical decisions.

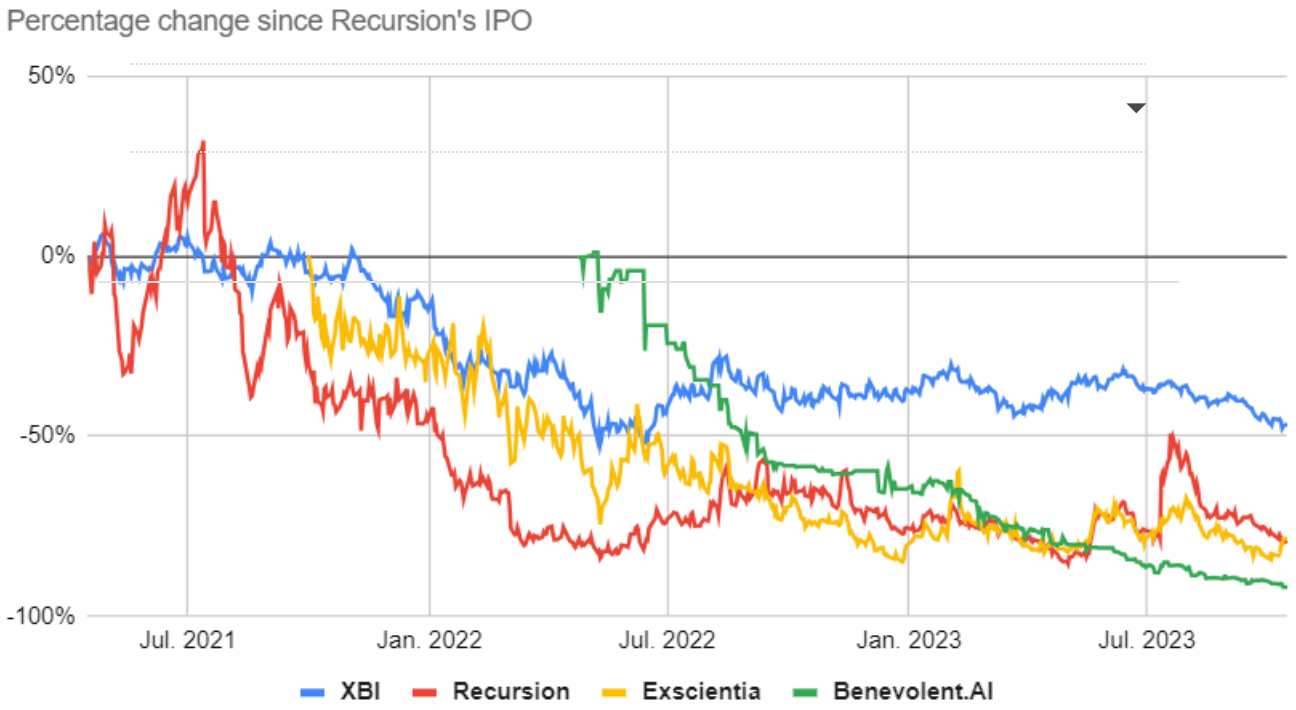
Exscientia had more control over the next drug, a cancer treatment called EXS-21546, which it brought into the clinic in December 2020. Earlier this month, the biotech said it was discontinuing an ongoing Phase I/II study with modeling suggesting "it will be challenging for '546 to reach a suitable therapeutic index."

Andrew Hopkins

Hopkins said that the trial didn't fail, as the biotech doesn't have full results back.

"It wasn't a clinical data decision," he said. "It was a strategic decision" to prioritize two other cancer drugs that the company believes have better chances.

"We don't want to be one of those companies that keeps pushing a program forward because it's the only thing they have," Hopkins said.



Recursion stands apart as the only one of the three to maintain a valuation above $1 billion today. (Exscientia is worth about $650 million, while BenevolentAI is valued at $117 million.) The biotech has had several positive Phase I readouts centered on safety and tolerability, such as a *C. diff* drug recently clearing a healthy volunteer study of 42 people. The company used AI to identify existing compounds rather than design new drugs for its early pipeline.

Dylan Reid

"The market has never known what to make of these companies," said Dylan Reid, a partner at Zetta Venture Partners. "They've been way too excited and way too down. At one point, they value the platform at X billions of dollars, and today, it's probably a drag on valuation."

While Recursion hasn't had a clinical failure, its development plans have had hiccups. It quietly dropped a rare disease drug last year, citing "noise in the potency" and delays in getting a trial going. Earlier this month, the Salt Lake City-based biotech slimmed down an ongoing Phase II study for an-

The AI clinical pipeline is full of other players as well, such as Verge Genomics' ALS drug, BPGbio's brain cancer treatment, Insilico Medicine's idiopathic pulmonary fibrosis drug, Generate:Biomedicines' Covid-19 antibody, and Auransa's liver cancer therapy.

Strategies evolve as more players emerge

AI backers say successful programs like Moderna's Covid-19 vaccine or Nimbus Therapeutics' TYK2 inhibitor used AI to a degree. But those drugmakers don't brand those medicines as AI-designed, while Exscientia, BenevolentAI, and Recursion market their approach as AI-driven or AI-enabled.

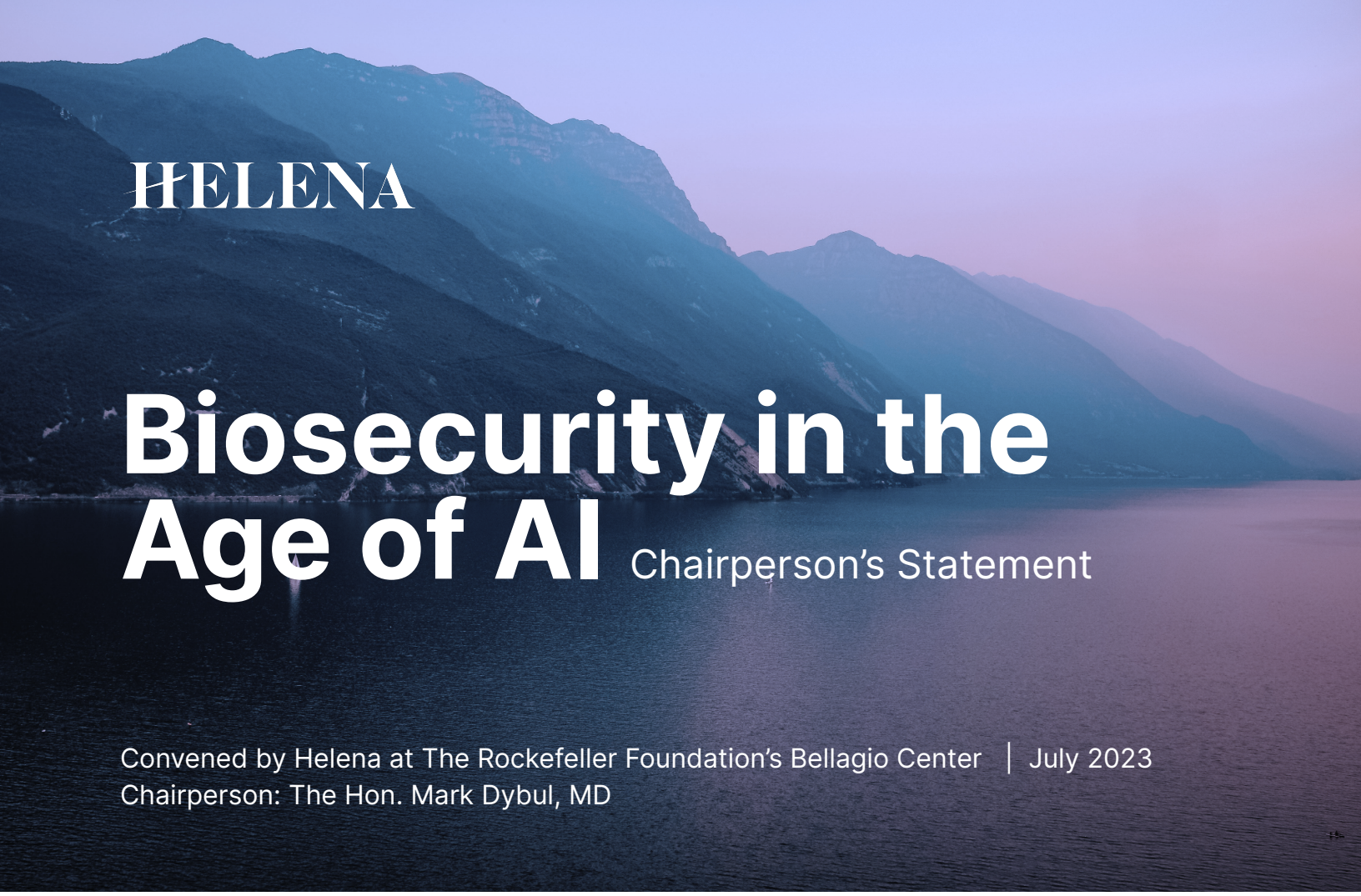
A decade in, leaders of the first-generation companies say they are still learning.

BenevolentAI, for instance, says its failed atopic dermatitis drug candidate didn't use the company's target identification approach, which is behind its ulcerative colitis drug that entered the clinic earlier this year.

Exscientia has incorporated more human tissue samples in its research process and hired experienced clinical hands like Michael Krams, Hopkins said.

"We've also now realized if we want to change the probability of success in the clinic, it's not just better molecules," Hopkins said. "We also need better translational models."

AUTHOR



HELENA

Biosecurity in the Age of AI

Chairperson's Statement

Convened by Helena at The Rockefeller Foundation's Bellagio Center | July 2023
Chairperson: The Hon. Mark Dybul, MD

A note on The Rockefeller Foundation Bellagio Center: For decades, the Bellagio Center has supported leaders to develop solutions and promote ideas that address the world's biggest challenges. It is a unique place and opportunity for leaders to have deep discussions that create new understanding. A convening or residency at the Bellagio Center does not imply an endorsement by The Rockefeller Foundation of the work and does not necessarily reflect positions or policies of The Rockefeller Foundation.

About Helena: Helena is a new breed of institution designed to meet the challenges of today and tomorrow. Through Helena Projects, we seek to implement solutions to critical societal problems. Helena's past and current work has addressed elements of climate change, governance reform, and existential risk mitigation, and encompassed nonprofit, for-profit, and legislative actions.

About the author: The Hon. Mark Dybul, MD, is a Helena member and professor of medicine at the Georgetown University Medical Center (GUMC). He was the head of both the U.S. PEPFAR and the Global Fund to Fight Aids, Tuberculosis, and Malaria (GFATM).

A note on the participants: Because of the positions held by several participants, it was determined that the names of those attending "Biosecurity in the Age of AI," convened under the Chatham House Rule, would not be published.

Table of Contents

Executive Summary

Introduction

Technological advancements in life sciences research – turbocharged by new and emerging Artificial Intelligence (AI) capabilities – are furnishing incredible breakthroughs in human health, sustainable development, and other fields. This convergence promises world-changing benefits for health and well-being, including opportunities to achieve global goals for pandemic preparedness and response, improve cancer detection and treatment, and alleviate chronic diseases such as diabetes. More broadly, AI holds the potential to transform sectors ranging from agriculture and food security to defense to climate change and energy production.

Despite the untold advantages these technologies will afford, prominent scientists and business leaders, – many of whom have been integral to the development of AI – have expressed serious concerns over the potential downside consequences of such innovation. Recently, more than 350 executives, researchers, and engineers signed a *Statement on AI Risk,*[1] which asserts that mitigating threats from AI "should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." In addition to concerns that have been enumerated in public and policy discussions, such as those relating to job displacement and risks to democracy from misinformation, members of the scientific community have recently begun to sound alarms over specific threats emerging from the intersection of AI and synthetic biology.

AI-Enabled Biology (also referred to in this report as "AI Bioconvergence") will profoundly augment our ability to prepare for and rapidly respond to naturally occurring, accidental, or weaponized pandemic threats. Moreover, expanded access to distributed biology tools and research has and will continue to furnish life-saving breakthroughs in diagnostics, vaccines, and other medical treatments. But this democratization of access will also transform the risk landscape by raising the ceiling on the potential harmfulness of engineered viruses and empowering a growing number of actors with the ability to modify or create pandemic-level pathogens.

In a recent academic exercise, an AI chatbot assisted graduate students in identifying four potential pandemic pathogens, issued step-by-step "lay-person" instructions for virus generation, and directed students toward companies that could manufacture synthetic DNA sequences for the purposes of engineering and producing these pathogens.[2] While it is an overstatement to suggest that such a set of instructions could result in an untrained actor engineering a pandemic-capable pathogen today – and it is possible that with additional training and time, the same results may have been achieved without AI – the example showcases the potential of technology to "upskill" individuals without relevant subject matter expertise or experience. By leveraging a small slice of well-intended research, ongoing technological advancement could result in an expanded and increasingly empowered pool of actors with the ability to manipulate pathogens. It may also increase the overall "ceiling" of harm by rendering

[1] "Statement on AI Risk: Cais." *Statement on AI Risk | CAIS*, www.safe.ai/statement-on-ai-risk. Accessed 21 July 2023.
[2] Soice, Emily H. *Can Large Language Models Democratize Access to Dual-Use Biotechnology?*, arxiv.org/ftp/arxiv/papers/2306/2306.03809.pdf. Accessed 21 July 2023.

pathogens more transmissible, deadlier, and/or more able to evade existing vaccines or therapies than their natural counterparts.[3]

Currently, a lack of informed and rigorous discourse around AI Bioconvergence raises concerns about two undesirable outcomes: no action at all, or an overreaction that prevents important scientific inquiry with the potential to create far-reaching advances for the future of humanity.

For decades, and accelerated by the onset of the Covid-19 pandemic, governments, regional, and global entities have been working to advance approaches to mitigate biological threats, whether naturally occurring or man-made. To date, however, these approaches have lagged significantly behind technological developments at the nexus of AI, synthetic biology, and life sciences research.

Among specific recommendations which will be delineated in this document, we therefore recommend swift action to close this gap by addressing AI Bioconvergence capabilities that may even now pose large scale risk to humanity.

Current Policy Landscape

Governments, regional entities, and global bodies including the United States,[4] United Kingdom,[5] African Union, European Commission,[6] and the World Health Organization (WHO) [7] have all developed recommendations to enhance biosafety and biosecurity, particularly as it pertains to Dual Use Research of Concern (DURC) and life science research with enhanced Pathogens of Pandemic Potential (ePPPs). Some countries – including the U.K., Canada, Switzerland, the Netherlands, and Germany – have made significant progress in establishing comprehensive oversight of laboratories working with pathogens.[8] Though more nascent, China has also publicly committed to strengthening its own biosecurity strategy to govern DURC activities at the national level. In the U.S., the Biden Administration recently announced that it secured voluntary commitments from leading technology developers to safeguard AI; this action represents an important first step that will inform ongoing efforts to combat AI Bioconvergence threats.

[3] Experts have outlined specific implications of AI Bioconvergence driving these potential negative outcomes, including: AI decreasing the time to engineer pathogens and toxins; AI supporting efforts to circumvent access barriers like screening; software enabling the design of new and custom agents; and modeling via AI, eliminating the need for home wet labs.
[4] "Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy." *The White House*, 12 Sept. 2022, www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/.
[5] Office, Cabinet. "UK Biological Security Strategy." *GOV.UK*, 11 June 2023, www.gov.uk/government/publications/uk-biological-security-strategy.
[6] The EU has been engaged in ongoing pandemic preparedness and biosecurity efforts through JA Terror. "About the Project." *JA TERROR*, 26 June 2023, www.jaterror.eu/about-the-project/.
[7] "Global Guidance Framework for the Responsible Use of the Life Sciences: Mitigating Biorisks and Governing Dual-Use Research." *World Health Organization*, www.who.int/publications/i/item/9789240056107. Accessed 21 July 2023.
[8] Rocco Casagrande, Testimony Before the House Committee on Energy and Commerce, 27 April 2023.

In tandem, governments[9] and entities like the WHO's Intergovernmental Negotiating Body (INB)[10] are working to outline clear goals to enhance pandemic detection and response capabilities. Resiliency strategies include investments into surveillance systems, improved coordination within and across countries, the development of and expanded access to medical countermeasures, increased personal protective equipment (PPE) stockpiles, and enhanced capacity to respond to pandemics, both within countries at high risk for zoonotic spillover and elsewhere.

While these efforts should be supported and advanced, to date, many policies regulating life sciences research have been limited to government-funded activities. Given the substantial commercial demand for breakthrough bioengineered products, it is essential that all research funders, including governments and those within the private sector, act urgently to mitigate bio risks. In doing so, stakeholders will need to expand the reach of new and existing policies to encompass work funded by the private sector, nonprofits and foundations, and wealthy individuals. These policies must be carefully designed and adequately resourced to adapt to the pace of change and remain effective over time, while not unduly impeding beneficial work.

Organizations such as the National Science Advisory Board for Biosecurity (NSABB) in the U.S.[11] and the WHO[12] have established thoughtful guidance to inform the continued development of policy frameworks and strategies that harness the "best of bio" while limiting downside consequences. Of particular relevance, the NSABB reinforced the need for increased oversight of publicly-funded research and advocated for the development of criteria to discern when additional scrutiny is required in synthetic biology, the removal of exclusions for research focused on surveillance and vaccine development, and greater transparency and risk-benefit analysis in life science activities. Their recommendations stopped short, however, of providing specific guidance for oversight of bioinformatics, modeling, and other in-silico research. This gap underscores the critical need for expanded efforts to address emerging threats and opportunities from AI Bioconvergence.

While global norms and standards to mitigate against the deliberate or accidental misuse of biotechnology exist, very few requirements proactively build biosafety and biosecurity into the review and design of new life sciences innovations, whether funded by companies, countries, or philanthropists.The advent of AI – which is significantly expanding access to specialized skills and information – makes expanded action imperative.

Biosecurity in the Age of AI

In late May 2023, the problem-solving organization Helena convened a small group of senior leaders from industry, government, think tanks, and academia to interrogate this rapidly evolving risk landscape and pressure-test courses of action. The meeting *Biosecurity in the Age of AI* took place over the course of two and a half days at The Rockefeller Foundation's Bellagio Center. The group's discussions

[9] *Biden-Harris Administration's National Security Strategy - the White House*, www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administrations-National-Security-Strategy-10.2022.pdf. Accessed 21 July 2023.
[10] "Intergovernmental Negotiating Body (INB)." *World Health Organization*, inb.who.int/. Accessed 21 July 2023.
[11] *Proposed Biosecurity Oversight Framework for the Future of Science*, osp.od.nih.gov/wp-content/uploads/2023/01/DRAFT-NSABB-WG-Report.pdf. Accessed 21 July 2023.
[12] "Global Guidance Framework for the Responsible Use of the Life Sciences: Mitigating Biorisks and Governing Dual-Use Research." *World Health Organization*, www.who.int/publications-detail-redirect/9789240056107. Accessed 21 July 2023.

were informed by interviews with dozens of technical experts and extensive review of existing policy frameworks and subject matter literature. While primarily focused on emerging threats related to AI, the group also addressed risks relating to DURC and research with ePPPs.

At the crux of the discussions was the following question:

Imagine it is five years from now, and we are living in a world that has embraced the promise of AI-Enabled Biology, yet remains safe and secure from biorisk. What governance and policy decisions must we make now to arrive at this optimal future?

The following Chairperson's Statement is a distillation of key recommendations to inform ongoing discussion and advance swift action. It is not intended to reflect consensus. In fact, the meeting in Bellagio highlighted critical areas requiring further reflection, analysis, and review. As one example, the group's conversations surfaced vigorous debate vis-à-vis how to appropriately balance information hazard risks arising from prospective policy actions against potential benefits. While this document will illuminate these and other tensions, resolution will require diligent investigation and continued and well-supported technical evaluation.

Currently, a lack of informed and rigorous discourse around AI Bioconvergence – its benefits, risks, and appropriate policy responses – raises concerns about two undesirable outcomes: no action at all, or an overreaction that prevents important scientific inquiry with the potential to create advances for the planet and beyond.

While one report will be inadequate to address the enormity of the challenges ahead, actions must be taken *now* to rapidly minimize misalignment between technological advancement and existing biosafety and biosecurity policies and tools.

This Chairperson's Statement will focus on the urgent tasks at hand as outlined in the following immediate actions:

Recommendation 1: Establish Public-Private AI Task Forces[13] and Subordinate Technical Working Groups:

In response to existing and rapidly evolving threats and opportunities from AI, the meeting at Bellagio surfaced the critical need to convene Public-Private AI Task Forces (PP AI Task Forces) at the highest executive levels across countries and regional and international entities. To mitigate against specific risks from AI Bioconvergence, these Task Forces should integrate advanced biosecurity expertise and be rapidly deployed in countries and regions with highly developed biotechnology sectors. Subordinate AI Bioconvergence technical working groups (TWGs) should

[13] While we will refer to these governance bodies as "Task Forces" throughout the doc, nomenclature (e.g., "task force" vs. "council") will reflect national, regional, and institutional differences. Such entities will also be compositionally variable depending on the place of origin.

incorporate expertise from across domestic and international domains (including health, defense, national security, cyber security, industry, and commerce) and sectors (public, private, and non-profit). Since AI extends beyond biosecurity, PP AI Task Forces should establish additional topic-specific TWGs to address critical areas as needed – such as AI's effects on information integrity and democracy.

Such efforts could be informed by United Nations Secretary General António Guterres's recent call to assemble a global AI watchdog, and expand on existing task forces, working groups, or equivalents, including those in the U.S., U.K., and E.U.

AI PP Task Forces and TWGs should draw on best-in-class public-private partnership models, including the Coalition for Epidemic Preparedness Innovations (CEPI), the African Union and African Center for Disease Control's (Africa CDC's) public/private COVID-19 response efforts, and Operation Warp Speed in the United States.

The U.K.'s recently established AI Foundation Model Taskforce[14] presents a significant and important opportunity for public-private engagement on AI Bioconvergence and should be supported by a subordinate working group providing technical analysis and recommendations. Likewise, the U.S.'s National Security Commission on Emerging Biotechnology[15] would benefit from the support of a working group focusing on this intersection.

When necessary, leaders should pursue legislative action to fund and institutionalize working group activities for the foreseeable future (five years at minimum). Among other areas requiring rapid decision-making, TWGs reporting to PP AI Task Forces should develop policy options to inform government and non-government action around the recommendations outlined in this report.

Recommendation 2: Safeguard the Digital-to-Physical Frontier, Starting with Mandatory DNA Synthesis Screening:

Governments, especially those with advanced synthetic biology and AI economies, should strengthen partnerships with the private sector to prevent digitally-designed threats from transforming into physical biological risks. AI technologies can enable discovery of harmful biological functions and furnish pathways to develop them, including by actors with fewer skills to perform this work safely, or those who may do so with malevolent intent. Therefore, the

[14] Department for Science, Innovation and Technology. "Tech Entrepreneur Ian Hogarth to Lead UK's AI Foundation Model Taskforce." *GOV.UK*, 18 June 2023, www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce.

[15] Reilly, Briana. "Lawmakers Name Appointees to New Emerging Biotech Panel. Inside Defense, 17 March 2022. https://insidedefense.com/insider/lawmakers-name-appointees-new-emerging-biotech-panel. Accessed 24 July 2023.

digital-to-physical boundary is increasingly fragile and remains critically important as the primary safeguard against misuse.

As an essential first step, governments should implement **mandatory screening policies** for DNA synthesis.To date, screening practices have been voluntary and primarily industry-led. While continued industry leadership and engagement will be critical to keep pace with development, thus far, these practices have not provided adequate coverage. In the development of mandatory requirements, governments could mirror "Know Your Customer" (KYC) and "Know Your Order" (KYO) policies utilized in the financial sector.

Countries should seek to implement screening requirements as quickly as possible, through executive order or equivalent action, or legislative action when necessary. Crucially, such policies and legislation will need to evolve over time to address emergent AI-Enabled Biology risks and other developments. To ensure sustained implementation, regulation should specify appropriate resourcing and expertise to support oversight. Given advancements in "desktop synthesis" that proliferate critical vulnerabilities in the digital-to–physical frontier, TWGs should concertedly assess and develop screening tools and policies that extend beyond publicly funded and industry-level actors and anticipate and mitigate against emerging threats from individuals operating benchtop synthesizers and assemblers.

Countries, regional, and international entities will need to adopt comparable highest-standard screening mechanisms, and should consider adopting universal approaches,[16] to minimize the potential for regulatory arbitrage.

Screening mechanisms should be available to all providers of DNA, ideally at low cost, or fully reimbursed or directly paid by governments.

To safeguard the digital-to-physical frontier over the long-term, governments should consider red-teaming screening mechanisms to surface and remediate vulnerabilities, and invest in "next generation" tools and methodologies that anticipate and counter efforts to bypass standard screening approaches.[17] At the same time, governments, with support from TWGs, should work to securely identify and mitigate hazards arising from the dissemination and sharing of high-risk sequence data, pathogen characterization, and research methods– including by increasing accountability methods and assessing the value of liability measures in this domain. TWGs should also evaluate data set generation and sharing in light of implications for AI models trained on this data and increased opportunities for digital-to-physical transcendence.

[16] Such as the International Common Mechanism for DNA Synthesis and SecureDNA.
[17] The U.S. Government's 'Functional Genomic and Computational Assessment of Threats," (FUN GCAT) provided support for such tool development from 2017-2022. IARPA - Fun GCAT

Recommendation 3: Appropriately Guardrail AI Technology, Including Large Language Models (LLMs) and Biological Design Tools (BDTs):

The use of AI tools in synthetic biology – including LLMs and BDTs – will expand access to pandemic-class biological agents and may allow a growing number of actors to enhance the lethality, host range, or transmissibility of these pathogens. Therefore, governments, biotechnology developers, and life science research funders should develop approaches to rapidly guardrail these technologies.

AI PP Task Forces should consider whether and how to control access to powerful AI models – perhaps through KYC and KYO policies linked to operational entities with enforcement capabilities. TWGs should also develop approaches to test and evaluate AI models and thoughtfully consider accountability mechanisms – including resource incentives (e.g., government contracts), as well as potential liability and regulatory measures that encourage the private sector to responsibly develop and deploy technologies.

In concert with TWGs focused on information integrity, AI Bioconvergence TWGs should also address threats posed by **AI to fuel mis- and disinformation in biology,** which could undermine confidence in the economic and functional value of biotechnologies and stoke chaos in the event of a biological incident.

Recommendation 4: Refine Policies Concerning ePPPs and Update and Reinforce Biorisk Policies to Mitigate Against Accidental and Deliberate Misuse:

Governments, in partnership with leading advisory boards and TWGs, should update definitions of concerning pathogens to accommodate technological advances in gene synthesis and manufacturing, increase oversight of research with ePPPs, and establish independent review mechanisms to enable more effective risk reduction in the field.

Health authorities and their advisory bodies (like the NSABB in the U.S. and global counterparts), as well as appropriate academic bodies, should advance additional efforts – ideally in partnership with TWGs – to improve, extend, and evolve oversight policies to address AI Bioconvergence. In addition, governments should engage with leading biosafety and biosecurity experts to decrease risks of accidental infection/transmission and deliberate misuse and ensure ongoing oversight and surveillance of novel research, biotechnology development, and science. In addition to detecting, assessing, and preventing immediate threats, tracking approaches should monitor unforeseen effects over the long-term.

Enhanced oversight measures should include increasing transparency around biosafety and biosecurity protocols, resourcing entities performing biological research and disease surveillance

to build sufficient biosafety and biosecurity capacity, establishing graduated reviews to triage research according to risk level, and conducting risk/benefit analyses to identify and fund alternatives to highest-risk research.

Recommendation 5: Enhance Biosecurity and Biosafety Norms to Explicitly Include AI-Enabled Biology and Promote International Organizations and Tools to Practically Implement Them:

The swift pace of technological development necessitates the evolution of biosafety and biosecurity norms, standards, and practical implementation. For decades, national and international tools to reduce biological risks have lagged significantly behind technology development. AI Bioconvergence advances are the latest to surge past existing risk reduction frameworks.

To meet and get ahead of emerging risks and opportunities, biotechnology and life sciences research funders must prioritize biosafety, biosecurity, and AI Bioconvergence as an integral component of their mandate. In addition, TWGs should recommend rigorous risk/benefit assessments in review processes for AI Bioconvergence research that carries the potential to cause large-scale harm. TWGs, in concert with governments and key advisors, should also seek to develop norms to address new tensions surfaced by AI Bioconvergence, such as the risks inherent to the development and dissemination of data sets. New norms should be developed and pressure tested in accordance with best practices from cybersecurity and nuclear security where relevant.

In concert, research and technology funders should commit to regular biosafety and biosecurity reviews and build additional funding into proposal and investment costs to support biosecurity-by-design approaches and accommodate more robust safety measures and requirements.

The creation of innovative tools that allow stakeholders from across government, foundations, and the private sector to mitigate risk in real time – *while* new technologies are being developed – will be essential.

Global initiatives, such as International Biosecurity and Biosafety Initiative for Science (IBBIS) and its associated Funders Compact, as well as regional efforts like the Asia Pacific Biosafety Association (APBA)[18] and African Biological Safety Association (AfBSA International),[19] can serve as hubs for public-private efforts to discuss and evolve norms and develop and disseminate novel tools.

[18] "Who Interim Guidance for Laboratory Biosafety Related to 2019-Ncov/COVID-19/SARS-COV-2." ..., a-pba.org/. Accessed 21 July 2023.
[19] *African Biological Safety Association (AfBSA) - International Biosafety*, internationalbiosafety.org/ifba_members/african-biological-safety-associationabsa/. Accessed 21 July 2023.

Recommendation 6: Resilience – Invest in Early Warning and Detection, Response Capacity, and Accountability Measures, and Build Biosafety and Biosecurity into These Approaches:

Many experts have warned about the possibility of another naturally-caused pandemic at the magnitude of COVID-19 within the next decade. Given added risks from AI Bioconvergence – including accidental and intentional release of synthetically-created pathogens – it is therefore imperative that governments, regional, and multinational entities strengthen surveillance and response capacity worldwide.

Governments and entities across continents have taken steps to bolster health security regulations, enhance early warning and detection systems, invest in PPE and medical countermeasures, and set aside financing to ensure readiness in the event of a pandemic. Many of these core priorities have been outlined by the WHO's Intergovernmental Negotiating Body (INB)[20] and the Independent Panel for Pandemic Readiness and Response,[21] and echoed by additional experts.[22] The institutionalization of resiliency efforts – as modeled in the launch of the WHO Pandemic Hub,[23] the 2022 U.S. National Biodefense Strategy,[24], the U.S. Center for Forecasting and Outbreak Analytics,[25] and the U.K.'s recent announcement of its intention to develop a Biothreat Radar[26] – are evidence of global commitments to increase detection and defense capabilities. The global Pandemic Fund,[27] based at the World Bank, has also prioritized biosafety and biosecurity considerations as an explicit part of its overall initial emphasis on disease surveillance, laboratory capacity, and health security workforce development.

By expanding the biological risk landscape, AI-Enabled Biology necessitates a corollary expansion of such resiliency efforts. Therefore, TWGs should engage with preparedness and health security

[20] "Intergovernmental Negotiating Body (INB)." *World Health Organization*, inb.who.int/. Accessed 21 July 2023.
[21] *The Independent Panel for Pandemic Preparedness and Response*, theindependentpanel.org/wp-content/uploads/2021/05/COVID-19-Make-it-the-Last-Pandemic_final.pdf. Accessed 21 July 2023.
[22] *Delay, Detect, Defend: Preparing for a Future in Which Thousands ... - GCSP*, dam.gcsp.ch/files/doc/gcsp-geneva-paper-29-22. Accessed 21 July 2023.
[23] "Pandemic Hub." *World Health Organization*, pandemichub.who.int/. Accessed 21 July 2023.
[24] *National Biodefense Strategy and Implementation Plan - the White House*, www.whitehouse.gov/wp-content/uploads/2022/10/National-Biodefense-Strategy-and-Implementation-Plan-Final.pdf. Accessed 21 July 2023.
[25] "About Us." *Centers for Disease Control and Prevention*, 20 Mar. 2023, www.cdc.gov/forecast-outbreak-analytics/about/index.html.
[26] "Dowden: World-Class Crisis Capabilities Deployed to Defeat Biological Threats of Tomorrow." *GOV.UK*, www.gov.uk/government/news/dowden-world-class-crisis-capabilities-deployed-to-defeat-biological-threats-of-tomorrow. Accessed 21 July 2023.
[27] "The Pandemic Fund." *World Bank*, fiftrustee.worldbank.org/en/about/unit/dfi/fiftrustee/fund-detail/pppr. Accessed 21 July 2023.

colleagues across sectors to ensure that detection and response approaches address novel threats posed by AI Bioconvergence, specifically by ensuring that development and investments in surveillance and detection are matched to the pace of AI-Enabled Biological risks.

Critically, resiliency strategies must be designed to withstand potential public mis- and disinformation campaigns fueled by the convergences of AI and social media.

Finally, governments should work with TWGs and leaders in the private sector and academia to develop attribution technologies[28] that can determine if a pathogen was engineered,[29] and by whom. Surveillance and attribution measures should be linked to appropriate accountability and enforcement mechanisms, with the goal of disincentivizing both accidents and intentional misuse.

[28] Lewis G, Jordan JL, Relman DA, et al. The Biosecurity Benefits of Genetic Engineering Attribution. *Nature Communications.* 2020 Dec;11(1):6294. DOI: 10.1038/s41467-020-19149-2. PMID: 33293537; PMCID: PMC7722838.
[29] Bioworks, Ginkgo. "IARPA, Ginkgo Bioworks and Draper Announce New Technologies to Detect Engineered DNA." *PR Newswire: Press Release Distribution, Targeting, Monitoring and Marketing*, 17 Oct. 2022, www.prnewswire.com/news-releases/iarpa-ginkgo-bioworks-and-draper-announce-new-technologies-to-detect-engineered-dna-301650505.html.

September issue: The hype, peril, and promise of AI

Why AI for biological design should be regulated differently than chatbots

By Matthew E. Walsh | September 1, 2023

By: Raevsky Lab/Adobe Stock

Share

MIT researchers recently contrived a scenario where non-scientist students used ChatGPT to help them obtain information on how to acquire DNA that could make pathogens with pandemic potential. These undergraduate students reportedly had limited biological know-how. But by using the chatbot, they were able to gain the knowledge to create dangerous material in the lab and evade biosecurity measures. This experiment drew attention to the impacts of artificial intelligence tools on the

biothreat landscape—and how such applications contribute to global catastrophic biological risks.

In recent weeks, scholars, the US policy community, and the public have been discussing the biosecurity implications and governance of AI-based tools. The White House recently released a fact sheet detailing the security measures that top large language model-based chatbot developers have voluntarily committed to—including internal and external security testing to guard against AI-based biosecurity risks. In mid-July, Sen. Edward J. Markey (D-Mass.) introduced legislation, the *Artificial Intelligence and Biosecurity Risk Assessment Act,* that, if enacted, would require the Assistant Secretary for Preparedness and Response to research how artificial intelligence tools could be used to generate biological weapons. And groups have published reports detailing recommendations to establish effective governance over artificial intelligence, such as the Helena Project report, *Biosecurity in the Age of AI*.

In an effort to include all the ways in which artificial intelligence tools influence the biothreat landscape, policy conversations often group together general-purpose chatbots with biology-specific, AI bio-design tools. Understanding how each category of AI tools work, what their capabilities and limitations are, and where they are in their commercial development is important to establish effective governance. But it is most critical to recognize that large language model-based chatbots and bio-design tools influence the biosecurity landscape in vastly different ways. Their governance, therefore, should be considered and developed independently.

Large language model-powered chatbots. These chatbots are a combination of a large language model and a user interface. Language models ingest vast amounts of data, typically text of human languages— what practitioners call “natural language.” Training these models consumes tremendous amounts of computation resources and time, often months. Through this process, the large language model learns the structure, or grammar, of the language in the data and commonly contains hundreds of billions of parameters. A user interface can be overlaid on the model, which then results in an easy-to-use AI tool, such as ChatGPT, Bard, or Claude. Based on the information in their training data, these tools respond to user queries with human-like responses. Because the training data is often scraped from the internet, the breadth of responses from these chatbots is vast and can range from restaurant recommendations to error fixes in programming code.

Large language model-based bio-design tools. These applications serve much more specific purposes than chatbots: They are built to help complete biological engineering tasks with varying levels of specificity. Recently developed large language model-based bio-design tools leverage the same methodology that chatbots use and are viewed as a promising application of the method. Instead of training the large language model on natural language, a bio-specific large language model is trained on the amino acid sequences of proteins or other biological sequences. This results in the application outputting biological sequences, instead of natural language.

These tools can learn the favorable properties of a biomolecule and make suggestions on promising options to test in the laboratory, decreasing the number of options needed to test before finding one with desirable properties. For example, the tool known as UniRep helps researchers engineer proteins based on their function, while ESMFold enables engineering based on structure. Both tools could be used, for example, to help design better therapies faster and to engineer proteins in organisms to improve the efficiency of biomanufacturing.

RELATED: **Introduction: The Hype, Peril, and Promise of Artificial Intelligence**

In addition to protein sequences, bio-specific language models have been trained on DNA sequences and even on glycan (sugar) sequences, simultaneously expanding their potential positive and negative impacts. Unlike the chatbots, bio-design tools that are publicly available generally lack a user interface and require computer programming knowledge to access and use, although there are efforts to make them easier to use.

Impact on the threat landscape. As evidenced in the MIT demonstration, general purpose chatbots can make it easier and quicker for people to access information that is prone to misuse. Because the output of chatbots is based on information found in their training data, these tools should currently not be considered as providing new abilities to malicious actors. For example, the students in the demonstration were asked to use ChatGPT to identify companies that were not members of the International Gene Synthesis Consortium, a group of synthetic DNA providers committed to best practices in biosecurity. The assumption was that if someone wanted to acquire harmful DNA, ordering it from a company not a part of the consortium would be more likely to

succeed than ordering it from one that was a part of the association. As expected, ChatGPT was able to provide a list of companies in moments. But without ChatGPT the user could still acquire the same information—by searching online for DNA synthesis providers and then cross-checking the list against those that are listed on the consortium website.

Some chatbots have been engineered to not provide responses that would be prone to misuse, including biological information, but researchers have shown that these restrictions can be overcome.

Bio-design tools, however, do provide new and improved abilities to their users that could be nefariously repurposed. Currently, these bio-specific tools can engineer one property of a biomolecule at a time. These tools can be used to predict function, ranging from improved binding ability of antibody variants to improved fluorescence of a protein. They can output a long list of probable options which can then be evaluated by a user for other properties, such as amino acid sequence. This gives a knowledgeable user the ability to essentially engineer multiple properties.

One example of misuse would be to use a bio-design tool to identify protein-based toxins that are predicted to be functionally similar to known toxins but are otherwise different enough from those found in nature that traditional safeguarding measures would be ineffective.

Moving forward. When considering the governance of chatbots and bio-design tools, it is important to recognize their differences. Doing so will allow for differentiation in future governance options. In the near term, governance of chatbots should be focused on preventing users from accessing *existing information prone to misuse*. There are ongoing efforts throughout the AI community towards such goals, including those in the voluntary commitments from tech companies outlined by the White House and organizations such as the Responsible Artificial Intelligence Institute. When addressing biosecurity concerns related to chatbots, biosecurity professionals should help inform what types of information could be misused to cause harm. Anthropic, the company behind Claude, for example, collaborated with biosecurity experts in developing their chatbot.

In contrast, governance of bio-design tools should be focused on preventing users from *generating harmful new information*. Technical biosecurity measures could be promoted through community norms and codes of conducts. These measures would be

aligned with existing efforts, such as the Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists. In a chemistry-based scenario that parallels bio-design tools, researchers were able to slightly adjust their existing chemical-design tool to maximize the predicted toxicity of chemicals instead of to minimize. Using this information, the researchers were able to identify chemicals predicted to be more toxic than even the most potent chemical weapons.

RELATED: To avoid an AI "arms race," the world needs to expand scientific collaboration

---

This scenario emphasized the relative ease with which nefarious actors could repurpose existing code that was originally built for beneficial purposes. But that does not mean AI tools must remain locked behind closed doors. Developers could overlay user interfaces, like chatbots, that allow others to use the tool as intended and without being able to make changes to the code. Practices like this should be discussed among the biosecurity community and considered for inclusion into future guidelines and codes of conduct.

Other governance measures, such as risk education and awareness raising of bio-design tools should be pursued. However, there are currently a few challenges in actually doing this. First, work is needed to develop and implement a categorization framework of bio-design tools that will be helpful in determining appropriate governance measures. Large language model based bio-design tools are just one type of bio-design tool. Other bio-design tools, such as AlphaFold2 and Rosetta are not built on large language models but can have the same applications as large language model-based bio-design tools. Governance pertaining to only large language model-based bio-design tools but not other tools with similar capability would be incomplete. Additionally, bio-design tools vary in the degree of user expertise they require (in both biology and computer programming) and in the types and amount of data, among others. A comprehensive framework that considers the multi-faceted landscape of bio-design tools would be very helpful in framing risk education and awareness raising initiatives.

Additionally, there is little, if any, peer-reviewed work analyzing the current impact of bio-design tools on the biothreat landscape. Bio-design tools will increase in capability over

time, and there is no sufficient risk assessment framework for mid- and long- term impacts. Because there is no published work attempting to reach a consensus among experts on what the impacts of large language model-based bio-design tools are on the biological threat landscape, policy makers will find it challenging to agree on what appropriate and commensurate governance measures are.

Lastly, there are few people in the world who have expertise in the differing subject areas of AI, engineered biology, and biosecurity. This means that the most effective and comprehensive work in this space needs to come from teams of experts who have to communicate across academic disciplines.

There is also a difference in the urgency of developing governance of large language model-based bio-design tools and chatbots. Chatbots are becoming increasingly commercialized and wide-spread, and consequently the window for establishing governance is closing. For developed technologies, like chatbots, more stringent governance measures, such as export controls or licensing, are generally more appropriate than they would be for nascent technologies like large language model-based bio-design tools. In the emerging arena of bio-design tools, there is still time to understand their implications and to work with technology developers to ensure that future tools are built with biosecurity considerations in mind—and with whistle-blowing channels for when they are not.

In grouping large language model based chatbots and large language model-based bio-design tools together, it will be challenging to identify one set of governance measures that would apply to both. This could potentially create an obstacle for the policy and scientific communities in aligning on what the appropriate governance measures are and needlessly stalling progress towards mitigating the risk associated with chatbots. Significant work is needed to fully understand and communicate the biosecurity impacts of bio-design tools. Underappreciation for the differences between these two applications, and their impacts on the biothreat landscape, could result in inappropriate or ineffective governance of each while simultaneously harming beneficial technological progress.

Keywords: artificial intelligence, bio-design tools, biosecurity, biosecurity threats, chatbots, governance, large language models
Topics: Disruptive Technologies

Share

Matthew E. Walsh

Matthew E. Walsh is a doctoral student in the department of Environmental Health and Engineering (health security track) at Johns Hopkins Bloomberg... Read More

Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools

Jonas B. Sandbrink[1*]

[1] *Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom*
[*] *Correspondence: jonas.sandbrink@trinity.ox.ac.uk*

Abstract

As advancements in artificial intelligence (AI) propel progress in the life sciences, they may also enable the weaponisation and misuse of biological agents. This article differentiates two classes of AI tools that pose such biosecurity risks: large language models (LLMs) and biological design tools (BDTs). LLMs, such as GPT-4, are already able to provide dual-use information that removes some barriers encountered by historical biological weapons efforts. As LLMs are turned into lab assistants and autonomous science tools, this will further increase their ability to support research. Thus, LLMs will in particular lower barriers to biological misuse. In contrast, BDTs will expand the capabilities of sophisticated actors. Concretely, BDTs may enable the creation of pandemic pathogens substantially worse than anything seen to date and could enable forms of more predictable and targeted biological weapons. In combination, LLMs and BDTs could raise the ceiling of harm from biological agents and could make them broadly accessible. A range of interventions would help to manage risks. Independent pre-release evaluations could ensure that developers have eliminated dangerous capabilities of new models. Risks from powerful science tools might be mitigated through providing differentiated access to legitimate researchers. Lastly, essential for mitigating risks will be universal and enhanced screening of gene synthesis products.

Introduction

Artificial intelligence (AI) has the potential to catalyse enormous advances in the life sciences and medicine. However, as AI accelerates the life sciences, it may also enable harmful and malicious applications of associated capabilities. Urbina *et al.* have demonstrated how an AI-powered drug discovery tool could be used to generate blueprints for plausible novel toxic chemicals that could serve as chemical weapons [1]. Similarly, AI may also empower the weaponisation and misuse of biological agents - and because of their potentially transmissible nature, risks from biological agents may exceed that of chemical ones.

This article differentiates two forms of AI which, in different ways, exacerbate biosecurity risks: large language models (LLMs) and biological design tools (BDTs). These classes of AI tools feature significantly different properties and risk profiles (see Table 1).

Next to the direct ways in which these tools could enable the creation of biological weapons, AI systems may also increase biosecurity risks through indirect avenues. For instance, LLMs could also exacerbate misinformation and disinformation challenges [2], which could negatively impact the response and

attribution of a biological event. Furthermore, LLMs might be misused as tools to radicalise and recruit or to coerce and manipulate scientists to share pathogen samples or acquire technical expertise for biological weapons development. These risks are less unique to biosecurity and are not the focus of this piece.

Risks from large language models (LLMs)

The first class of AI tools that might enable misuse of biology are large language models (LLMs) that have been trained on large amounts of text, including scientific documents and discussion forums. LLMs and related "AI assistants" can provide scientific information, access relevant online resources and tools, and instruct research. Examples include foundation models (e.g. GPT-4/ChatGPT), language models optimised for assisting scientific work (e.g. BioGPT) [3], and LLM-based applications for interfacing with other scientific tools and laboratory robots [4,5]. While foundation models are products of large and expensive training runs and are currently developed by a small number of companies, LLM-based applications have been developed by more resource-constrained academic researchers [4,5].

LLMs might impact the risks of biological misuse in several ways. A key theme is that LLMs increase the accessibility to existing knowledge and capabilities, and thus may lower the barriers to biological misuse (see Figure 1b).

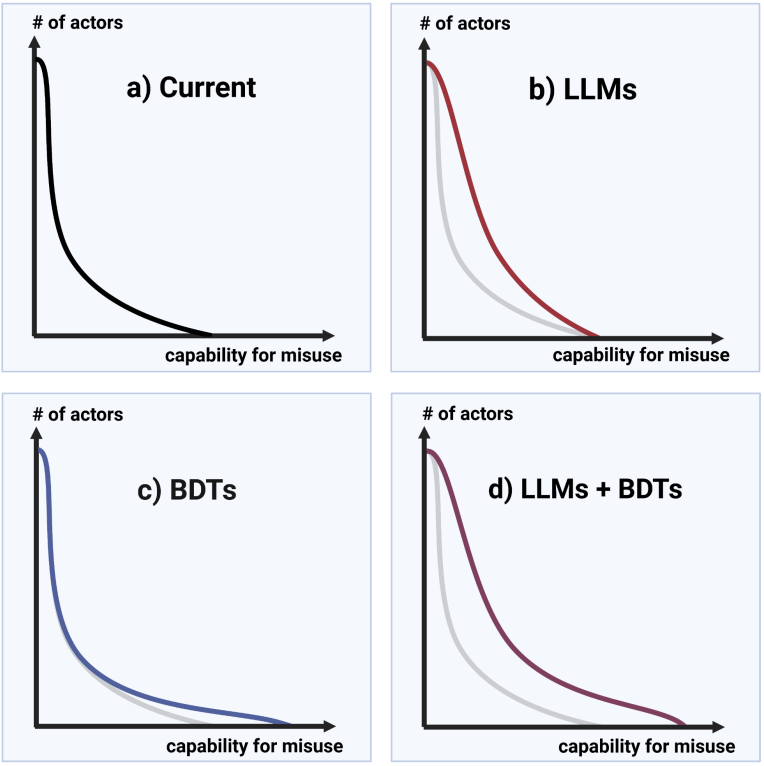


Figure 1: Schematic of effects on LLMs and BDTs on capabilities for biological misuse
Illustrative schematic of how artificial intelligence tools impact capabilities across the spectrum of actors with the potential to misuse biology. a) Currently most individuals are not able to access biological agents, and only a small number of actors are capable of causing large-scale harm. b) Large language models (LLMs) will increase capabilities across the spectrum of actors but are less likely to substantially

raise the ceiling of capabilities. c) Biological design tools (BDTs) will increase the ceiling of capabilities. d) The combination of LLMs and BDTs will increase the ceiling of capabilities and make such capabilities accessible to a significant number of individuals.

1. Teaching about dual-use topics

First, LLMs will enable efficient learning about "dual-use" knowledge which can be used for informing legitimate research but also for causing harm. In contrast to internet search engines, LLMs can answer high-level and specific questions relevant to biological weapons development, can draw across and combine sources, and can relay the information in a way that builds on the existing knowledge of the user. This could enable smaller biological weapons efforts to overcome key bottlenecks. For instance, one hypothesised factor for the failed bioweapons efforts of the Japanese doomsday cult Aum Shinrikyo is that it's lead scientist Seichii Endo, a PhD virologist, failed to appreciate the difference between the bacterium *Clostridium botulinum* and the deadly botulinum toxin it produces [6]. ChatGPT readily outlines the importance of "harvesting and separation" of toxin-containing supernatant from cells and further steps for concentration, purification, and formulation. Similarly, LLMs might have helped Al-Qaeda's lead scientist Rauf Ahmed, a microbiologist specialising in food production, to learn about anthrax and other promising bioweapons agents, or they could have instructed Iraq's bioweapons researchers on how to successfully turn its liquid anthrax into a more dangerous powdered form [6,7]. It remains an open question how much LLMs are actually better than internet search engines at teaching about dual-use topics.

2. Identifying specific avenues to biological misuse

Second, LLMs can help with the ideation and planning of how to attain, modify, and disseminate biological agents. Already now, LLMs are able to identify how existing supply chains can be exploited to illicitly acquire biological agents. In a recent one-hour exercise, LLMs enabled non-scientist students to identify four potential pandemic pathogens and how they can be synthesised, which companies supply synthetic DNA without screening customers and orders, and the potential to engage contract service providers for relevant laboratory work [8]. In the longer term, LLMs could also generate ideas for how to design biological agents tailored for a specific goal, such as what molecular targets would be best suited to produce a particular pathology.

3. Step-by-step instructions and trouble-shooting experiments

Additionally, LLMs could become very effective laboratory assistants which can provide step-by-step instructions for experiments and guidance for troubleshooting experiments. Such AI lab assistants will have many beneficial applications for helping less experienced researchers and replicating experimental methods from publications. However, these AI lab assistants might also support laboratory work for malicious purposes. For instance, a key reason for Aum Shinrikyo's failure to weaponise anthrax was that Seiichi Endo did not succeed at turning a benign vaccine strain of the bacterium into its pathogenic form, despite access to relevant protocols for plasmid insertion. Endo might have succeeded with an AI lab assistant to provide tailored instructions and help with troubleshooting. One crucial open question is how much of an additional barrier "tacit knowledge" plays, knowledge that cannot easily be put into words, such as how to hold a pipette or recognise when cells look ready for the next step of laboratory work. However, what is clear

is that if AI lab assistants create the perception that performing a laboratory feat is more achievable, more groups and individuals might try their hand - which increases the risk that one of them actually succeeds.

4. Autonomous science capability

In the longer term, as LLMs and related AI tools improve their ability to do scientific work with minimal human input, this could potentially transform barriers to biological weapons. Firstly, LLMs can instruct laboratory robots based on natural language commands, which will make them easier to use [9]. Secondly, LLMs can serve as the basis for autonomous science agents, which break tasks into manageable pieces, interface with relevant specialised computational tools, and instruct laboratory robots [5]. Challenges relating to coordinating large teams under secrecy limited the Soviet and Iraq bioweapons programs and likely has also served as a barrier for terrorist groups [6]. If autonomous science capabilities enable individuals and small groups to achieve large-scale scientific work, this will likely empower covert bioweapons programs.

Risks from biological design tools (BDTs)

The second class of AI tools that might pose a risk of misuse are biological design tools (BDTs). These BDTs are trained on biological data and can help design new proteins or other biological agents. Examples include RFDiffusion, as well as protein language models like ProGen2 and Ankh [10–12]. These BDTs are frequently open sourced, regardless of whether they are developed by academia (RFDiffusion) or industry (ProGen2, Ankh). Next to tools for protein or organism design, there are also other machine learning tools with related dual-use implications, such as tools that shed light on host-pathogen interactions through predicting properties like immune evasion [13] or through advancing functional understanding of the human genome [14]. Currently, biological design tools are still limited to creating proteins with relatively simple, single functions. However, eventually, relevant tools likely will be able to create proteins, enzymes, and potentially even whole organisms optimised across different functions.

There are three key ways in which BDTs might impact risks of biological misuse. In contrast to LLMs which mainly increase the accessibility of biological weapons, BDTs may increase the ceiling of capabilities and thus the ceiling of harm posed by biological weapons (see Figure 1c).

1. Sophisticated groups and increased worst-case scenario risks

First, as biological design tools advance biological design, this will likely increase the ceiling of harm that biological misuse could cause. It has been hypothesised that for evolutionary reasons naturally emerging pathogens feature a trade-off between transmissibility and virulence [15]. BDTs might enable overcoming this trade-off and allow the creation of pathogens optimised across both of these properties. Such pathogens might be released accidentally or deliberately, including by groups like Aum Shinrikyo. Bioterrorism with such designed pathogens is a low-probability scenario, because very few people have relevant motivations and - even with AI tools - designing an optimised pathogen will require significant skills, time, and resources. However, these barriers to using BDTs will decrease with advances in large language models and other AI lab

assistants. Thus, humanity might face the threat of pathogens substantially worse than anything nature might create, including pathogens capable of posing an existential threat.

2. State actors and new capabilities

Second, biological design tools may be a key contributor to raising biological engineering capabilities in a way that makes biological weapons more attractive for state actors. The United States never included bioweapons developed during the 1960s in its war plans due to their short shelf life and the risk of harming friendly troops [6]. Iraq never deployed its bioweapons, likely because of a lack of certainty around its effectiveness and fear of retaliatory measures. If AI tools push the ceiling of biological design to make biological agents more predictable and targetable to specific geographic areas or populations, this could increase the attractiveness of biological weapons.

3. Circumventing sequence-based biosecurity measures

In the near term, biological design tools will challenge existing measures to control access to dangerous agents. Examples include the taxonomy-based Australia Group List for export controls and the genetic sequence-based screening of synthetic DNA products. BDTs will make it easier to design potentially harmful agents that do not resemble the function or sequence of any known toxin or pathogen. An example includes "recoding" the function of a known toxin in a substantially different genetic sequence, which current or near-future open source BDTs might already be capable of. Thus, taxonomy or sequence similarity-based controls will not be sufficient to prevent illicit access to harmful biological agents in an age of AI-powered biological design.

Takeaways for risk mitigation

The properties of LLMs and BDTs and their risk profiles have important implications for risk mitigation. Mitigating risks from LLMs requires urgent action, because LLMs are already posing biosecurity risks and LLM capabilities may advance very fast and unpredictably [16]. In contrast, risks from biological design tools are still more ill-defined and advances are somewhat more gradual. For both types of AI tools, governments need to engage tool developers through which they can monitor risks and can create nimble governance strategies. One crucial area to follow is how LLMs interact with BDTs to make advanced biological design capabilities more accessible (see Figure 1d). Possible mechanisms include LLMs providing natural language interfaces to using BDTs, AI lab assistants helping to turn biological designs into physical agents, and eventually LLMs becoming more powerful at biological design than specialised tools.

Pre-release model evaluations

Biosecurity risks from cutting-edge LLMs can be mitigated by ensuring that they do not feature dangerous biological capabilities at release. Leading companies and AI governance scholars are coalescing around pre-release model evaluations as a key tool for identifying dangerous capabilities of new models [17]. OpenAI performed a prototype version of such pre-release model evaluations before the release of GPT-4 [18]. Ideally, pre-release model evaluations would involve an external and independent audit of foundation models with a structured set of tests, including relating to the ability to help with planning or execution of a biological attack. This would incentivise developers to remove harmful model behaviour throughout

training and deployment. Even if a powerful LLM is found to be safe at release, once it is open sourced it can be fine-tuned to develop dangerous capabilities. Thus, it is critical that sufficiently powerful LLMs are not open sourced and their model weights are held securely.

Controlling access to dual-use science capabilities

A crucial question is who should be able to access dual-use scientific capabilities of different models. While it is clearly undesirable that an LLM helps to plan a biological attack, this is less obvious for scientific capabilities with legitimate and harmful applications, such as the synthesis of influenza virus. Arguably, LLMs accessed by the general public do not need to help with such dual-use science tasks. Thus, limiting relevant capabilities for public model versions likely features greater benefits and downsides. In contrast, to reduce risks from LLMs or BDTs developed to help with legitimate research, more differentiated access controls could be explored. Powerful lab assistants and BDTs could require user authentication and, where appropriate, documentation of biosafety and biosecurity review. This would require moving away from open source publishing for such tools [19]. This might for instance make sense for protein design tools that are able to create functional equivalents of controlled toxins and pathogens. Where access controls are imposed, it will be crucial to ensure equitable access across the globe.

Mandatory gene synthesis screening

Lastly, the most effective way to mitigate increased risks from LLMs and BDTs might be to strengthen biosecurity measures at the boundary from the digital to the physical. Access to synthetic DNA is critical for translating any biological design into a physical agent. Industry leaders are already voluntarily screening gene synthesis orders and are calling for a regulatory baseline [20]. Such a mandatory baseline for the screening of gene synthesis orders and other synthetic biology services would be a very effective measure to prevent illicit access to biological agents. At the same time, screening tools need to be improved in step with advances in biological design. For example, it may be possible for future synthesis screening tools to predict the function of novel sequences. To this end, AI developers, biosecurity experts, and companies providing synthesis products could collaborate to develop appropriate screening tools.

Conclusion

It is yet uncertain how and to what extent advances in artificial intelligence will exacerbate biosecurity risks. However, already now, risks at the intersection of AI and biosecurity have policy implications that go beyond their immediate mitigation. Biosecurity risks have become a concrete instantiation of a broader set of artificial intelligence risks that could catalyse general AI governance measures. At the same time, as AI makes the misuse of biology more accessible, this strengthens the need for mitigating dual-use risks in the life sciences more generally. If risks from AI can be effectively mitigated, this sets the groundwork for enabling AI to realise its very positive implications for the life sciences and human health.

Acknowledgements

Markus Anderljung, Anemone Franz, Nicole Wheeler, and others for comments on the manuscript and helpful discussions. This piece only represents the opinion of the author, and not that of any of the organisations that they are working with. The author's doctoral research is funded by Open Philanthropy.

Bibliography

1. Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell. 2022;4: 189–191. doi:10.1038/s42256-022-00465-9
2. Goldstein JA, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv; 2023. Available: http://arxiv.org/abs/2301.04246
3. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23: bbac409. doi:10.1093/bib/bbac409
4. Bran AM, Cox S, White AD, Schwaller P. ChemCrow: Augmenting large-language models with chemistry tools. arXiv; 2023. doi:10.48550/arXiv.2304.05376
5. Boiko DA, MacKnight R, Gomes G. Emergent autonomous scientific research capabilities of large language models. arXiv; 2023. doi:10.48550/arXiv.2304.05332
6. Ouagrham-Gormley SB. Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development. 1st edition. Ithaca: Cornell University Press; 2014.
7. Warrick J. Suspect and A Setback In Al-Qaeda Anthrax Case <span class=. Washington Post. 31 Oct 2006. Available: https://www.washingtonpost.com/archive/politics/2006/10/31/suspect-and-a-setback-in-al-qaeda-anthrax-case-span-classbankheadscientist-with-ties-to-group-goes-freespan/eeb4e5a1-9d08-4dfa-bccc-5c18e311502a/. Accessed 11 May 2023.
8. Soice EH, Rocha R, Cordova K, Specter M, Esvelt KM. Can large language models democratize access to dual-use biotechnology? arXiv; 2023. doi:10.48550/arXiv.2306.03809
9. Inagaki T, Kato A, Takahashi K, Ozaki H, Kanda GN. LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation. arXiv; 2023. doi:10.48550/arXiv.2304.10267
10. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. bioRxiv; 2022. p. 2022.12.09.519842. doi:10.1101/2022.12.09.519842
11. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023; 1–8. doi:10.1038/s41587-022-01618-2
12. Elnaggar A, Essam H, Salah-Eldin W, Moustafa W, Elkerdawy M, Rochereau C, et al. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. arXiv; 2023. doi:10.48550/arXiv.2301.06568
13. Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Sander C, et al. Learning from pre-pandemic data to forecast viral escape. bioRxiv; 2023. p. 2022.07.21.501023. doi:10.1101/2022.07.21.501023
14. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Carranza NL, Grzywaczewski AH, Oteri F, et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. bioRxiv; 2023. p. 2023.01.11.523679. doi:10.1101/2023.01.11.523679
15. Alizon S, Hurford A, Mideo N, Van Baalen M. Virulence evolution and the trade-off hypothesis:

history, current state of affairs and the future. J Evol Biol. 2009;22: 245–259. doi:10.1111/j.1420-9101.2008.01658.x
16. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. arXiv; 2022. doi:10.48550/arXiv.2206.07682
17. Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, Leung J, et al. Model evaluation for extreme risks. arXiv; 2023. doi:10.48550/arXiv.2305.15324
18. OpenAI. GPT-4 Technical Report. arXiv; 2023. Available: http://arxiv.org/abs/2303.08774
19. Smith JA, Sandbrink JB. Biosecurity in an age of open science. PLOS Biol. 2022;20: e3001600. doi:10.1371/journal.pbio.3001600
20. Carter SR, Yassif J, Isaac CR. Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance. Nuclear Threat Initiative; 2023 May. Available: https://www.nti.org/analysis/articles/benchtop-dna-synthesis-devices-capabilities-biosecurity-implications-and-governance/

Table 1: Summary of characteristics, risks, and risk mitigation options for LLMs and BDTs

	Large language models (LLMs)	**Biological design tools (BDTs)**
Definition	Tools trained primarily on natural language which can instruct and conduct research.	Tools trained on biological data that are used for designing new proteins or other biological agents.
Examples	• Foundation models (e.g. GPT-4/ChatGPT) • Language models for assisting scientific work (e.g. BioGPT) • Language model-based tools for (autonomous) scientific research (e.g. ChemCrow [4], Boiko et al. 2023)	• ProteinMPNN, RFdiffusion • Protein language models trained on genetic sequences (e.g. ProGen2) • Smaller and more specialised tools (e.g. Ogden et al 2019)
Developers	• Foundation models: few well-resourced companies. • Science-specific models or LLM applications: distributed (academic) creators.	• Most biological design tools: distributed and open-source • Small number of large models: well-resourced companies.
Major risks	Lower barriers to accessing and misusing biological agents: • Providing information on dual-use topics • Providing lab assistance and, eventually, autonomous research • Identifying avenues for misuse • Creating a perception of increased accessibility In the future, LLMs/autonomous science tools may also increase the ceiling of capabilities.	Increased ceiling of capabilities for sophisticated actors: • Enabling creation and misuse of pathogens much worse than anything known today • Enabling biological weapons targeted to populations or geographies In the short term, enabling the creation of hazardous proteins that are not picked up by existing gene synthesis screening.
Risk mitigation	• Pre-release evaluations by third parties and post-release reporting of hazards for foundation models • Do not open source powerful LLMs and hold model weights securely • Provide differentiated access to dual-use AI tools for science based on authentication of users	• Monitoring of capabilities and risks • For general-purpose BDTs, move away from open source to differentiated access • Universal screening of gene synthesis orders and advancement of functional screening

Defining the scope of AI regulations

*Jonas Schuett**

(Forthcoming in *Law, Innovation and Technology*, Volume 15, Issue 1)

The paper argues that the material scope of AI regulations should not rely on the term 'artificial intelligence (AI)'. The argument is developed by proposing a number of requirements for legal definitions, surveying existing AI definitions, and then discussing the extent to which they meet the proposed requirements. It is shown that existing definitions of AI do not meet the most important requirements for legal definitions. Next, the paper argues that a risk-based approach would be preferable. Rather than using the term AI, policy makers should focus on the specific risks they want to reduce. It is shown that the requirements for legal definitions can be better met by defining the main sources of relevant risks: certain technical approaches (e.g. reinforcement learning), applications (e.g. facial recognition), and capabilities (e.g. the ability to physically interact with the environment). Finally, the paper discusses the extent to which this approach can also be applied to more advanced AI systems.

1. Introduction

Policy makers around the world are currently working on AI regulations.[1] In 2021, the European Commission published a proposal for an Artificial Intelligence Act (AI Act),[2] which is generally seen as the first comprehensive attempt to regulate AI in a major jurisdiction. The US has been more hesitant so far.

---

* Research Fellow, Centre for the Governance of AI, Oxford, UK; Research Affiliate, Legal Priorities Project, Cambridge, MA, USA; PhD Candidate, Faculty of Law, Goethe University Frankfurt, Germany; jonas.schuett@governance.ai.

[1] By 'regulation', I mean 'sustained and focused attempts to change the behaviour of others in order to address a collective problem or attain an identified end or ends, usually but not always through a combination of rules or norms and some means for their implementation and enforcement, which can be legal or non-legal', Julia Black and Andrew D Murray, 'Regulating AI and Machine Learning: Setting the Regulatory Agenda' (2019) 10 *European Journal of Law and Technology*, https://perma.cc/A456-QPHH; see also Christel Koop and Martin Lodge, 'What is Regulation? An Interdisciplinary Concept Analysis' (2017) 11 *Regulation and Governance* 95, https://doi.org/10.1111/rego.12094.

[2] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC.

Under the Trump administration, the focus was more on removing regulatory barriers,[3] but this focus has shifted under the Biden administration. Although there has been some work at the federal level,[4] most efforts to regulate AI seem to take place at the state level.[5] China's approach to AI regulation has also changed over the past two years. Initially, the focus was on voluntary AI ethics principles similar to those published by Western institutions.[6] But since 2020, more stringent regulations for AI companies were introduced, while state use of AI remains completely unrestricted.[7] This global dynamic has already been framed as a 'race to regulate AI'.[8]

One challenge faced by all policy makers who work on AI regulation is how to define the scope of application, which determines whether or not a regulation is applicable in a particular case. The scope of application defines *what* is regulated (material scope), *who* is regulated (personal scope), *where* the regulation applies (territorial scope), and *when* it applies (temporal scope). In this paper, I focus on the material scope. The territorial and temporal scope depend on jurisdiction-specific details, and defining the personal scope is a difficult question which deserves a paper on its own. The scope of application is described in the body of the regulation, using terms typically defined elsewhere in the regulation. These definitions are called legal definitions. The distinction between the terms that are used to define the scope of application ('this regulation applies to AI') and the definitions of these terms ('AI means…') will be important throughout this paper because the core argument is based on the conjunction between the two ('policy makers should only use the term AI for the scope definition if there is a good definition of AI').

Defining the scope of AI regulations is particularly challenging because the term AI is used for so many different systems—'it isn't any one thing'.[9] It can

---

[3] The White House, 'Maintaining American Leadership in Artificial Intelligence' (2019) Executive Order 13859, https://perma.cc/MAN8-7TJJ.

[4] E.g. Eric Lander and Alondra Nelson, 'Americans Need a Bill of Rights for an AI-Powered World' (*The White House*, 22 October 2021) https://perma.cc/6ZRX-Q9ZB.

[5] See National Conference of State Legislatures, 'Legislation Related to Artificial Intelligence' (2022) https://perma.cc/49NS-WE9Y.

[6] E.g. Beijing Academy of Artificial Intelligence, 'Beijing AI Principles' (2019) https://perma.cc/PHA3-NUGY.

[7] Jennifer Conrad and Will Knight, 'China Is About to Regulate AI—and the World Is Watching' (2022) https://perma.cc/6ACT-WW4M.

[8] Nathalie A Smuha, 'From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence' (2021) 13 *Law, Innovation and Technology* 57, https://doi.org/10.1080/17579961.2021.1898300.

[9] Peter Stone and others, 'Artificial Intelligence and Life in 2030' (*Stanford University*, 2016) https://perma.cc/36VX-Y6MM, 48.

refer to systems that play games,[10] produce coherent text,[11] predict protein structures,[12] diagnose eye diseases,[13] or control nuclear fusion reactors.[14] From a regulatory perspective, these systems have very different risk profiles and therefore must be treated differently. To further complicate things, the term AI is highly ambiguous. There is a vast spectrum of definitions,[15] and its meaning changes over time. As famously put by John McCarthy: 'as soon as it works, no one calls it AI any more'.[16]

The question of how to define AI in legal terms—especially in a regulatory context—has been raised by many legal scholars. While some have suggested the need for a single legal definition of AI,[17] others have argued that this is not

---

[10] E.g. Oriol Vinyals and others, 'Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning' (2019) 575 *Nature* 350, https://doi.org/10.1038/s41586-019-1724-z; Julian Schrittwieser and others, 'Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model' (2020) 588 *Nature* 604, https://doi.org/10.1038/s41586-020-03051-4; OpenAI and others, 'Dota 2 with Large Scale Deep Reinforcement Learning' (2019) https://arxiv.org/abs/1912.06680.

[11] E.g. Jacob Devlin and others, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding' (2018) https://arxiv.org/abs/1810.04805; Tom B Brown and others, 'Language Models are Few-Shot Learners' (2020) https://arxiv.org/abs/2005.14165; Jack W Rae and others, 'Scaling Language Models: Methods, Analysis & Insights from Training Gopher' (2021) https://arxiv.org/abs/2112.11446; Jordan Hoffmann and others, 'Training Compute-Optimal Large Language Models' (2022) https://arxiv.org/abs/2203.15556; Aakanksha Chowdhery and others, 'PaLM: Scaling Language Modeling with Pathways' (2022) https://arxiv.org/abs/2204.02311.

[12] E.g. Andrew W Senior and others, 'Improved Protein Structure Prediction Using Potentials from Deep Learning' (2020) 577 *Nature* 706, https://doi.org/10.1038/s41586-019-1923-7; Kathryn Tunyasuvunakool and others, 'Highly Accurate Protein Structure Prediction for the Human Proteome' (2021) 596 *Nature* 590, https://doi.org/10.1038/s41586-021-03828-1.

[13] E.g. Jason Yim and others, 'Predicting Conversion to Wet Age-Related Macular Degeneration Using Deep Learning' (2020) 26 *Nature Medicine* 892, https://doi.org/10.1038/s41591-020-0867-7.

[14] E.g. Jonas Degrave and others, 'Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning' (2022) 602 *Nature* 414, https://doi.org/10.1038/s41586-021-04301-9.

[15] Shane Legg and Marcus Hutter, 'A Collection of Definitions of Intelligence' (2007) https://arxiv.org/abs/0706.3639; Sofia Samoili and others, 'AI Watch: Defining Artificial Intelligence' (*European Commission*, 2020) https://doi.org/10.2760/382730.

[16] Bertrand Meyer, 'John McCarthy' (*Communications of the ACM*, 28 October 2011) https://perma.cc/49S8-3GM6.

[17] Gary Lea, 'Why We Need a Legal Definition of Artificial Intelligence' (*The Conversation*, 2 September 2015) https://perma.cc/6NZG-5KCS; Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019) https://doi.org/10.1007/978-3-319-96235-1, 7-8; Rex Martinez, 'Artificial Intelligence: Distinguishing between Types & Definitions' (2019) 19 *Nevada Law Journal* 1015, https://perma.cc/F8YN-7RKZ, 1022.

feasible.[18] However, there are three notable gaps in the current literature. First, although most arguments rely on certain requirements for legal definitions (e.g. being future-proof), there seems to be no meta-discussion about these requirements. They tend to be treated as something given, without any justification of their legal origin or appropriateness. Second, there is no comprehensive discussion of all requirements; different scholars focus on different requirements. Third, there is only limited discussion of alternative approaches.

The paper proceeds as follows. First, I argue that policy makers should not rely on the term AI to define the material scope of AI regulations. Next, I argue that policy makers should instead consider using certain technical approaches, applications, and capabilities, following a risk-based approach. Finally, I discuss the extent to which this approach can also be applied to more advanced AI systems.

2. Should policy makers use the term AI to define the material scope of AI regulations?

The most obvious way to define the material scope of AI regulations would be to use the term AI. For example, Article 2(1) of the AI Act uses the following formulation:

> This Regulation applies to (a) providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country; (b) users of AI systems located within the Union; (c) providers and users of AI systems that are located in a third country, where the output produced by the system is used in the Union.[19]

But policy makers should only use the term AI to define the scope of application if they can also define it in a way that is appropriate for regulatory purposes. The question is: does such a definition exist? To answer this question, I propose a set of requirements for legal definitions generally, survey existing AI definitions, and then discuss the extent to which they meet the requirements for legal definitions.

---

[18] Chris Reed, 'How Should We Regulate Artificial Intelligence?' (2018) 376 *Philosophical Transactions of the Royal Society A* 1, https://doi.org/10.1098/rsta.2017.0360, 2; Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 288; Miriam C Buiten, 'Towards Intelligent Regulation of Artificial Intelligence' (2019) 10 *European Journal of Risk Regulation* 41, https://doi.org/10.1017/err.2019.8, 45; Urs Gasser and Virgilio AF Almeida, 'A Layered Model for AI Governance' (2017) 21 *IEEE Internet Computing* 58, https://doi.org/10.1109/MIC.2017.4180835.

[19] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC, Art. 2(1).

2.1. Requirements for legal definitions

In democratic countries, policy makers are bound by higher-ranking sources of law, such as constitutional law and general legal principles. If regulations violate these laws or principles, they can be void or invalid—the particular effects are of course jurisdiction-specific. Here, I give a brief overview of relevant laws and principles in the EU and US and distil them into a list of requirements for legal definitions (Table 1).

Table 1: Requirements for legal definitions

Title	Description	Origin
Over-inclusiveness	Legal definitions must not be over-inclusive. A definition is over-inclusive if it includes cases which are not in need of regulation according to the regulation's objective.[20] Simply put, this is a case of too much regulation.	Principle of proportionality
Under-inclusiveness	Legal definitions must not be under-inclusive. A definition is under-inclusive if cases which should have been included are not included.[21] This is a case of too little regulation.	Effectiveness
Precision	Legal definitions must be precise. It must be possible to determine clearly whether or not a particular case falls under the definition.	Principle of legal certainty, vagueness doctrine
Understandability	Legal definitions must be understandable. Ideally, the definition should be based on the existing meaning of terms and comply with the natural use of language. At least in principle, people without expert knowledge should be able to apply the definition.	Principle of legal certainty, vagueness doctrine
Practicability	Legal definitions should be practicable. It should be possible to determine with little effort whether or not a concrete case falls under the definition. The assessment of every element of the definition should be possible on the basis of the information typically available to legal practitioners.	Good legislative practice (helps to maintain the efficiency of the judicial system)
Flexibility	Legal definitions should be flexible. They should be able to accommodate technical progress. They should only contain elements which are unlikely to change in the foreseeable future.	Good legislative practice (helps to prevent the need for regulatory updating)

[20] Robert Baldwin, Martin Cave and Martin Lodge, *Understanding Regulation: Theory, Strategy, and Practice* (Oxford University Press 2011), 70.

[21] Ibid.

Regulations in the EU must comply with the *principle of proportionality*. Pursuant to Article 5(4) of the Treaty on European Union, 'the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties.' Although proportionality has not been used as a general principle of constitutional law in the US, it has nonetheless been recognized as an element of constitutional doctrine in several areas of contemporary constitutional law.[22]

EU regulations must further comply with the *principle of legal certainty*. According to the Court of Justice of the European Union, policy makers are required to ensure 'that Community rules enable those concerned to know precisely the extent of the obligations which are imposed on them. Individuals must be able to ascertain unequivocally what their rights and obligations are and take steps accordingly.'[23]

The US *vagueness doctrine*, which is rooted in due process considerations, has similar implications. According to the US Supreme Court, 'a statute which either forbids or requires the doing of an act in terms so vague that men of common intelligence must necessarily guess at its meaning and differ as to its application violates the first essential of due process of law.'[24] Put differently, 'legal protection requires that texts intended in the first place for use by lawyers should be easily understandable by every citizen.'[25]

Finally, regulations should be *effective*. Here, effectiveness refers to the degree to which a given regulation achieves or progresses towards its objectives. It is worth noting that the concept of effectiveness is highly controversial within legal research,[26] but for the purposes of this paper, the debate has no relevant implications.

To the best of my knowledge, a list similar to Table 1 does not currently exist. Existing lists of requirements for AI definitions[27] and scientific definitions in general[28] do not take a legal perspective. And although most of the

---

[22] Vicki C Jackson, 'Constitutional Law in an Age of Proportionality' (2015) 124 *Yale Law Journal* 2680, https://perma.cc/B7HB-5NW4, 3104.

[23] Case C-345/06, *Gottfried Heinrich* (2009) ECR I-01659, https://perma.cc/6YML-D4BW.

[24] *Connally v. General Construction Co.* (1926) 269 US 385, https://perma.cc/C2WR-9H2Q.

[25] Heikki ES Mattila, *Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas* (Routledge 2013), 46; Jeanne Price, 'Wagging, Not Barking: Statutory Definitions' (2013) 60 *Cleveland State Law Review* 999, https://perma.cc/VAH7-YNBP, 1031.

[26] See Maria De Benedetto, 'Effective Law from a Regulatory and Administrative Law Perspective' (2018) 9 *European Journal of Risk Regulation* 391, https://doi.org/10.1017/err.2018.52.

[27] Pei Wang, 'On Defining Artificial Intelligence' (2019) 10 *Journal of Artificial General Intelligence* 1, https://doi.org/10.2478/jagi-2019-0002, 3-6.

[28] Rudolf Carnap, *Logical Foundations of Probability* (University of Chicago Press 1950), https://perma.cc/QE4G-YAZ5, 7.

above mentioned requirements have been discussed in legal scholarship,[29] there seems to be no comprehensive discussion of all requirements. As mentioned above, different scholars focus on different requirements, which tend to be treated as something given and are rarely, if ever, linked to their legal origin.

It is worth noting that the list of requirements should be taken with a grain of salt for two reasons. First, this discussion of the legal origins considers only EU and US laws and principles. Consideration of other jurisdictions was beyond the scope of this paper. However, since the underlying rationale is often not jurisdiction-specific, I expect the list to be useful in other jurisdictions as well. Second, this list is unlikely to be exhaustive. There will likely be further requirements in certain jurisdictions. Similarly, some of the requirements might not be as relevant in some jurisdictions as they are in others, or they might take a slightly different form. For example, it seems plausible that different applications of proportionality analysis lead to different interpretations of over-inclusiveness.[30] But these variations seem to be a necessary consequence of my attempt to define requirements that are relevant for policy makers worldwide. In any case, the requirements can be used to evaluate existing definitions of AI and can be adapted to the requirements of different jurisdictions.

---

[29] The problem of over- and under-inclusive AI definitions is discussed by Lyria B Moses, 'Recurring Dilemmas: The Law's Race to Keep up with Technological Change' (2007) 2 *Journal of Law, Technology & Policy* 239, https://perma.cc/4EKU-RV6J, 260-264; Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 361-362, 373; Chris Reed, 'How Should We Regulate Artificial Intelligence?' (2018) 376 *Philosophical Transactions of the Royal Society A* 1, https://doi.org/10.1098/rsta.2017.0360, 2; Rex Martinez, 'Artificial Intelligence: Distinguishing between Types & Definitions' (2019) 19 *Nevada Law Journal* 1015, https://perma.cc/F8YN-7RKZ, 1038; Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 325, 327-328; Miriam C Buiten, 'Towards Intelligent Regulation of Artificial Intelligence' (2019) 10 *European Journal of Risk Regulation* 41, https://doi.org/10.1017/err.2019.8, 45. Precision and understandability are addressed by Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 373; Rex Martinez, 'Artificial Intelligence: Distinguishing between Types & Definitions' (2019) 19 *Nevada Law Journal* 1015, https://perma.cc/F8YN-7RKZ, 1035; and flexibility by Lyria B Moses, 'Recurring Dilemmas: The Law's Race to Keep up with Technological Change' (2007) 2 *Journal of Law, Technology & Policy* 239, https://perma.cc/4EKU-RV6J; Rex Martinez, 'Artificial Intelligence: Distinguishing between Types & Definitions' (2019) 19 *Nevada Law Journal* 1015, https://perma.cc/F8YN-7RKZ, 1017; Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 357.

[30] See Vicki C Jackson, 'Constitutional Law in an Age of Proportionality' (2015) 124 *Yale Law Journal* 2680, https://perma.cc/B7HB-5NW4.

2.2. Existing definitions of AI

There is no generally accepted definition of the term AI. Since its first usage by McCarthy et al.,[31] a vast spectrum of definitions has emerged. Below, I provide an overview of existing AI definitions. A more comprehensive collection of definitions can be found in relevant literature.[32] Categorizations of different AI definitions have been proposed by Russell and Norvig,[33] Wang,[34] and Bhatnagar et al.[35] The OECD has also published a Framework for the Classification of AI Systems, which is explicitly targeted at policy makers.[36]

The following list contains popular AI definitions which have been proposed by computer scientists and philosophers:

> The science of making machines do things that would require intelligence if done by men.[37]
>
> The art of creating machines that perform functions that require intelligence when performed by people.[38]
>
> The science and engineering of making intelligent machines, especially intelligent computer programs … Intelligence is the computational part of the ability to achieve goals in the world.[39]
>
> That activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.[40]
>
> The study of agents that receive percepts from the environment and perform actions.[41]

Some legal scholars have also proposed definitions of AI:[42]

---

[31] John McCarthy and others, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence' (1955) https://perma.cc/S9DU-GWFF.

[32] Shane Legg and Marcus Hutter, 'A Collection of Definitions of Intelligence' (2007) https://arxiv.org/abs/0706.3639; Sofia Samoili and others, 'AI Watch: Defining Artificial Intelligence' (*European Commission*, 2020) https://doi.org/10.2760/382730.

[33] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2020).

[34] Pei Wang, 'On Defining Artificial Intelligence' (2019) 10 *Journal of Artificial General Intelligence* 1, https://doi.org/10.2478/jagi-2019-0002.

[35] Sankalp Bhatnagar and others, 'Mapping Intelligence: Requirements and Possibilities' in Vincent C Müller (ed), *Philosophy and Theory of Artificial Intelligence* (Springer 2018), https://doi.org/10.1007/978-3-319-96448-5_13.

[36] OECD, 'Framework for the Classification of AI Systems' (2022) https://doi.org/10.1787/cb6d9eca-en.

[37] Marvin Minsky, *Semantic Information Processing* (MIT Press 1969), v.

[38] Ray Kurzweil, *The Age of Intelligent Machines* (MIT Press 1990), 14.

[39] John McCarthy, 'What is Artificial Intelligence?' (12 November 2007) https://perma.cc/QL9Y-AY8A, 2.

[40] Nils J Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (Cambridge University Press 2009), https://perma.cc/CQV7-N233, xiii.

[41] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2020), vii.

[42] It is worth noting that none of the definitions is intended to be used to define the scope of AI regulations. Scherer only wants to 'discuss the definitional problems that regulators will have to confront' (p. 359), while Turner's definition is meant as a 'core definition which

Machines that are capable of performing tasks that, if performed by a human, would be said to require intelligence.[43]

The ability of a non-natural entity to make choices by an evaluative process.[44]

A system, program, software, or algorithm that acts autonomously to think rationally, think humanely, act rationally, act humanely, make decisions, or provide outputs.[45]

AI definitions in policy proposals are particularly relevant for this paper:

Software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.[46]

(1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets. (2) An artificial system developed in computer software, physical hardware, or another context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action. (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks. (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task. (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.[47]

A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.[48]

The use of digital technology to create systems capable of performing tasks commonly thought to require intelligence.[49]

It is worth highlighting a few characteristics of these definitions before continuing with the legal analysis. For example, some of the proposed definitions refer to disciplines ('the science of', 'the art of', 'the study of') and others to

---

captures the essence of a term, without delimiting its precise boundaries' (p. 21), and Martinez acknowledges that his definition 'is going to be under- or over-inclusive' (p. 1038).

[43] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 362.

[44] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019) https://doi.org/10.1007/978-3-319-96235-1, 16.

[45] Rex Martinez, 'Artificial Intelligence: Distinguishing between Types & Definitions' (2019) 19 *Nevada Law Journal* 1015, https://perma.cc/F8YN-7RKZ, 1038.

[46] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC, Art. 3(1).

[47] Section 238(g) of the FY2019 National Defense Authorization Act; also used by Russell T Vought, 'Guidance for Regulation of Artificial Intelligence Applications' (*The White House*, 17 November 2020) https://perma.cc/U2V3-LGV6, 1.

[48] OECD, 'Recommendation of the Council on Artificial Intelligence' (2019) OECD/LEGAL/0449, https://perma.cc/M6Z7-BESV, 7

[49] Office for AI, 'A Guide to Using Artificial Intelligence in the Public Sector' (2019) https://perma.cc/8XQU-LRNB, 6.

systems ('software system', 'artificial system', 'machine-based system'). Most serve academic purposes, while only a few are intended to be used in regulations. One might therefore be tempted to only focus on the definitions by policy makers; however, these definitions are often inspired by academic definitions—for example, the definition in Section 238(g) of the FY2019 National Defense Authorization Act is heavily influenced by Russell and Norvig[50]—thus it seems worthwhile to discuss a wider range of definitions.

2.3. Do existing AI definitions meet the requirements for legal definitions?

As outlined above, legal definitions must meet a number of requirements that can be derived from prior-ranking law, or are at least considered good legislative practice. In Table 2, I discuss the extent to which existing AI definitions meet these requirements using the evaluation options 'Yes', 'No', 'Debatable', and 'Unknown'. Although these options give the false impression that the requirements are binary, they are used for convenience. Since courts ultimately have to make yes-or-no decisions (e.g. whether or not a provision is proportionate), this simplification seems acceptable. It goes without saying that the evaluation is necessarily subjective.

Table 2: Do existing AI definitions meet the requirements for legal definitions?

Requirements	Existing definitions of AI
Over-inclusiveness	*No.* Existing AI definitions are highly over-inclusive. For example, many systems that are able to achieve goals in the world are clearly not in need of regulation (e.g. game-playing agents). The same holds true for systems that can, for a given set of human-defined objectives, generate outputs that influence their environment.
Under-inclusiveness	*No.* Some AI definitions are also under-inclusive. For example, systems which do not achieve their goals—like an autonomous vehicle that is unable to reliably identify pedestrians—would be excluded, even though they can pose significant risks.[51] Similarly, the Turing test[52] excludes systems that do not communicate in natural language, even though such systems may need regulation (e.g. autonomous vehicles).
Precision	*No.* Existing AI definitions are highly vague. Many of them define AI in comparison to human intelligence, even though it is highly disputed how

[50] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2020), 1-5.

[51] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 362.

[52] Alan M Turing, 'Computing Machinery and Intelligence' (1950) 59 *Mind* 433, https://doi.org/10.1093/mind/LIX.236.433.

	human intelligence should be defined.[53] Other definitions simply replace one difficult-to-define term ('intelligence') with another ('goal').[54] Russell and Norvig's rational agent definition[55] is equally vague, especially with regards to its notion of limited rationality. In complex environments, agents are often unable to take the optimal action. It is therefore sufficient if they take the action that is optimal *in expectation*. However, in many cases, it is impossible to determine ex-ante whether or not a concrete action is expected to be optimal because ground truth is unattainable. Even if it were, no system can always select the optimal action. How often does a system need to take the optimal action in order to be considered rational?
Understandability	*Debatable.* It is debatable whether existing definitions are understandable. The term seems intuitive at first glance—it is simply a compound of two commonly used terms: 'artificial' and 'intelligence'. However, as mentioned above, it is far from obvious what intelligence actually means. The intuitive meaning may also be misleading. Due to pop-cultural illustrations of AI, people might anthropomorphize AI.[56]
Practicability	*Debatable.* The practicability of many definitions is also debatable. It may be possible to determine whether or not a system is able to achieve its goals on the basis of typically available information. The Turing test,[57] however, would be highly impracticable. Courts would not be able to conduct the test every time they have to decide whether or not a system is considered AI by the law.
Flexibility	*Yes.* The definitions seem sufficiently flexible. The fact that some of them are decades old suggests that they can accommodate technical progress. They also seem relatively general and technology-neutral. One could argue that the so-called 'AI effect' speaks against their flexibility. As McCarthy puts it: 'as soon as it works, no one calls it AI any more'.[58] However, this effect only applies to what is generally considered to be AI. It does not necessarily provide a counterargument against the flexibility of specific definitions.

[53] Shane Legg and Marcus Hutter, 'A Collection of Definitions of Intelligence' (2007) https://arxiv.org/abs/0706.3639.

[54] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 361.

[55] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2020).

[56] Arleen Salles, Kathinka Evers and Michele Farisco, 'Anthropomorphism in AI' (2020) 11 *AJOB Neuroscience* 88, https://doi.org/10.1080/21507740.2020.1740350; Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 353-355.

[57] Alan M Turing, 'Computing Machinery and Intelligence' (1950) 59 *Mind* 433, https://doi.org/10.1093/mind/LIX.236.433.

[58] Bertrand Meyer, 'John McCarthy' (*Communications of the ACM*, 28 October 2011) https://perma.cc/49S8-3GM6.

Taken together, existing definitions of AI do not meet the most important requirements for legal definitions. They are highly over-inclusive and vague, while their understandability and practicability are debatable. I doubt that there even is a definition which meets all of the requirements. I would argue that definitions of the term AI are inherently over-inclusive and vague. Due to its broadness, the term will always include many different systems with very different risk profiles which must be treated differently. 'We just need a better definition' would therefore be the wrong conclusion. Relatedly, it would be wrong to deploy a social definition of AI, according to which 'AI is what people generally consider to be AI'.[59] Such a definition would not only be circular, it would also not meet the requirements for legal definitions, as it is inherently vague.

One might object that vagueness is an inherent property of many legal definitions.[60] Many laws use imprecise language, but courts have been able to deal with it. Why should the term AI be any different? My response to this objection is twofold. First, vagueness is a matter of degree. It would be wrong to assume that, simply because courts have been able to deal with imprecise language in the past, policy makers can ignore the issue completely. It might be necessary to use terms that are somewhat imprecise, but I would argue that the term AI is close to the edge of the vagueness spectrum. Second, even if policy makers used a single definition of AI, the above mentioned problems would simply be deferred to the judiciary. Courts would have to develop a casuistry which would also have to meet the requirements detailed above. This would not change the nature of the problem, only the actor who has to solve it.

One might insist that the judiciary would in fact be better suited to develop a precise definition of AI.[61] I do not argue against this claim, as it seems to be a matter of legal tradition. Scholars from civil law countries (like me) tend to favour statutory definitions, while common law scholars are more used to definitions developed by courts.

Finally, one might point out that the proposed AI Act does use a single definition of AI.[62] Am I really suggesting that the proposal does not meet the

---

[59] Peter Cihon and others, 'Corporate Governance of Artificial Intelligence in the Public Interest' (2021) 12 *Information*, https://doi.org/10.3390/info12070275.

[60] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 373.

[61] Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 341-344; Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019) https://doi.org/10.1007/978-3-319-96235-1, 21.

[62] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC, Art. 3(1).

requirements for legal definitions? Again, my response would be twofold. First, I would argue that their definition of AI mostly serves symbolic purposes. The substance lies in Annex I, which contains a list of technical approaches, and Annex III, which contains a list of high-risk applications. In other words, the material scope is only superficially defined by the term AI. Upon closer examination, the term is an 'empty shell'.[63] Overall, their approach is similar to the one I will suggest below. Second, the European Commission was well aware of the above mentioned requirements. The fact that they explain at length why their approach is future-proof, proportionate, and increases legal certainty[64] suggests that, in their view, other approaches might not meet these requirements.

In summary, the results of my discussion seem defensible against plausible objections. I therefore recommend that the material scope should not rely on the term AI. Having said that, I do believe that there is value in using the term for communication purposes. For example, policy makers can still call it 'AI regulation'. They might even use the term to define the material scope, as long as it does not play a substantive role.[65]

3. What should they do instead?

For the substance of the scope definition, policy makers should take a risk-based approach. Risk-based regulation tries to achieve policy objectives by targeting activities that pose the highest risk, while leaving lower-risk activities unencumbered.[66] The scope of such regulations is defined by the risks they want to address. As Turner puts it, policy makers should not ask 'what is AI?', but 'why do we need to define AI at all?', and 'what is the unique factor of AI that needs regulation?'[67] Or in the words of Casey and Lemley: 'We don't need rules that decide whether a car with certain autonomous features is or is not a

---

[63] It is worth noting that the term is not completely 'empty'. For example, the fact that doing Bayesian statistics on paper is not covered by the scope is because the AI definition in Art. 3(1) requires an AI system to be software.

[64] Ibid, 3, 7, 10.

[65] If the term is indeed used for the scope definition, it is important that the corresponding definition of AI is very broad and does not exclude relevant systems. Many of the above-mentioned definitions seem to meet these requirements.

[66] See Robert Baldwin and Julia Black, 'Driving Priorities in Risk-based Regulation: What's the Problem?' (2016) 43 *Journal of Law and Society* 565, https://doi.org/10.1111/jols.12003, 565.

[67] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019), https://doi.org/10.1007/978-3-319-96235-1, 8, 15.

robot. What we actually need are rules that regulate unsafe driving behaviour.'[68]

This approach is in line with existing policy proposals. For example, in their proposal for an AI Act, the European Commission focuses on high-risk applications, with almost no requirements for systems with low or minimal risk.[69] They also report that most of the respondents to their stakeholder consultation were explicitly in favour of a risk-based approach.[70] Similarly, the German Data Ethics Commission proposes a pyramid of five levels of criticality.[71]

There is an extensive body of literature on risks from AI. Risks have been conceptualised as accident risks,[72] misuse risks,[73] and structural risks.[74] One could also distinguish between near-term and long-term risks, but some scholars have argued convincingly that this distinction is not always useful, mainly because many ethics and safety issues span different time horizons.[75]

There has also been some work on AI risk factors, broadly defined as all factors that contribute to risks from AI. Most notably, Hernández-Orallo et al. have conducted a survey of known safety-relevant characteristics of AI.[76] They distinguish between (1) internal characteristics (e.g. interpretability), (2) effect of the external environment on the system (e.g. the ability of the operator to

---

[68] Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 342-343.

[69] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC, 12.

[70] Ibid, 8.

[71] German Data Ethics Commission, 'Opinion of the Data Ethics Commission' (2019) https://perma.cc/23QM-JNLJ, 177.

[72] Dario Amodei and others, 'Concrete Problems in AI Safety' (2016) https://arxiv.org/abs/1606.06565; Zachary Arnold and Helen Toner, 'AI Accidents: An Emerging Threat' (*Center for Security and Emerging Technology*, July 2021) https://doi.org/10.51593/20200072.

[73] Miles Brundage and others, 'Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (2018) https://arxiv.org/abs/1802.07228.

[74] Remco Zwetsloot and Allan Dafoe, 'Thinking About Risks from AI: Accidents, Misuse and Structure' (*Lawfare*, 11 February 2019) https://perma.cc/7S3K-6L4U.

[75] Seth D Baum, 'Reconciliation between Factions Focused on Near-Term and Long-Term Artificial Intelligence' (2018) 33 *AI & Society* 565, https://doi.org/10.1007/s00146-017-0734-3; Stephen Cave and Seán ÓhÉigeartaigh, 'Bridging Near- and Long-Term Concerns About AI' (2019) 1 *Nature Machine Intelligence* 5, https://doi.org/10.1038/s42256-018-0003-2; Carina Prunkl and Jess Whittlestone, 'Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society' (2020) *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 138, https://doi.org/10.1145/3375627.3375803.

[76] José Hernández-Orallo and others, 'Surveying Safety-relevant AI Characteristics' (2019) *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019* 57, https://perma.cc/HDS9-3LA2.

intervene during operation), and (3) effect of the system on the external environment (e.g. whether the system influences a safety-critical setting).

Although their categorization is convincing, I do not use it below, mainly because it serves a different purpose. Theirs is intended to reveal neglected areas of research and to suggest design choices for reducing certain safety concerns, whereas I am interested in defining sources of AI risk in a way that meets the requirements for legal definitions. Their categorization also excludes risks caused by 'the malicious or careless use of a correctly-functioning system', which would be relevant in a regulatory context. For similar reasons, I also do not use the categorization by Burden and Hernández-Orallo.[77]

Instead, I use my own simple categorization of the main sources of risks from AI. I distinguish between (1) technical approaches ('how it is made'), (2) applications ('what it is used for'), and (3) capabilities ('what it can do').[78] In the following, I explain each of the three categories along with examples and discuss the extent to which they meet the requirements for legal definitions.

3.1. Technical approaches

Some AI risks are directly linked to certain technical approaches. One such approach is *reinforcement learning*, which is used in games,[79] robotics,[80] recommender systems[81] and nuclear fusion reactors.[82] But using this approach poses a number of inherent risks. For example, if the objective function of a reinforcement learning agent contains explicit specifications only regarding the main goal, it might implicitly express indifference towards other aspects of the environment. This can lead to situations where the agent disturbs its environment in negative ways while pursuing its main goal. This problem is typically

---

[77] John Burden and José Hernández-Orallo, 'Exploring AI Safety in Degrees: Generality, Capability and Control' (2020) *Proceedings of the Workshop on Artificial Intelligence Safety co-located with 34th AAAI Conference on Artificial Intelligence* 36, https://perma.cc/98QU-FVBM.

[78] Note that the categorisation is not intended to be mutually exclusive. As I will discuss below, I recommend using elements of multiple categories to narrow down the scope. The list is probably also not exhaustive, although I do believe that it captures the vast majority of relevant sources of risks.

[79] Julian Schrittwieser and others, 'Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model' (2020) 588 *Nature* 604, https://doi.org/10.1038/s41586-020-03051-4.

[80] Julian Ibarz and others, 'How to Train Your Robot with Deep Reinforcement Learning: Lessons We have Learned' (2020) 14 *The International Journal of Robotics Research* 698, https://doi.org/10.1177/0278364920987859.

[81] M Medhi Afsar, Trafford Crump and Behrouz Far, 'Reinforcement Learning Based Recommender Systems: A Survey' (2021) https://arxiv.org/abs/2101.06286.

[82] Jonas Degrave and others, 'Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning' (2022) 602 *Nature* 414, https://doi.org/10.1038/s41586-021-04301-9.

referred to as 'negative side effects'.[83] Another problem is 'reward hacking', the exploitation of unintended loopholes in the reward function.[84] A third problem is how we can ensure that agents can be safely interrupted at any time.[85] Policy makers who want to address these risks could use the following definition:

> 'Reinforcement learning' means the machine learning task of learning a policy from reward signals that maximises a value function.[86]

Policy makers could also use the terms *supervised learning* and *unsupervised learning* to define the material scope of AI regulations. These approaches are used in a wide range of different systems, including systems that support judicial decision-making[87] or select employees.[88] However, both approaches can lead to discrimination by reproducing biases contained in the training data.[89] They can be defined as follows:

> 'Supervised learning' means the machine learning task of learning a function that maps from an input to an output based on labelled input-output pairs.[90]
>
> 'Unsupervised learning' means the machine learning task of learning patterns in an input even though no explicit feedback is supplied.[91]

Although it can be important to specify certain technical approaches, they should usually not be the main element of the scope definition. I expect them to be more relevant at lower levels of abstractions, assuming that vague provisions are specified in guidelines or standards.

---

[83] Dario Amodei and others, 'Concrete Problems in AI Safety' (2016) https://arxiv.org/abs/1606.06565, 4-7; Victoria Krakovna and others, 'Penalizing Side Effects Using Stepwise Relative Reachability' (2018) https://arxiv.org/abs/1806.01186.

[84] Jack Clark and Dario Amodei, 'Faulty Reward Functions in the Wild' (*OpenAI*, 21 December 2016) https://perma.cc/6HAB-4BWZ.

[85] Laurent Orseau and Stuart Armstrong, 'Safely Interruptible Agents' (*Machine Intelligence Research Institute*, 28 October 2016) https://perma.cc/RYV9-34QL.

[86] Richard S Sutton and Andrew G Barto, *Reinforcement Learning: An Introduction* (MIT Press 2018), 6.

[87] Julia Angwin and others, 'Machine Bias' (*ProPublica*, 23 May 2016) https://perma.cc/KA6N-WG37.

[88] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women' (*Reuters*, 11 October 2018) https://perma.cc/CMT4-L468.

[89] Tolga Bolukbasi and others, 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings' (2016) *Proceedings of the 30th International Conference on Neural Information Processing Systems* 4356, https://perma.cc/V6PD-TN5Z; Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 77, https://perma.cc/8TEZ-M3GQ.

[90] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2020), 652-653.

[91] Ibid.

3.2. Applications

Other risks are not linked to technical approaches, but certain applications. By 'application', I mean a system's use-case within its socio-technical context, including what it is used for, how it is used, who uses it, and what their intentions are.[92] Although the concept is a bit fuzzy, we can distinguish between different subcategories of applications, such as a system's general task (e.g. making recommendations or generating content),[93] its sector-specific use-case (e.g. autonomous driving or automated trading), or its role in the deployment process (e.g. whether it is a foundation model[94] or a fine-tuned model).[95]

Autonomous driving is a typical example of an application. Policy makers may want to reduce the risks that autonomous driving poses to road safety and security, physical integrity, and property rights. The material scope of such regulations could be defined using six levels of automation, as described in the technical standard SAE J3016.[96] These definitions have already been adopted by policy makers in the US[97] and the EU.[98]

Policy makers may also want to reduce the specific risks of *facial recognition technology*. A number of studies show that facial recognition technology can have gender or race biases.[99] This is particularly worrying if such systems are used for law enforcement purposes. In the US, some municipalities have therefore started to ban state use of facial recognition technology for law enforcement purposes, including San Francisco[100] and Boston.[101] The European Commission has proposed a similar ban in the EU, with a few narrow

---

[92] This is related to the socio-technical characteristics specified by NIST, 'AI Risk Management Framework: Initial Draft' (2022) https://perma.cc/FGM8-5TTG, 10-12.

[93] I owe this idea to Markus Anderljung.

[94] 'Foundation models' are large pre-trained models that can serve as the foundation for a wide array of down-stream applications. Some predict that their use will be increasingly widespread, Rishi Bommasani and others, 'On the Opportunities and Risks of Foundation Models' (2021) https://arxiv.org/abs/2108.07258.

[95] Note that the subcategories are neither mutually exclusive nor exhaustive.

[96] SAE International, 'Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles' (2021) J3016_202104, https://perma.cc/7LD6-48VG.

[97] US Department of Transportation, 'Preparing for the Future of Transportation' (2018) https://perma.cc/FPJ3-VELU.

[98] European Commission, 'On the Road to Automated Mobility: An EU Strategy for Mobility of the Future' COM (2018) 283 final, https://perma.cc/5ZXR-YUXD.

[99] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 77, https://perma.cc/8TEZ-M3GQ.

[100] Kate Conger, Richard Fausset and Serge F Kovaleski, 'San Francisco Bans Facial Recognition Technology' (*The New York Times*, 14 May 2019) https://perma.cc/B8KQ-P8WX.

[101] Khari Johnson, 'Boston Bans Facial Recognition Due to Concern About Racial Bias' (*VentureBeat*, 24 June 2020) https://perma.cc/G635-AGCY.

exceptions.[102] In addition to discrimination risks, facial recognition also raises severe privacy concerns.[103] Policy makers who want to address these risks could use the following definition:

> 'Facial recognition' means the automatic processing of digital images which contain the faces of individuals for identification, authentication/verification or categorisation of those individuals.[104]

Overall, I expect applications to be the most important element of the scope definition, especially tasks and sectoral use-cases.

3.3. Capabilities

A third category of sources of AI risk is a system's capabilities. For example, policy makers may want to limit the material scope to systems which can *physically interact with their environment* via robotic hands or other actuators. Only embodied systems can directly cause physical harm or damage property.[105] This ability could be defined as follows:

> 'Physical interaction' means the ability to use sensors to perceive the physical environment and effectors to manipulate this environment.[106]

Another capability-related source of AI risk is the *ability to make automated decisions*. This element can be used to exclude systems which only make suggestions while humans make the final decision. One could call systems with this ability 'self-executive'. Policy makers could use this element to address certain risks resulting from a loss of control[107] and other assurance risks—those risks which stem from an operator's inability to understand and control AI

---

[102] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC.

[103] Zekeriya Erkin and others, 'Privacy-Preserving Face Recognition' (2009) *Proceedings of the 9th International Symposium on Privacy Enhancing Technologies* 235, https://doi.org/10.1007/978-3-642-03168-7_14.

[104] Article 29 Data Protection Working Party, 'Opinion 02/2012 on Facial Recognition in Online and Mobile Services' (2012) WP192, https://perma.cc/Y72E-WAX3, 2.

[105] José Hernández-Orallo and others, 'Surveying Safety-relevant AI Characteristics' (2019) *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019* 57, https://perma.cc/HDS9-3LA2, 58.

[106] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson 2009).

[107] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 *Harvard Journal of Law & Technology* 353, https://perma.cc/2CK2-59EK, 366-369.

systems during operation.[108] This element is already being used in Articles 13(2)(f), 14(2)(g) and 15(1)(h) of the GDPR. It can be defined as follows:

> 'Automated decision-making' means the ability to make decisions by technological means without human involvement.[109]

A third example of a capability is the *ability to make decisions which have a legal or similarly significant effect*. Consider two virtual assistants: one reminds you on your friends' birthdays, the other is able to buy products. Clearly, the two systems have very different risk profiles (the latter may require some degree of consumer protection, for example). This element is already being used in Article 22 of the GDPR. The European Data Protection Board has endorsed the definition by the Article 29 Data Protection Working Party:[110]

> 'Legal effect' means any impact on a person's legal status or their legal rights.
> 'Similarly significant effect' means any equivalent impact on a person's circumstances, behaviour, or choices. This may include their financial circumstances, access to health services, employment opportunities or access to education.

The main role of this class of elements is to narrow down the scope. It should not be the central element.

3.4. Do definitions of certain technical approaches, applications, and capabilities meet the requirements for legal definitions?

Let us now examine to what extent definitions of certain technical approaches, applications, and capabilities meet the requirements for legal definitions. Table 3 breaks down the discussion by category and requirement.

---

[108] Pedro A Ortega and Vishal Maini, 'Building Safe Artificial Intelligence: Specification, Robustness, and Assurance' (*Medium*, 27 September 2018) https://perma.cc/L7PK-LC46.

[109] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (2018) WP251rev.01, https://perma.cc/864E-R5MJ, 8.

[110] Ibid, 21-22.

Table 3: Do definitions of certain technical approaches, applications, and capabilities meet the requirements for legal definitions?

Requirements	Technical approaches	Applications	Capabilities
Over-inclusiveness	*No.* There will always be systems that use one of the above mentioned technical approaches, but should not be subject to regulation (e.g. game-playing agents based on reinforcement learning).	*Yes.* In many cases, the main regulatory goal will be to reduce certain application-specific risks (e.g. discriminatory recommender systems used to support judicial decision-making).	*No.* Not all systems with certain capabilities pose risks which are in need of regulation. For example, industrial robots and vending machines both have the ability to physically manipulate their environment, but their risk profile is very different.
Under-inclusiveness	*No.* Relevant risks can not be attributed to a single technical approach. For example, supervised learning is not inherently risky. And if a definition lists many technical approaches, it would likely be over-inclusive.	*No.* Not all systems that are applied in a specific context pose the same risks. Many of the risks also depend on the technical approach.	*No.* Relevant risks can not be attributed to a certain capability alone. By its very nature, capabilities need to be combined with other elements (‘capability of something’).
Precision	*Yes.* It is easy to determine whether or not a system is based on a certain technical approach.	*Yes.* Applications can be defined precisely. This is by no means a novel challenge for the law.	*Yes.* In many cases, capabilities can be defined in a binary way (e.g. a system either can physically manipulate its environment or not).
Understandability	*Yes.* For developers it will be easy to understand definitions of certain technical approaches. One can expect the same from non-technical people who are responsible for the development, deployment, or use of systems.	*Yes.* There are no apparent reasons for why definitions of applications are not understandable.	*Yes.* Most capabilities are intuitive (e.g. the ability to physically manipulate its environment).

Practicability	*Yes.* The required information about the technical approach is easy to obtain.	*Yes.* The required information about the application is easy to obtain.	*Yes.* Some capabilities already have established legal definitions (e.g. the ability to make decisions which have a legal or similarly significant effect).
Flexibility	*Unknown.* It is highly uncertain whether today's technical approaches will be used in the future. Definitions will be more flexible if the technical approach is defined broadly, but they will also be less precise.	*Debatable.* While some applications are unlikely to change in the future, almost certainly new applications will emerge.	*Yes.* Definitions of capabilities seem to be able to accommodate technical progress.

In summary, definitions of certain technical approaches, applications, and capabilities meet more of the requirements for legal definition than definitions of the term AI (see Table 2). This suggests that policy makers should favour a risk-based approach over the 'classical' approach.

One might be tempted to simply pick one of three categories for the scope definition, but I would argue that a *multi-element approach* is often preferable.[111] The following example illustrates the idea:

> This regulation applies to facial recognition systems for law enforcement purposes based on supervised learning.

In the example, the material scope is defined by a certain application (facial recognition for law enforcement purposes) and a certain technical approach (supervised learning). This approach allows policy makers to target risks in a more fine-grained way and thereby reduce over-inclusiveness and increase precision.

Relatedly, there will always be cases which fall under the scope definition, but should not be included. To further reduce over-inclusiveness, policy makers can use exemptions.[112] For example, Article 2(3) of the AI Act contains the following exemption:

---

[111] Bryan Casey and Mark A Lemley, 'You Might Be a Robot' (2019) 105 *Cornell Law Review* 287, https://perma.cc/Y989-ZDXG, 356.

[112] Note that there is a difference between 'not covered by the scope' and 'exempt from scope'. In the first case, the regulation is not applicable. In the second case, the regulation is applicable, but it explicitly states that it should not apply to a specific case.

> This Regulation shall not apply to AI systems developed or used exclusively for military purposes.[113]

Exemptions can be located at the beginning of the regulation (e.g. within the definition of the material scope), in the body of the regulation (e.g. within particular chapters or norms), or both. In the context of risk-based AI regulation, exemptions will typically cover cases where negative effects are low-probability, low-impact (scale), where systems only affect a small number of people (scope), and/or where people can decide not to be subject to the effects of the system (optionality).[114] Exemptions may also be used to exclude areas which are already governed by other areas of law (e.g. military use of AI). Overall, I expect most AI regulations to benefit from exemptions in one way or another.

The proposed AI Act takes a similar approach to the one I recommend.[115] As mentioned above, the material scope is not really defined by the term AI. Instead, the scope definition combines a number of technical approaches (Annex I) with certain high-risk applications (Annex III).[116] Although this seems like a reasonable approach, I would point out three potential areas for improvement. First, to the extent that my observation is correct, the European Commission should consider making it explicit that their scope definition does not rely on the term AI (e.g. in the recitals). This could help to prevent misconceptions among laypeople (e.g. the false interpretation that the regulation would apply to any use of Bayesian statistics[117]). Second, they should consider distinguishing between different technical approaches. In the current version, it is sufficient if a system is based on *any* of the technical approaches listed in Annex I. However, a recruiting system based on a simple statistical approach would not pose the same risks as a system based on supervised learning. Third, they should consider defining capabilities, as doing so could further reduce over-inclusiveness and increase precision.

---

[113] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC, Art. 2(3).

[114] See OECD, 'Framework for the Classification of AI Systems' (2022) https://doi.org/10.1787/cb6d9eca-en, 67.

[115] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC.

[116] The third element—the ability to generate outputs that influence environments—seems to not play any meaningful role.

[117] As implied by Bob Carpenter, 'EU Proposing to Regulate the Use of Bayesian Estimation' (*Statistical Modeling, Causal Inference, and Social Science*, 22 April 2021) https://perma.cc/6FTJ-9FZB.

4. Can this approach also be applied to AGI regulations?

Future AI systems that achieve or exceed human performance in a wide range of cognitive tasks have been referred to as 'artificial general intelligence (AGI)'.[118] Even though the prospect of AGI is speculative, and some people remain sceptical,[119] a number of surveys show that many AI researchers do take it seriously.[120]

While the development of AGI could be overwhelmingly beneficial for humanity, it could also pose significant risks. Potential risks from AGI have been studied, among others, by Nick Bostrom,[121] Allan Dafoe,[122] Stuart Russell,[123] and Toby Ord.[124] There are also a number of public figures, such as Stephen Hawking,[125] Elon Musk,[126] and Bill Gates,[127] who have warned against the dangers of AGI. Against this background, it is not surprising that policy makers have started taking AGI more seriously. For example, the UK National AI Strategy contains the following passage:

> The [UK] government takes the long-term risk of non-aligned Artificial General Intelligence, and the unforeseeable changes that it would mean for the UK and the world, seriously.[128]

---

[118] Ben Goertzel and Cassio Pennachin, *Artificial General Intelligence* (Springer 2007), https://doi.org/10.1007/978-3-540-68677-4.

[119] Oren Etzioni, 'No, the Experts Don't Think Superintelligent AI is a Threat to Humanity' (*MIT Technology Review*, 20 September 2016) https://perma.cc/ZME3-UGXW; Melanie Mitchell, 'Why AI is Harder than We Think' (2021) https://arxiv.org/abs/2104.12871.

[120] Seth D Baum, Ben Goertzel and Ted G Goertzel, 'How Long Until Human-Level AI? Results from an Expert Assessment' (2011) 78 *Technological Forecasting and Social Change* 185, https://doi.org/10.1016/j.techfore.2010.09.006; Vincent C Müller and Nick Bostrom, 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion' in Vincent C Müller (ed), *Fundamental Issues of Artificial Intelligence* (Springer 2016), https://doi.org/10.1007/978-3-319-26485-1_33; Katja Grace and others, 'When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 *Journal of Artificial Intelligence Research* 729, https://doi.org/10.1613/jair.1.11222.

[121] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014).

[122] Allan Dafoe, 'AI Governance: A Research Agenda' (*Centre for the Governance of AI*, 27 August 2018) https://perma.cc/SA6T-F6XW.

[123] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Random House 2019).

[124] Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books 2020).

[125] Rory Cellan-Jones, 'Stephen Hawking Warns Artificial Intelligence Could End Mankind' (*BBC*, 2 December 2014) https://perma.cc/GA8A-BT5M.

[126] Samuel Gibbs, 'Elon Musk: Artificial Intelligence is our Biggest Existential Threat' (*The Guardian*, 27 October 2014) https://perma.cc/8WLM-LN4G.

[127] Kevin Rawlinson, 'Microsoft's Bill Gates Insists AI is a Threat' (*BBC*, 29 January 2015) https://perma.cc/N6Y8-CG2S.

[128] UK Government, 'National AI Strategy' (2021) https://perma.cc/RYN4-EEBR, 60.

If and when it becomes evident that AGI is in fact possible, policy makers may want to reduce the associated risks via regulation. This would again raise the question of how they should define the material scope of such AGI regulations. Would a risk-based approach be applicable to define all sorts of AI, including AGI?

It seems very likely that the *technical approach* that is used to build AGI will significantly influence its risks and potential risk mitigation strategies. For example, if AGI is developed using reinforcement learning,[129] we might use an approach called 'reward modelling' to align it to human values.[130] One might therefore be tempted to rely on technical approaches when defining the material scope of AGI regulations. However, there is an ongoing debate about whether today's technical approaches are sufficient to build AGI. While some AI researchers think this is reasonable,[131] others remain sceptical.[132] Given the high degree of uncertainty, policy makers should probably not rely exclusively on specific technical approaches.

Since AGI is characterised by the generality of its intelligence, it seems less fruitful to define specific *applications*. However, one could nonetheless distinguish between different types of AGI, such as question-answering, command-executing, or non-goal-directed systems.[133] Since these types could influence the feasibility and desirability of different safety precautions,[134] policy makers may want to use them to define the material scope of AGI regulations.

As mentioned above, the decisive *capability* of AGI is the generality of its intelligence. If a system exceeds human intelligence across the board, humanity would become the second most intelligent species on Earth[135] and might permanently lose its influence over the future.[136] However, I doubt that there is a definition of this capability that meets the requirements for legal definitions, mainly because I expect it to be highly vague. Instead, policy makers

---

[129] David Silver and others, 'Reward is Enough' (2021) 299 *Artificial Intelligence*, https://doi.org/10.1016/j.artint.2021.103535.

[130] Jan Leike and others, 'Scalable Agent Alignment via Reward Modeling: A Research Direction' (2018) https://arxiv.org/abs/1811.07871.

[131] Paul Christiano, 'Prosaic AI Alignment' (*AI Alignment*, 19 November 2016) https://perma.cc/43ED-M735.

[132] Daniel Kokotajlo, 'A Dilemma for Prosaic AI Alignment' (*AI Alignment Forum*, 17 December 2019) https://perma.cc/SG22-SCVW.

[133] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014), 177-193.

[134] Ibid, 191-192.

[135] Richard Ngo, 'AGI Safety from First Principles' (*AI Alignment Forum*, 28 September 2020) https://perma.cc/8JEE-ZDH8.

[136] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014); Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books 2020).

may want to define capabilities that could lead to the development of AGI.[137] These capabilities seem easier to define, but would still capture relevant AGI risks. One such capability could be the ability to recursively self-improve:[138]

> 'Recursive self-improvement' means an agent's ability to iteratively improve its own performance.

In summary, it seems plausible that policy makers could follow a risk-based approach to define the material scope of AGI regulations, though the focus might shift from technical approaches to capabilities.

A final note on foreseeability: It seems highly unlikely that we are able to foresee all future risks from AI. It is therefore necessary to make the scope definition scalable and adaptable. One possible approach could be to make the substance of the scope definition more future-proof. For example, one could include a catch-all definition of AI risk ('any other instance of significant risk caused or exacerbated by…'), which regulators and courts could then use to fill future regulatory gaps. Another approach would be to make it easier to update the scope definition as we identify new risks (e.g. via sunset clauses or built-in revision schedules). One could also combine more general definitions at the legislative level with more specific definitions at the sub-legislative level—which is very similar to what the proposed AI Act does.[139] Regardless of the particular approach policy makers choose, they need to closely monitor the AI landscape[140] and pay close attention to early warning signs.[141]

5. Conclusion

In this paper, I have shown that existing definitions of AI do not meet the most important requirements for legal definitions. Therefore, policy makers should not rely substantially on the term AI to define the material scope of AI. I have also shown that definitions of the main sources of relevant risks—certain technical approaches, applications, and capabilities—meet more of the

---

[137] More specifically, the scope definition could focus on the *intention* to develop certain capabilities (e.g. 'any serious and promising attempt').

[138] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014), 409; Irving J Good, 'Speculations Concerning the First Ultraintelligent Machine' (1966) 6 *Advances in Computers* 31, https://doi.org/10.1016/S0065-2458(08)60418-0, 33.

[139] Pursuant to Art. 4, the European Commission can adopt delegated acts to amend the list of techniques and approaches listed in Annex I.

[140] Jess Whittlestone and Jack Clark, 'Why and How Governments Should Monitor AI Development' (2021) https://arxiv.org/abs/2108.12427.

[141] Carla Z Cremer and Jess Whittlestone, 'Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI' (2020) 6 *International Journal of Interactive Multimedia and Artificial Intelligence* 100, http://doi.org/10.9781/ijimai.2021.02.011.

requirements for legal definitions than definitions of the term AI. Finally, I have argued that this approach can, in principle, also be used to define the material scope of AGI regulations.

The paper has made four main contributions. First, it has provided a comprehensive legal argument for why policy makers should not rely on the term AI for regulatory purposes and why a risk-based approach would be preferable. Second, it has proposed a list of specific requirements for legal definitions which can also be used to evaluate other definitions. Third, the paper has suggested a new categorization of the main sources of AI risks that policy makers may want to address. And fourth, it can be seen as a first step towards the study of AGI regulation, which I expect will turn into its own field of interest for policy makers and researchers in the future.

The findings of this paper are relevant for policy makers worldwide. They support the European Commission's risk-based approach.[142] The suggested definitions of certain technical approaches, applications, and capabilities can also be used to amend or substantiate the list of techniques and approaches in Annex I and high-risk applications in Annex III. But I expect the findings to be even more relevant for policy makers who have not yet drafted concrete proposals. Defining the material scope of AI regulations requires careful consideration. I hope this paper comes at the right time to help policy makers rise to this challenge.

Acknowledgements

I am grateful for valuable comments and feedback from Seth Baum, Conor Griffin, Renan Araújo, Leonie Koessler, Nick Hollman, Suzanne Van Arsdale, Markus Anderljung, Matthijs Maas, and Sébastien Krier. I also thank the participants of a seminar hosted by the Legal Priorities Project in February 2021. All remaining errors are my own.

[142] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' COM (2021) 206 final, https://perma.cc/V2RH-6KGC.

Legal Priorities Research: A Research Agenda

Christoph Winter[1,2,3], Jonas Schuett[1,4], Eric Martínez[1,5], Suzanne Van Arsdale[1], Renan Araújo[1], Nick Hollman[1], Jeff Sebo[6], Andrew Stawasz[3], Cullen O'Keefe[1,7,8], Giuliana Rotola[9,10]

[1] *Legal Priorities Project*
[2] *Instituto Tecnológico Autónomo de México (ITAM)*
[3] *Harvard University*
[4] *Goethe University Frankfurt*
[5] *Massachusetts Institute of Technology (MIT)*
[6] *New York University (NYU)*
[7] *OpenAI*
[8] *University of Oxford*
[9] *Open Lunar Foundation*
[10] *Foresight Institute*

January 2021

We are particularly grateful for valuable comments, feedback and support from Alfredo Parra, Mark Eccleston-Turner, Leonie Koessler, Sean Richardson, Tyler John, Olavo Bittencourt, Lisa Forsberg, Kevin Tobia, Bradly Condon, Markus Anderljung, Seth Baum, Tobias Baumann, Haydn Belfield, Alexis Carlier, Devin Mauney, Peter Cihon, Frans von der Dunk, Dov Greenbaum, Ameen Jauhar, David Koplow, Sebastien Krier, David Manheim, Tanja Masson-Zwaan, Cecil Abungu, Gabriel Bankman-Fried, Gregory Lewis, Janvi Ahuja, Sarah Carter, Alasdair Phillips-Robins, Caleb Huffman, and the participants at presentations of the individual parts of this agenda. All views and errors are our own.

Introduction

Humanity has seen relatively stable improvements in quality of life over time. Present generations benefit from the accomplishments of past generations, and future generations benefit from advanced knowledge, economic growth, stronger institutions, and other improved conditions for welfare created by present generations. This trend, however, might change.

Our ever-advancing knowledge, based on the exchange of ideas throughout space and time, has led to technologies that threaten the very existence of future generations. Yet, while humanity has been aware of the first anthropogenic existential threat for some time (the use of nuclear weapons) and is slowly realizing the dangers of climate change, the ongoing COVID-19 pandemic has shown that we are not prepared for some of the greatest threats of this century. For example, although scientific knowledge allowed us to encode the genome of the novel coronavirus within days, and an effective vaccine was discovered shortly thereafter, most national and international institutions have not been able to challenge the spread of the virus effectively.

More deadly and contagious pandemics, natural or engineered, may well pose much greater, possibly even existential threats to the future of humanity. Whether we address these and other risks—such as those resulting from advanced artificial intelligence, runaway climate change, or synthetic biology—will drastically affect the well-being of future generations, so much so that we may be at a very unusual point in history: For the first time, the future of sentient life heavily depends on those in the present. Even more so, its very existence may be at stake during what has been referred to as "the precipice" (Ord, 2020). Although our actions (and inactions) may have historically unique consequences for future generations, their interests are not represented in current political and economic systems, and human intuitions have not yet been updated accordingly. This calls for fundamental legal change.

Given that some of the risks and opportunities to positively shape the lives of countless future individuals are much greater than others, prioritization is of utmost importance. What are the greatest risks and opportunities for humanity, and what is the role that multidisciplinary-informed legal research can take? How can we prioritize so as to increase the chance of a flourishing and long-lasting future of

humanity? How can we cooperate most effectively with those whom we will never meet, but whose lives lie in our hands? Choosing to address these questions and prioritizing carefully among them may be one of the great opportunities of our time to positively change the human trajectory, and will be the guiding theme of this agenda.

Part 1 outlines the various empirical and philosophical foundations underlying both our research agenda and legal priorities research generally. In particular, we highlight prioritization efforts as an important and neglected tool for legal scholarship (Section 1) and emphasize the importance of taking into account the long-term consequences of laws and legal research during prioritization (Section 2). Finally, we offer a rigorous yet flexible, and potentially ever-evolving methodological framework for deciding which problems to work on and how to tackle them (Section 3).

In Part 2 of this agenda, we explore a number of specific cause areas in more detail and identify promising research projects within each. We recognize that many of these projects are relatively broad, and further work is often needed to articulate a more specific research question that would naturally correspond to an individual research paper. We also provide an overview of relevant literature at the end of each individual subsection. This Part covers the law and governance of artificial intelligence (Section 4), synthetic biology and biorisk (Section 5), and institutional design (Section 6). Since choosing the right research project is one of the most important factors that determines the impact of legal research, we have also identified a number of meta-research projects (Section 7). Research in this area tackles problems that legal researchers encounter when prioritizing, such as whether to focus on international, comparative, or national law.

Part 3 follows the structure of Part 2. Here, we outline further cause areas that also fit our methodology criteria but for which further research is needed to more precisely compare them with other cause areas. This Part covers space governance (Section 8) and animal law (Section 9). Though we refer to these as cause areas for further engagement, we encourage interested researchers to pursue projects in these fields, both at the meta- and object-level, and may integrate them into our main cause areas in future iterations of this agenda.

Legal priorities research is by its very nature an interdisciplinary affair. We therefore include an appendix which aims to give an overview of some of the most closely related areas of existing literature that are likely to be particularly useful for legal priorities research. This appendix is organized around the general academic disciplines of philosophy (A), economics (B), psychology (C), macrohistory (D), and political science (E). Within each discipline we identify both general examples of interdisciplinary research between law and that respective discipline, as well as more specific research areas within those disciplines.

Identifying the most important research projects is accompanied by high degrees of both normative and empirical uncertainty. Although we develop specific criteria in Section 3 to account for this, a substantial amount of holistic uncertainty remains and must be acknowledged. This leads us even more so to appreciate feedback from the wider community of legal scholars who are interested in prioritization, law, and the long-term future. In fact, it would not have been possible to write this agenda without the helpful feedback and comments from and conversations with various experts in the first place. The transparency of this agenda's philosophical and empirical assumptions in its first Section is very much motivated by the idea of continuing and encouraging a fruitful culture of feedback. This said, the agenda is a common project in a different way as well: We aim at inspiring and encouraging the legal community to take up the outlined challenges. Anyone interested in using the agenda to get ideas and guidance on potential projects should feel free to do so.

Part 2

Exploration by Cause Areas

In this part, we explore a number of cause areas in more detail. This includes the law and governance of artificial intelligence (Section 4), synthetic biology and biorisk (Section 5), and institutional design (Section 6). Since choosing the right research project is one of the most important factors that determines the impact of legal research, we are also engaging in a number of meta-research projects (Section 7). Instead of competing with the existing organizations, our research in this area is significantly more specific in that it exclusively tackles problems that legal researchers encounter when prioritizing, such as whether to focus on international, comparative, or national law.

4 Artificial Intelligence

Artificial intelligence (AI)[80] could significantly shape the long-term future. On the one hand, it could enable scientific breakthroughs[81] and the accumulation of unprecedented wealth. On the other hand, it could pose existential risks and cause

---

[80] There is no generally accepted definition of the term "artificial intelligence." Since its first usage by McCarthy et al. (1955), a vast spectrum of definitions has emerged. Popular definitions have been proposed, among others, by Kurzweil et al. (1990), McCarthy (2007), Minsky (1969), Nilsson (2009), and Russell and Norvig (2020). For surveys of AI definitions, see Legg and Hutter (2007a, 2007b) and Monett and Lewis (2018). Recently, policy makers have started to develop their own definitions (European Commission, 2018; Federal Government of Germany, 2019; High-Level Expert Group on AI, 2019; Organisation for Economic Co-operation and Development [OECD], 2019; Office for AI, 2020). For more information about the term "AI" in the legal context, see Martinez (2019), Scherer (2016), Schuett (2019), and Turner (2019). More advanced AI systems have been referred to as "Transformative AI, TAI" (Dafoe, 2018; Gruetzemacher et al., 2019; Gruetzemacher & Whittlestone, 2019; Karnofsky, 2016b), "Artificial General Intelligence, AGI" (Goertzel & Pennachin, 2007; Muehlhauser, 2013), and "Superintelligence" (Bostrom, 1998, 2003b, 2014).

[81] See, for instance, DeepMind's latest progress in solving the "protein folding problem" (Jumper et al., 2020).

suffering on an astronomical scale. There seems to be a general consensus in prioritization research that positively shaping the development of AI is one of the world's most pressing problems (Gloor, 2016b; Karnofsky, 2016a; Wiblin, 2017). Even though the law seems to play an important role in this respect, there is surprisingly little legal research focused on the long-term implications of AI.[82] We have identified four areas of research which seem particularly promising: reducing existential risks from AI (Section 4.1), reducing suffering risks from AI (Section 4.2), sharing the benefits of AI (Section 4.3), and meta-research in AI (Section 4.4).

4.1 Reducing Existential Risks from AI

It has been argued that AI could pose existential risks for humanity (Bostrom, 2014; Christian, 2020; Ord, 2020; Russell, 2019).[83] Ord (2020) estimates that there is a 10% chance that AI will cause an existential catastrophe within the next 100 years. Similarly, Wiblin (2017) estimates that the risk of a serious catastrophe caused by machine intelligence within the next 100 years is between 1% and 10%. A recent survey of leading AI safety and governance researchers reveals similar estimates (Carlier et al., 2020).[84] Risks from AI have been conceptualized as (a) accident risks, (b) misuse risks, and (c) structural risks (Zwetsloot & Dafoe, 2019).[85] The following research projects detail promising mechanisms through which the law could help to reduce each of these risks.[86]

---

82 Notable exceptions include Flynn (2020), Liu et al. (2018), Maas (2019a, 2019b), O'Keefe (2018, 2020a, 2020b), and O'Keefe et al. (2020).

83 Recall that "existential risks" are risks where an adverse outcome would either annihilate Earth-originating intelligent life, or permanently and drastically curtail its potential (Bostrom, 2002). For more information on existential risks, see Section 3.2.1.

84 Note that subjective probability estimates of existential catastrophes should be taken with a grain of salt (Baum, 2020b; Beard et al., 2020a; see also Morgan, 2014). We therefore advise against putting too much emphasis on the precise numbers. However, the estimates do suggest that leading experts think that the probability is sufficiently high to take the risks seriously.

85 It is worth noting that accident and misuse risks are dichotomous (unintentional vs. intentional harm), whereas structural risks can overlap with both accident and misuse risks.

86 For a more general analysis of potential responses to extinction risks, see Cotton-Barratt et al. (2020).

RESEARCH PROJECTS

4.1.1 Reducing Accident Risks

"AI accidents" can be defined as any unintended and harmful behavior of an AI system (Amodei et al., 2016).[87] Specific scenarios in which AI accidents cause an existential catastrophe have been described by Bostrom (2014) and Yudkowsky (2008a),[88] as well as Christiano (2019).[89] A major challenge in all scenarios is to ensure that advanced AI systems are properly aligned with human values (Bostrom, 2014; Christian, 2020; Christiano, 2018b; Gabriel, 2020; Russell, 2019; Soares, 2016a; Soares & Fallenstein, 2014; Taylor et al., 2016). This problem, which is typically called the "alignment problem," involves a technical and a normative challenge (Gabriel, 2020).

The technical challenge is how to encode values in a given AI system so that it reliably does what it ought to do.[90] Proposed solutions include "iterated amplification" (Christiano, 2018; Cotra, 2018) and "debate" (Irving et al., 2018), though the problem ultimately remains unsolved. The law can help to ensure that these or

---

87 Accident risks can be further broken down into (a) specification problems, (b) robustness problems, and (c) assurance problems (Ortega & Maini, 2018). Specification ensures that an AI system's behavior aligns with the operator's true intentions. For more information on specification problems, see Clark and Amodei (2016), Everitt et al. (2019), Krakovna et al. (2019a, 2019b), and Leike et al. (2018). Robustness ensures that an AI system continues to operate within safe limits upon encountering perturbations. For more information on robustness problems, see García and Fernández (2015), Goodfellow et al. (2015), Kohli et al. (2019), Quiñonero-Candela et al. (2009), and Szegedy et al. (2014). Assurance ensures that we can understand and control AI systems during operations. For more information on assurance problems, see Orseau and Armstrong (2016) and Doshi-Velez and Kim (2017).

88 In this scenario a single AI system with goals that are hostile to humanity quickly becomes sufficiently capable of complete world domination and causes the future to contain very little of what we value. The scenario has been criticized, among others, by Baum (2018b), Baum et al. (2017), Calo (2017), Christiano (2018a), Davis and Marcus (2019), Drexler (2019), Goertzel (2015), and Shah (2018). For reviews of *Superintelligence* in academic journals, see Brundage (2015), Thorn (2015), and Thomas (2016). For informal discussion, see Fodor (2018) and Grace (2014).

89 This scenario, which Christano refers to as "part 2," involves multiple AIs accidentally being trained to seek influence, and then failing catastrophically once they are sufficiently capable, causing humans to become extinct or otherwise permanently lose all influence over the future. For informal discussion, see Hubinger et al. (2019) and in parts Carlier and Davidson (2020). See Manheim (2019) on the dynamics that make the multi-agent scenario more complex and difficult to understand even in the short run.

90 Bostrom (2014, p. 185) calls this the "value loading problem."

other solutions are actually implemented or slow down the development before certain safety standards are met. For example, there could be corresponding AI safety regulations.[91] How should such regulations be formed? Will EU regulation diffuse globally via the so-called "Brussels effect" (Bradford, 2020), or will there be a global race to the bottom with regards to minimum safety standards (Askell et al., 2019; Smuha, 2019)? Is there a need for new regulatory bodies (Calo, 2014; Erdélyi & Goldsmith, 2018; Scherer, 2016)? How should the scope of AI safety regulations be defined (Schuett, 2019)? Do we need new regulatory instruments (Clark & Hadfield, 2018)? How can compliance be monitored and enforced? Is there a need for stronger forms of supervision (Bostrom, 2019; Garfinkel, 2018)? If so, would they violate civil rights and liberties? What is the relationship between hard and soft law (Villasenor, 2020)? In particular, what role should professional self-regulation (O'Keefe, 2020a) and other forms of soft-law play (Cihon, 2019; Cihon et al., 2020; Jobin et al., 2019)? Do existing criminal law provisions penalize the (concrete or abstract) increase of existential accident risks (e.g., Section 221 of the German Criminal Code)? How effective are liability regimes to tackle existential accident risks? Which other legal mechanisms are conceivable (e.g., Farquhar et al., 2017)?

The normative challenge is what values, if any, we ought to encode in a given AI system. A possible answer to this question is to use some aggregate of the ethical views of society (Baum, 2017).[92] How can legal research contribute to the related challenges, such as whose ethical views to include, how to identify their views, and how to combine individual views to a single view? What can we learn from techniques to balance conflicting legal interests, such as the principles of "proportionality" or "balancing" respectively? To what extent can the law itself be used as a proxy for desirable values?

---

91 See the discussion around the "White Paper on AI" (European Commission, 2020) in the EU, for example, Abecassis et al. (2020), Belfield et al. (2020), Centre for the Governance of AI (2020), and Future of Life Institute (2020), as well as the "Ethics Guidelines for Trustworthy AI" (High-Level Expert Group on AI, 2019), for example, Avin and Belfield (2019). Also see the responses to planned government regulation in the UK, for example, Beard et al. (2017), Belfield and Ó hÉigeartaigh (2017), Belfield et al. (2020), and Cave (2017).

92 More precisely, one could seek to have the AI derive its values from the values of other ethical agents. This mechanism has been called "coherent extrapolated volition" (Bostrom, 2014; Muehlhauser & Helm, 2012; Yudkowsky, 2004). Alternatively, one could follow a "bottom-up" approach, i.e., AI designed to learn ethics as it interacts with its environment and with other ethical agents (Allen et al., 2000; Allen et al., 2005; Wallach & Allen, 2008; Wallach et al., 2008).

4.1.2 Reducing Misuse Risks

"AI misuse" means any use of an AI system with the intention of causing harm (Brundage et al., 2018).[93] A possible risk scenario involves a malevolent actor (for example, a terrorist organization or rogue state) who gains control over powerful AI-based weapons (for example, lethal autonomous weapons). How can the law be used to reduce existential risks in this scenario? In particular, what role should criminal law and law enforcement play? Is there a need to legally restrict certain types of scientific knowledge to prevent malevolent actors from gaining control over potentially dangerous AI technologies (Bostrom, 2017; Ovadya & Whittlestone, 2019; Shevlane & Dafoe, 2020; Whittlestone & Ovadya, 2020)? If so, how could this be done most effectively? To what extent is restricting scientific knowledge consistent with the relevant provisions of constitutional law?

Another misuse scenario involves an authoritarian government that uses AI-based surveillance techniques to permanently suppress opposition (Caplan, 2011; Garfinkel, 2018; Ord, 2020; Young et al., 2019). For such Orwellian surveillance states, the term "digital authoritarianism" has been coined. If they lock in the conditions for welfare on an extremely low level, they could constitute an existential risk (see Section 3.2.1). How can the law prevent the emergence of such regimes? Should certain surveillance techniques be banned?[94] Which limits does constitutional law place on the use of facial recognition technologies for state surveillance purposes (Ferguson, 2019)? Inversely, in which cases can stronger forms of state surveillance be justified in order to reduce other types of risk (Bostrom, 2019; Garfinkel, 2018)? How should international law respond to such threats?

The judicial system will likely play an important role in a digital authoritarian state. With the development of advanced artificial judicial intelligence (Winter, 2021a), values, laws, and other norms could be implemented into a primarily AI-based judiciary that becomes resistant to change. This type of lock-in effect has been called "technological-legal lock-in" (Crootof, 2019) and has been argued to result from current limitations of AI systems to adapt to social changes and institutional factors such as path dependence (Bernstein, 2006; Crootof, 2019; Re & Solow-Niederman, 2019). How does this conception of technological-legal lock-in scale with advancements in AI capabilities and potential solutions to the alignment problem, in particular to the normative challenge (Gabriel, 2020)? What other

---

93 Brundage et al. (2018) prefer the term "malicious use," but there seems to be no difference. For more information on misuse risks, see Belfield (2019), Dafoe (2018), and Karnofsky (2016a).

94 In the US, some municipalities have already started to ban state use of facial recognition technology for law enforcement purposes, including San Francisco (Conger et al., 2019) and Boston (Johnson, 2020).

institutional factors contribute to technological-legal lock-in? Which challenges would artificial judicial decision-making pose for liberal democracy (Winter, 2021a)? How can we uphold liberal democratic values in general and the separation of powers in particular within an AI judiciary? How should these long-term risks be balanced with potential short-term benefits, such as improved access to justice, transparency and fairness (Winter, 2020a; Winter, 2021a)? Which other long-term effects from AI in the judiciary are conceivable (Hollman et al., 2021)?

4.1.3 Reducing Structural Risks

AI could also shape the broader environment in harmful ways that do not fall into the accident-misuse dichotomy. These risks have been called "structural risks" (Zwetsloot & Dafoe, 2019). They typically result from the destabilizing effects of AI and could also be seen as risk factors (see Section 3.2.1).[95] A possible scenario involves some kind of war exacerbated by developments in AI (Aguirre, 2020; Allen & Chan, 2017; Avin & Amadae, 2019; Boulanin et al., 2020; Dafoe, 2018; Geist & Lohn, 2018; Horowitz, 2019; Horowitz et al., 2019; Jayanti & Avin, 2020; Lieber & Press, 2017; Maas, 2019a, 2019b).[96] For example, if AI systems could be used to detect retaliation capabilities, the equilibrium of mutual assured destruction would be disturbed, which would drastically increase the risk of a nuclear war (Bostrom, 2019; Horowitz, 2019; Lieber & Press, 2017). How effective are international treaties at banning certain AI applications (Castel & Castel, 2016; Maas, 2019a; Nindler, 2019; Wilson, 2013)? Can operators of lethal autonomous weapons be held criminally responsible (Bo, 2020)? How else can the law be used to reduce structural risks in a war scenario?

Race dynamics are another destabilizing factor (Armstrong et al., 2016; Askell et al., 2019; Bostrom, 2017; Hogarth, 2018; Naudé & Dimitri, 2020; Soares, 2016b). If competing actors think that AI could lead to some kind of economic, military or technological supremacy, and gains from AI result from their relative strength over other actors, then a race dynamic will commence in which actors might be willing to sacrifice safety in order to "win the race" (Askell et al., 2019). Such a dynamic could increase the risk that advanced AI systems are unaligned, thereby increasing the risk of an existential accident. How can the law reduce such race dynamics?

---

95 For more information on AI risk factors, see Hernández-Orallo et al. (2019) and Burden & Hernández-Orallo (2020).

96 Another scenario has been described in Part 1 of *What failure looks like* (Christiano, 2019). This scenario involves multiple AIs pursuing easy-to-measure goals, rather than the goals humans actually care about, causing us to permanently lose some influence over the future. For informal discussion, see Clarke (2020), Grue_Slinky (2019), Hanson (2019), and Pace (2020).

Which legal mechanisms can help to increase trust among competing actors (Brundage et al., 2020)? For example, there could be regulations intended to prevent a race to the bottom with regards to minimum safety standards (Smuha, 2019). There could also be auditing and certifications schemes (Cihon et al., 2020), or contractual obligations to develop AI responsibly (Askell et al., 2019). What are the most effective means to reduce structural risk?

As governments realize the power of AGI, they may seek to gain control over its development and deployment, leading to a new kind of geopolitics which has been referred to as "AI nationalism" (Hogarth, 2018). Increasing economic and political tensions between states like the US and China could then increase other types of risks, such as the risk of great power wars. How can the law reduce such tensions and foster cooperation between states? How effective are economic treaties at preventing related protectionist trade policies? How can the law help to make AI a global public good (Hogarth, 2018)? Does this require a new global organization (Cihon et al., 2020a, 2020b; Erdélyi & Goldsmith, 2018; Kemp et al., 2019)? It is worth noting that private actors currently dominate AI research and development, which leads to the question of who should govern the development of advanced AI systems (Leung, 2019). When is governmental control desirable (Leung, 2018) and what form should it take? To the extent that government control or influence is undesirable, which modes of influence (O'Keefe, 2020b) and possible defensive measures exist? Under what circumstances would it be preferable if governments were unaware of the development of advanced AI systems?

EXISTING ACADEMIC LITERATURE

Allen, G., & Chan, T. (2017, July). *Artificial intelligence and national security*. Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. https://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI and Society*, *31*, 201–206. https://doi.org/10.1007/s00146-015-0590-y

Askell, A., Brundage, M., & Hadfield, G. (2019). *The role of cooperation in responsible AI development*. arXiv. https://arxiv.org/abs/1907.04534

Avin, S., & Amadae, S. M. (2019). Autonomy and machine learning as risk factors at the interface of nuclear weapons, computers and people. In V. Boulanin, (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk* (pp. 105–118). Stockholm International Peace Research Institute. https://doi.org/10.17863/CAM.44758

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI and Society*, *35*, 165–176. https://doi.org/10.1007/s00146-017-0760-1

Bernstein, G. (2006). When new technologies are still new: Windows of opportunity for privacy protection. *Villanova Law Review*, *51*(4), 921–950. https://digitalcommons.law.villanova.edu/vlr/vol51/iss4/8

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135–148. https://doi.org/10.1111/1758-5899.12403

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455–476. https://doi.org/10.1111/1758-5899.12718

Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Carlsson, M. P. (2020). *Artificial intelligence, strategic stability and nuclear risk*. Stockholm International Peace Research Institute. https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). *Malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. https://arxiv.org/abs/1802.07228

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensbold, J., O'Keefe, C., Koren, . . . Anderljung, M. (2020). *Toward trustworthy AI development: Mechanisms for supporting verifiable claims*. arXiv. https://arxiv.org/abs/2004.07213

Calo, R. (2014, September 15). *The case for a federal robotics commission*. Brookings Institution. https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission

Caplan, B. (201). The totalitarian threat. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 504–519). Oxford University Press.

Carlier, A., Clarke, S., & Schuett, J. (2020). *AI risk survey* [Unpublished manuscript].

Castel, J.-G., & Castel, M. E. (2016). The road to artificial superintelligence: Has international law a role to play? *Canadian Journal of Law and Technology*, *14*(1), 1–15. https://ojs.library.dal.ca/CJLT/article/view/7211

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.

Christiano, P., Shlegeris, B., & Amodei, D. (2018). *Supervising strong learners by amplifying weak experts*. arXiv. https://arxiv.org/abs/1810.08575

Cihon, P. (2019). *Standards for AI governance: International standards to enable global coordination in AI research & development* [Technical report]. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Cihon, P., Maas, M. M., & Kemp, L. (2020a). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy*, *11*(5). https://doi.org/10.1111/1758-5899.12890

Cihon, P., Maas, M. M., & Kemp, L. (2020b). Should artificial intelligence governance be centralised? Design lessons from history. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–234. https://doi.org/10.1145/3375627.3375857

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2020). *AI certification: Advancing ethical practice by reducing information asymmetries* [Manuscript submitted for publication].

Clarke, J., & Hadfield, G. K. (2018). *Regulatory markets for AI safety*. arXiv. https://arxiv.org/abs/2001.00078

Crootof, R. (2019). "Cyborg Justice" and the risk of technological-legal lock-in. *Columbia Law Review Forum*, *119*(7), 233–251. https://columbialawreview.org/content/cyborg-justice-and-the-risk-of-technological-legal-lock-in

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf

Erdélyi, O. J., & Goldsmith, J. A. (2018). Regulating artificial intelligence: Proposal for a global solution. *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. https://doi.org/10.1145/3278721.3278731

Farquhar, S., Cotton-Barratt, O., & Snyder-Beattie, A. (2019). Pricing externalities to balance public risks and benefits of research. *Health Security*, *15*(4), 401–408. https://doi.org/10.1089/hs.2016.0118

Ferguson, A. G. (2019). Facial recognition and the fourth amendment. *Minnesota Law Review*, *105*. http://dx.doi.org/10.2139/ssrn.3473423

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*. https://doi.org/10.1007/s11023-020-09539-2

Geist, E., & Lohn, A. J. (2018). *How might artificial intelligence affect the risk of nuclear war?* RAND Corporation. https://www.rand.org/pubs/perspectives/PE296.html

Hollman, N., Winter, C. K., & Jauhar, A. (2021). *Long-term challenges of AI for the judiciary* [Unpublished manuscript].

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Horowitz, M. C., Scharre, P., & Velez-Green, A. (2019). *A stable nuclear future? The impact of autonomous systems and artificial intelligence*. arXiv. https://arxiv.org/abs/1912.05291

Irving, G., Christiano, P., & Amodei, D. (2018). *AI safety via debate*. arXiv. https://arxiv.org/abs/1805.00899

Jayanti, A., & Avin, S. (2020). *It takes a village: The shared responsibility of "raising" an autonomous weapon*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/it-takes-village

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kemp, L., Cihon, P., Maas, M. M., Belfield, H. Ó hÉigeartaigh, S., Leung, J., & Cremer, C. Z. (2019). *UN high-level panel on digital cooperation: A proposal for international AI governance*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/proposal-international-ai-governance

Leung, J. (2019). *Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies* [Doctoral dissertation]. University of Oxford. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665

Liu, H.-Y., Lauta, K. C., & Maas, M. M. (2018). Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, *102*, 16–19. https://doi.org/10.1016/j.futures.2018.04.009

Maas, M. M. (2019a). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, *40*(3), 285–311. https://doi.org/10.1080/13523260.2019.1576464

Maas, M. M. (2019b). Innovation-proof global governance for military artificial intelligence? *Journal of International Humanitarian Legal Studies*, *10*(1), 129–157. https://doi.org/10.1163/18781527-01001006

Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy. *AI and Society*, *35*, 367–379. https://doi.org/10.1007/s00146-019-00887-x

Nindler, R. (2019). The United Nation's capability to manage existential risks with a focus on artificial intelligence. *International Community Law Review*, *21*(1), 5–34.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Ovadya, A., & Whittlestone, J. (2019). *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*. arXiv. https://arxiv.org/abs/1907.11274

O'Keefe, C. (2020a). *Antitrust-compliant AI industry self-regulation* [Unpublished manuscript]. https://cullenokeefe.com/blog/antitrust-compliant-ai-industry-self-regulation

O'Keefe, C. (2020b). *How will national security considerations affect antitrust decisions in AI? An examination of historical precedents* [Technical report]. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/How-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-Okeefe.pdf

Re, R. M., & Solow-Niederman, A. (2019). Developing artificially intelligent justice. *Stanford Technology Law Review*, *22*(2), 242–289.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin Random House.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law and Technology*, *29*(2), 353–400. http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf

Schuett, J. (2019). *A legal definition of AI*. arXiv. https://arxiv.org/abs/1909.01095

Shevlane, T., & Dafoe, A. (2020). The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 173–179. https://doi.org/10.1145/3375627.3375815

Smuha, N. A. (2019). *From a 'Race to AI' to a 'Race to AI Regulation': Regulatory competition for artificial intelligence*. SSRN. http://dx.doi.org/10.2139/ssrn.3501410

Soares, N. (2016a). *The value learning problem*. Machine Intelligence Research Institute. https://intelligence.org/files/ValueLearningProblem.pdf

Soares, N., & Fallenstein, B. (2014). Agent foundations for aligning machine intelligence with human interests: A technical research agenda. In V. Callaghan, J. Miller, R.

Yampolskiy, & S. Armstrong (Eds.), *The technological singularity: Managing the journey* (pp. 103–125). Springer. https://doi.org/10.1007/978-3-662-54033-6_5

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). *Alignment for advanced machine learning systems*. Machine Intelligence Research Institute. https://intelligence.org/files/AlignmentMachineLearning.pdf

Villasenor, J. (2020, July 31). *Soft law as a complement to AI regulation*. Brookings Institution. https://www.brookings.edu/research/soft-law-as-a-complement-to-ai-regulation

Whittlestone, J., & Ovadya, A. (2020). *The tension between openness and prudence in responsible AI research*. arXiv. https://arxiv.org/abs/1910.01170

Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. *Virginia Environmental Law Journal*, *31*(2), 307–364. https://www.jstor.org/stable/44679544

Winter, C. K. (2020a). The value of behavioral economics for EU judicial decision-making. *German Law Journal*, *21*(2), 240–264. https://doi.org/10.1017/glj.2020.3

Winter, C. K. (2021a). Exploring the challenges of artificial judicial decision-making for liberal democracy [Forthcoming]. In P. Bystranowski, P. Janik, & M. Próchnicki (Eds.), *Judicial decision-making: Integrating empirical and theoretical perspectives*. https://www.christophwinter.net/s/AI-Judiciary.pdf

Young, M., Katell, M., & Krafft, P. M. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data and Society*, *6*(2), 1–14 https://doi.org/10.1177/2053951719868492

Yudkowsky, E. (2008a). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf

EXISTING INFORMAL DISCUSSION

Aguirre, A. (2020, November 11). *Why those who care about catastrophic and existential risk should care about autonomous weapons* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2

Bo, M. (2020, December 18). *Meaningful human control over autonomous weapon systems: An (international) criminal law account*. Opinio Juris. http://opiniojuris.org/2020/12/18/meaningful-human-control-over-autonomous-weapon-systems-an-international-criminal-law-account

Christiano, P. (2018, April 7). *Clarifying "AI alignment*". AI Alignment. https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6

Christiano, P. (2019, March 17). *What failure looks like [*Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Cotra, A. (2018, March 4). *Iterated distillation and amplification*. AI Alignment. https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616

Garfinkel, B. (2018, October 12). *The future of surveillance*. Effective Altruism. https://www.effectivealtruism.org/articles/ea-global-2018-the-future-of-surveillance

Hogarth, I. (2018, June 13). *AI nationalism*. https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism

Karnofsky, H. (2016a, May 6). *Potential risks from advanced artificial intelligence: The philanthropic opportunity*. Open Philanthropy. https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

Leung, J. (2018, September 28). *Analyzing AI actors*. Effective Altruism. https://www.effectivealtruism.org/articles/ea-global-2018-analyzing-ai-actors

Soares, N. (2016b, July 23). *Submission to the OSTP on AI outcomes*. Machine Intelligence Research Institute. https://intelligence.org/2016/07/23/ostp

Wiblin, R. (2017b, March). *Positively shaping the development of artificial intelligence*. 80,000 Hours. https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence

Zwetsloot R., & Dafoe, A. (2019, February 11). *Thinking about risks from AI: Accidents, misuse and structure*. Lawfare. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure

4.2 Reducing Suffering Risks from AI

AI could cause suffering on an astronomical scale (Althaus & Baumann, 2020; Althaus & Gloor, 2016; Baumann, 2017a, 2017b, 2018a, 2018b; Clifton, 2020; Daniel, 2017; Gloor, 2016b; Tomasik, 2018, 2019b).[97] If a state of astronomical suffering is permanently locked in, it could be worse than extinction, making such scenarios the worst kind of existential risks (Daniel, 2017). Against this backdrop, it is worrying that suffering risks (s-risks) from AI are highly neglected, especially in legal research. Given our high degree of uncertainty, disentanglement research seems particularly important (see Flynn, 2017). Besides that, we think that the following research directions are worth considering.

RESEARCH PROJECTS

4.2.1 Near Misses

A potential s-risk scenario involves an AGI which is only slightly misaligned with human values (Tomasik, 2018).[98] Such a scenario, which has been called "near miss," could cause astronomical amounts of suffering. For example, suppose an AGI has the goal of creating as many "happy minds" as possible, but its slightly askew interpretation of this goal results in vast numbers of minds with severe mental-

[97] Recall that "suffering risks" are risks where an adverse outcome would bring about suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far (Althaus & Gloor, 2016). For more information on suffering risks, see Section 3.2.1.

[98] For more information on the "alignment problem," see Section 4.1.

health problems like depression or anxiety. We have already outlined potential ways in which the law could help to solve the alignment problem in Section 4.1.

4.2.2 Mind Crime

S-risks could also result from situations in which artificial minds are made to suffer for instrumental purposes, for instance in order to simulate evolution or perform experiments (Tomasik, 2019b). This scenario has been called "mind crime" (Bostrom, 2014, p. 125). If one assumes that artificial minds have a relevant moral status (Danaher, 2019; Gunkel, 2018; Schwitzgebel & Garza, 2015; Shulman & Bostrom, 2020; Tomasik, 2014), and that there could be vast numbers of them, suffering can reach astronomical scales. What is the threshold above which artificial minds should be legally protected (Chesterman, 2020; Hubbard, 2011; Kurki, 2019)? How should the law deal with uncertainties about their moral status (cf. MacAskill et al., 2020)? What can we learn from the related debate on animal welfare (see Section 9)?

4.2.3 Agential S-Risks

"Agential s-risks" involve agents that actively and intentionally want to cause harm (Althaus & Baumann, 2020; Baumann, 2017b, 2018b).[99] It seems at least somewhat plausible that artificial agents might exhibit behavior that resembles malevolent traits like psychopathy or sadism.[100] Their occurrence in some humans suggests that they may have provided evolutionary fitness advantages (Book et al., 2015; Jonason et al., 2015; McDonald et al., 2012; Nell, 2006). If these traits prove useful in a given environment, then advanced AI systems that are trained on this environment might learn corresponding behavior with potentially catastrophic consequences. For example, an agent might cause suffering as a strategic threat in an escalating conflict (Baumann, 2018b). One possible intervention would be to expand the scope of extortion laws. To the extent that the agent making such threats is controlled by states, international treaties banning such strategies could be another lever. Besides that, it is unclear how the law could reduce such risks. There is a need for exploratory research that structures the problem and identifies relevant questions for legal research.

---

[99] This is the s-risk equivalent of existential misuse risks as outlined in Section 4.1.2.

[100] Note that one should not make the mistake of anthropomorphizing AI (see Salles et al. 2020). The notion of malevolence might be of limited value in the context of an artificial agent (Althaus & Baumann, 2020).

EXISTING ACADEMIC LITERATURE

Althaus, D., & Gloor, L. (2016, September). *Reducing risks of astronomical suffering: A neglected priority*. Center on Long-Term Risk. https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly*, *69(4)*, 819–844. https://doi.org/10.1017/S0020589320000366

Clifton, J. (2020, March). *Cooperation, conflict, and transformative artificial intelligence: A research agenda*. Center on Long-Term Risk. https://longtermrisk.org/research-agenda

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, *26*, 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Gloor, L. (2016b, November). *Altruists should prioritize artificial intelligence*. Center on Long-Term Risk. https://longtermrisk.org/altruists-should-prioritize-artificial-intelligence

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hubbard, F. P. (2011). "Do Androids Dream?": Personhood and intelligent artifacts. *Temple Law Review*, *83*(2), 405–474. https://www.templelawreview.org/article/83-2_hubbard

Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, *39*, 98–119. https://doi.org/10.1111/misp.12032

Tomasik, B. (2014). *Do artificial reinforcement-learning agents matter morally?* arXiv. https://arxiv.org/abs/1410.8233

Tomasik, B. (2019b, July 2). *Risks of astronomical future suffering*. Center on Long-Term Risk. https://longtermrisk.org/risks-of-astronomical-future-suffering

EXISTING INFORMAL DISCUSSION

Althaus, D., & Baumann, T. (2020, April 29). *Reducing long-term risks from malevolent actors* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors

Baumann, T. (2017a, September 16). *Focus areas of worst-case AI safety*. Reducing Risks of Future Suffering. https://s-risks.org/focus-areas-of-worst-case-ai-safety

Baumann, T. (2017b, December 15). *Using surrogate goals to deflect threats*. Reducing Risks of Future Suffering. https://s-risks.org/using-surrogate-goals-to-deflect-threats

Baumann, T. (2018a, July 5). *An introduction to worst-case AI safety*. Reducing Risks of Future Suffering. https://s-risks.org/an-introduction-to-worst-case-ai-safety

Baumann, T. (2018b). *A typology of s-risks*. Center for Reducing Suffering. http://centerforreducingsuffering.org/a-typology-of-s-risks

Daniel, M. (2017). *S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017)*. Center on Long-Term Risk. https://longtermrisk.org/s-risks-talk-eag-boston-2017

Tomasik, B. (2018, December 13). *Astronomical suffering from slightly misaligned artificial intelligence*. Essays on Reducing Suffering. https://reducing-suffering.org/near-miss

4.3 Sharing the Benefits of AI

AI could create wealth on an astronomical scale with far-reaching implications for every sector of the economy (Bostrom, 2003a; Hanson, 2001; Makridakis, 2017; Trajtenberg, 2018; Trammel & Korinek, 2020). However, by default those benefits may be captured by a small set of actors, and due to some lock-in effects, the initial distribution of wealth may be hard to change in certain circumstances. If the initial distribution is suboptimal, humanity could permanently lose a significant fraction of its potential, thus constituting a p-risk (see Section 3.2.1). The question of how the gains of AI ought to be distributed—and how to design mechanisms to approach an ideal distribution of gains—may therefore be one of the most important economic questions of our time.

RESEARCH PROJECTS

4.3.1 Distributing Windfall Profits

It seems plausible that AI will enable the accumulation of unprecedented wealth in the hands of a few firms. "Windfall profits" are profits greater than a substantial fraction of the world's total economic output (O'Keefe et al., 2020). How should these profits be distributed? A possible solution is the so-called "Windfall Clause," a voluntary but binding agreement to donate a meaningful portion of profits if they earn a historically unprecedented economic windfall from the development of advanced AI (Bostrom, 2014; O'Keefe et al., 2020). Which other mechanisms are conceivable (see also the Shared Prosperity Initiative)?

4.3.2 Economic Regulation of AI

Technology industries are highly concentrated (Varian, 2001) and AI services may have features of a natural monopoly. Many competition authorities are therefore concerned with avoiding harm to consumers and deadweight loss associated with monopolized AI markets, especially if these markets dominate the world economy. How can antitrust/competition law (U.S. House Judiciary Subcommittee on Antitrust, Commercial and Administrative Law, 2020), utility ratemaking, and other options be used as a tool to check the power of large AI companies, and avoid excessive pricing of AI services without excessively reducing incentives to innovate (Belfield, 2020b; Hua & Belfield, 2020; O'Keefe, 2020b; see also Khan, 2016)? Another promising area concerns investor-state treaty disputes. As large AI companies and governments might use private arbitration to resolve disputes, how can

we ensure that important implications for the long-term future are duly taken into consideration?

4.3.3 Corporate Governance and Firm Incentives

Firms' incentives shape their behavior. Still, profit-maximization alone seems unlikely to be the best incentive structure for firms aiming to develop advanced AI systems. What other firm structures might be desirable to ensure that safety and ethical concerns are given due consideration (Brockman et al., 2019; Feldman et al., 2020)? How can employees (Belfield, 2020a), investors (Belfield, 2020c, 2020d) and other actors (Cihon, et al., 2020) influence corporate decision-making? In particular, which legal instruments are at their disposal (for example, unionization, shareholder resolutions, replacing the board of directors)?

4.3.4 International Coordination and Distribution of Benefits

AI development is concentrated in a small number of already-wealthy countries, but is likely to affect the entire world in the long-run. A maximally beneficial distribution of the benefits from AI will necessarily cross borders (Ó hÉigeartaigh et al., 2020). Yet it is unclear whether existing international institutions responsible for equitably distributing benefits from AI are adequate for this task. What would adequate institutions look like? Will their form and mission vary geographically, and if so, how? How would they interact with governments, NGOs, private AI developers, and existing international bodies? What would beneficiaries' rights against such distributor bodies be?

4.3.5 Intellectual Property

IP regimes may have a significant influence over the development of advanced AI. AI is expensive to produce (Amodei & Hernandez, 2018), but comparatively cheap to copy once produced, making it a prototypical candidate for IP protections. Yet, the IP protections for AI are currently patchwork (Calvin & Leung, 2020), unsettled, and evolving. Reliance on trade secrets also means that AI may be protected indefinitely, unlike copyrighted or patented systems, thus potentially depriving the general public of gains from lower-cost copies of original systems after IP protections expire. It may also create difficulties for regulatory auditing of algorithms (Kroll et al., 2017, p. 658; Tsamados et al., 2020, p. 18). Furthermore, the data-intensity of training AI systems raises questions about infringement during training (O'Keefe et al., 2019). Structuring the IP of AI systems properly may influence both the rate of progress in the field and the magnitude and distribution of economic gains from IP-protected systems. Are the current IP regimes adequate to

balance incentives for innovation and widespread adoption, or ought they be revised to accommodate for unique dynamics in AI? If so, will existing international IP treaties allow such tailoring?

EXISTING ACADEMIC LITERATURE

Belfield, H. (2020a). Activism by the AI community: Analysing recent achievements and future prospects. In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 15–21). https://doi.org/10.1145/3375627.3375814

Belfield, H. (2020b). *From tech giants to a tech colossus: Antitrust objections to the Windfall Clause* [Manuscript submitted for publication].

Belfield, H. (2020c). *Financing our final hour (Pt. I): Institutional investors' Obligations to manage global risks* [Unpublished manuscript].

Belfield, H. (2020d). *Financing our final hour (Pt. II): Institutional investors' strategies for managing global risks* [Unpublished manuscript].

Bostrom, N. (2003a). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(3), 308–314. https://doi.org/10.1017/S0953820800004076

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Calvin, N., & Leung, J. (2020). *Who owns artificial intelligence? A preliminary analysis of corporate intellectual property strategies and why they matter* (Working paper). Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-working-paper-Who-owns-AI-Apr2020.pdf

Cihon, P., Schuett, J., & Baum, S. D. (2020). *Corporate governance of artificial intelligence in the public interest* [Manuscript submitted for publication].

Hanson, R. (2001). *Economic growth given machine intelligence* (Technical Report). University of California, Berkeley. http://mason.gmu.edu/~rhanson/aigrow.pdf

Hua, S.-S., & Belfield, H. (2020). *AI & antitrust: Reconciling tensions between competition law and cooperative AI development* [Manuscript submitted for publication].

Khan, L. M. (2016). Amazon's antitrust paradox. *Yale Law Journal*, *126*(3), 710–805. https://digitalcommons.law.yale.edu/ylj/vol126/iss3/3

Korinek, A. (2019). *Integrating ethical values and economic value to steer progress in artificial intelligence* [Working paper]. National Bureau of Economic Research. https://doi.org/10.3386/w26130

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Ó hÉigeartaigh, S., Whittlestone, J., Liu, Y., Zeng, Y., & Liu, Z. (2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology*, *33*, 571–593. https://doi.org/10.1007/s13347-020-00402-x

O'Keefe, C. (2020b). *How will national security considerations affect antitrust decisions in AI? An examination of historical precedents* [Technical report]. Centre for the Gover-

nance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/How-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-OKeefe.pdf

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020, February). *The windfall clause: Distributing the benefits of AI for the common good* [Technical report]. Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/windfallclause

Trajtenberg, M. (2018). *AI as the next GPT: A political-economy perspective* (Working Paper 24245). National Bureau of Economic Research. https://doi.org/10.3386/w24245

Trammel, P., & Korinek, A. (2020, October). *Economic growth under transformative AI: A guide to the vast range of possibilities for output growth, wages, and the labor share* (GPI Working Paper No. 8-2020). Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/philip-trammell-and-anton-korinek-economic-growth-under-transformative-ai

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2020). *The ethics of algorithms: Key problems and solutions*. SSRN. http://dx.doi.org/10.2139/ssrn.3662302

Varian, H. R. (2001). High-technology industries and market structure. *Proceedings – Economic Policy Symposium – Jackson Hole, Federal Reserve Bank of Kansas City*, 65–101. Archived at https://perma.cc/DZ2B-E7GT

EXISTING INFORMAL DISCUSSION

Amodei, D., & Hernandez, D. (2018, May 16). *AI and compute*. OpenAI. https://openai.com/blog/ai-and-compute

Brockman, G., Sutskever, I., & OpenAI (2019, March 11). *OpenAI LP* [Press release]. https://openai.com/blog/openai-lp

O'Keefe, C., Lansky, D., Clark, J., & Payne, C. (2019). *Before the United States Patent and Trademark Office Department of Commerce: Comment regarding request for comments on intellectual property protection for artificial intelligence*, *Innovation Docket No. PTO–C–2019–0038, Comment of OpenAI, LP addressing question 3*. https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf

4.4 Meta-Research in AI

Shaping the development of advanced AI involves substantial uncertainties. For example, views on AI timelines vary widely (Baum et al., 2011; Grace et al., 2018; Müller & Bostrom, 2016) and researchers disagree on why AI might pose an existential risk (Adamczewski, 2019; Carlier et al., 2020; Cottier & Shah, 2019; Dai, 2018, 2019; Garfinkel, 2018, 2020; Ngo, 2019, 2020). There is no simple answer to the question of how legal scholarship can best contribute to the raised issues. Some resources should therefore be dedicated towards "meta-research," that is to say, addressing high-level uncertainties and methodological questions that arise in

prioritizing legal research. In the following, we list promising AI-specific meta-research projects, while Section 7 concerns meta-research in general.

RESEARCH PROJECTS

4.4.1 Improving Our Ability to Shape the Development of AI in the Future

If one believes that future generations will have more effective ways to shape the development of AI, then one should consider improving their ability to do so.[101] Should we, for example, wait to regulate AI in order to prevent a regulatory backlash (Baum, 2016; Gurkaynak et al., 2016)? How can we ensure that the law remains adaptive to future AI technologies (Maas, 2019b; Moses, 2011)? In particular, how can law-related path dependencies be prevented? What measures can we take today that make governing AI in the future easier? For example, it might be useful to establish AI registers which contain detailed information about potentially harmful AI systems (Floridi, 2020). The law could also help to accumulate resources over a substantial length of time (see Trammell, 2020). To this end, what role do foundation law and tax law play?

4.4.2 Predicting How the Law Will Shape the Development of AI

Predicting AI progress is an important challenge that has received considerable attention (Armstrong & Sotala, 2015; Cremer & Whittlestone, 2020; Etzioni, 2020; Gruetzemacher, 2020; Gruetzemacher et al., 2020; Page et al., 2020). However, there is much less work, if any, that tries to predict how the law will shape the development of AI, even though the law will likely have a significant influence. How has the law shaped the development of other general purpose technologies? To what extent should regulatory impact assessments (OECD, 2009) include long-term implications of AI (see Calvo et al., 2020)?

4.4.3 Clarifying Legal Researchers' Views on the Long-Term Implications of AI

Currently, legal research is mainly concerned with legal questions about today's AI systems (for example, regarding liability, data protection, or anti-discrimination). It is unclear what their views on the long-term implications of AI are, in particular on existential risks, suffering risks, and extreme benefits. Clarifying these views, for example, by conducting specific literature reviews or surveys,

[101] For more information on the underlying view called "patient longtermism," see MacAskill (2020b), Todd (2020a), and Trammell (2020).

would therefore be a valuable research project that could unlock future research opportunities (see Section 3.2.2).

EXISTING ACADEMIC LITERATURE

Armstrong, S., & Sotala, K. (2015). How we're predicting AI—or failing to. In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence* (pp. 11–29). Springer. https://doi.org/10.1007/978-3-319-09668-1_2

Baum, S. D. (2016). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society*, *32*(4), 543–551. https://doi.org/10.1007/s00146-016-0677-0

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, *78*(1), 185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, *2*, 89–91. https://doi.org/10.1038/s42256-020-0151-z

Carlier, A., Clarke, S., & Schuett, J. (2020). *AI risk survey* [Manuscript in preparation].

Cremer, C. Z., & Whittlestone, J. (2020). Canaries in technology mines: Warning signs of discontinuous progress in AI. *Evaluating Progress in AI Workshop – ECAI 2020*. http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_4.pdf

Floridi, L. (2020). Artificial intelligence as a public service: Learning from Amsterdam and Helsinki. *Philosophy & Technology*, *33*, 541–546. https://doi.org/10.1007/s13347-020-00434-3

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Gruetzemacher, R. (2020). *Forecasting transformative AI* [Doctoral dissertation]. Auburn University. https://etd.auburn.edu/handle/10415/7338

Gruetzemacher, R., Dorner, F., Bernaola-Alvarez, N., Giattino, C., & Manheim, D. (2020). *Forecasting AI progress: A research agenda*. arXiv. https://arxiv.org/abs/2008.01848

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review*, *32*(5), 749–758. https://doi.org/10.1016/j.clsr.2016.05.003

Maas, M. M. (2019b). Innovation-proof global governance for military artificial intelligence? *Journal of International Humanitarian Legal Studies*, *10*(1), 129–157. https://doi.org/10.1163/18781527-01001006

Moses, L. B. (2011). *Recurring dilemmas: The law's race to keep up with technological change*. SSRN. http://dx.doi.org/10.2139/ssrn.979861

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Springer. https://doi.org/10.1007/978-3-319-26485-1_33

Page, M., Aiken, C., & Murdick, D. (2020, October). *Future indices: How crowd forecasting can inform the big picture*. Center for Security and Emerging Technology, Georgetown University. https://cset.georgetown.edu/research/future-indices

Trammell, P. (2020). *Patience and philanthropy*. Global Priorities Institute, University of Oxford. https://philiptrammell.com/static/PatienceAndPhilanthropy.pdf

EXISTING INFORMAL DISCUSSION

Adamczewski, T. (2019, May 25). *A shift in arguments for AI risk*. Fragile Credences. https://fragile-credences.github.io/prioritising-ai

Cottier, B., & Shah, R. (2019, August 15). *Clarifying some key hypotheses in AI alignment* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment

Dai, W. (2018, December 16). *Two neglected problems in human-AI safety* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/HTgakSs6JpnogD6c2/two-neglected-problems-in-human-ai-safety

Dai, W. (2019, February 10). *The argument from philosophical difficulty* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/w6d7XBCegc96kz4n3/the-argument-from-philosophical-difficulty

Garfinkel, B. (2019, February 9). *How sure are we about this AI stuff? [*Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/9sBAW3qKppnoG3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff

Garfinkel, B. (2020, July 9). *On scrutinising classic AI risk arguments*. 80,000 Hours. https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments

Ngo, R. (2019, February 21). *Disentangling arguments for the importance of AI safety* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/w6d7XBCegc96kz4n3/the-argument-from-philosophical-difficulty

Ngo, R (2020, September 28). *AGI safety from first principles [*Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

5 Synthetic Biology and Biorisk

Synthetic biology[102] has great potential to shape the long-term future, promising numerous beneficial applications in medicine, fuel, materials science, agriculture, and other industries. Synthetic biology also poses global catastrophic risks to human-originating civilization, threatening serious loss of well-being and life on a global scale and constituting a risk factor.[103] Some extreme cases of this are

---

[102] There is no generally accepted definition of "synthetic biology." The term emerged at the turn of the millenia as an extension of recombinant DNA and genetic engineering in the 1970s and has continued to evolve. For an overview of its development, see National Academies of Sciences, Engineering, and Medicine (NASEM, 2018a), Chapter 2, Acevedo-Rocha (2016), and Way et al. (2014). Today, synthetic biology is frequently defined as applying engineering concepts and approaches to biology (see, e.g., Agapakis, 2014). Another common definition offers two main elements: (a) the design and construction of new biological components and systems, and (b) the redesign of existing, natural biological organisms and systems for useful purposes (Engineering Biology Research Consortium, 2020; Evans, 2014). Policy makers have surveyed and proposed their own definitions (see, e.g., European Commission, 2014; Secretariat of the Convention on Biological Diversity, 2015). For a further survey of definitions, see Nature Biotechnology (2009), and for discussions on other core principles of synthetic biology, see Benner and Sismour (2005), Benner et al. (2011), Endy (2015), Le Feuvre and Scrutton (2018), and Oldham et al. (2012).

Synthetic biology encompasses diverse tools, techniques, and applications from a variety of scientific disciplines and industries. For a discussion of uses and applications, see, for example, König et al. (2013), Pray et al. (2011), and Schmidt and Pei (2010).

[103] There are several definitions for global catastrophic risk, such as those set forth in Bostrom and Ćirković (2007, p. 1) ("The term 'global catastrophic risk' lacks a sharp definition. We use it to refer, loosely, to a risk that might have the potential to inflict serious damage to human well-being on a global scale. On this definition, an immensely diverse collection of events could constitute global catastrophes: potential candidates range from volcanic eruptions to pandemic infections, nuclear accidents to worldwide tyrannies, out-of-control scientific experiment to climate changes, and cosmic hazards to economic collapse."), Cotton-Barratt et al. (2016, p. 1) ("risk of events or processes that would lead to the deaths of approximately a tenth of the world's population, or have a comparable impact."), Millett and Snyder-Beattie (2017) ("We loosely define global catastrophic risk as being 100 million fatalities, and existential risk as being the total extinction of humanity."), Open Philanthropy (2020b) ("We use the term 'global catastrophic risks' to refer to risks that could be globally destabilizing enough to permanently worsen humanity's future or lead to human extinction."), Palmer et al.

existential risks; Ord (2020) estimates that there is a 1 in 30 chance that engineered pandemics will cause an existential catastrophe within the next 100 years.[104] Prioritization research has identified the related fields of biosecurity and governance of synthetic biology and biotechnology as major global priorities (Centre for the Study of Existential Risk, 2020; Future of Humanity Institute, 2020; Lewis, 2020; Open Philanthropy, 2020b; Watson, 2018).[105] Such governance must bridge boundaries between legal and scientific disciplines, between national and international law, between international and national geopolitical areas, and between professionals and amateurs as technology, education, and information become increasingly accessible.

This Section begins with a focus on how the law can reduce existential risk, first by minimizing the likelihood of intentional or unintentional release through preventive measures (Section 5.1) and second by minimizing the negative outcomes upon release through coordination and response (Section 5.2).[106] While we believe legal research to address these existential risks is most important, it also seems worth considering how to steer scientific research and distribute benefits and risks,

---

(2017), Schoch-Spana et al. (2017, p. 1) ("The Johns Hopkins Center for Health Security's working definition of *global catastrophic biological risks* (GCBRs): those events in which biological agents—whether naturally emerging or reemerging, deliberately created and released, or laboratory engineered and escaped—could lead to sudden, extraordinary, widespread disaster beyond the collective capability of national and international governments and the private sector to control. If unchecked, GCBRs would lead to great suffering, loss of life, and sustained damage to national governments, international relationships, economies, societal stability, or global security."), Yassif (2017) ("A GCR is something that could permanently alter the trajectory of human civilization in a way that would undermine its long-term potential or, in the most extreme case, threaten its survival.").

104 For additional estimates of existential risk, see footnote 109. Note that probability estimates of existential catastrophes should be taken with caution (Beard et al., 2020a; see also Baum, 2020b; Beard et al., 2020b; Morgan, 2014; Yudkowsky, 2008b). Ord (2020) acknowledges that there is "significant uncertainty in these estimates, and they should be treated as representing the right order of magnitude" (p. 167). For a discussion of types of uncertainties in estimating natural pandemic risk, see Manheim (2018), and for an estimate that addresses those concerns, see Snyder-Beattie et al. (2019).

105 Governance of synthetic biology has also received considerable attention from the scientific community (see, e.g., Douglas & Stemerding, 2013; Kelle, 2013; Ribeiro & Shapira, 2019; Stirling et al. 2018; Wallach, 2018) and legal community (see, e.g., Mandel & Marchant, 2014), albeit with less attention to the far future.

106 This categorization is presented in NASEM (2018a), Chapter 8, but other, similar typologies may be useful in considering the broad range of risks and how to address them (see Avin, 2018, p. 2; Cotton-Barratt et al., 2020; Farquhar et al., 2017, p. 17; Schoch-Spana et al., 2017).

which may implicate existential risks through loss of potential, as well as pleasure risks and suffering risks (Section 5.3).[107]

5.1 Preventing Intentional or Accidental Release of a Biological Agent

The most desirable outcome is to avoid a catastrophic event entirely (Cotton-Barratt et al., 2020, p. 273). If we can prevent the intentional or accidental release[108] of a biological organism that poses catastrophic or existential risk, human-originating civilization can avoid that harm and retain resources that would have been expended in responding to and mitigating the threat. It seems worthwhile to focus on these anthropogenic risks—those arising out of human activity, such as engineered pathogens—because they may pose much greater existential risk than natural ones (see Lewis, 2020; Ord, 2020, p. 167; Sandberg & Bostrom, 2008).[109] The following avenues of research seem promising:

RESEARCH PROJECTS

5.1.1 Reducing Misuse Risks (Biowarfare, Bioterrorism)

"Misuse" here means any use of synthetic biology with the intention of causing harm. One challenge of preventing misuse in synthetic biology is the evolving risk landscape. Over time, less powerful, non-state actors may pose existential risk, as increasingly powerful tools become more available, less expensive, and easier to use.[110] How might the law address this more distributed and democratized biology

---

107 Notably, now may be a particularly good time for legal research to reduce biorisk. We may be in a window of opportunity for government and private interest and policy change in light of COVID-19; however, this window may be short, focused on natural risks, and tempered by the need to respond to immediate needs (Joshi, 2020; cf. World Bank, 2017, p. 17).

108 For a portrayal of biological risks on a spectrum ranging from natural to accidental to intentional, see Husbands (2018, Figure 1).

109 Ord (2020) estimates that engineered pandemics are roughly 330 times more likely to cause an existential catastrophe by 2120 than naturally arising pandemics. He estimates that the x-risk from natural pandemics is 1 in 10,000 (.01%) and from engineered pandemics is 1 in 30 (3.3%).

Similar results were found in an informal survey conducted at the 2008 Oxford Global Catastrophic Risk Conference, where participants estimated that an engineered pandemic was 40 times more likely to cause human extinction by 2100. The median risk estimate of participants for natural pandemics was .05% and for engineered pandemics was 2% (Sandberg & Bostrom, 2008).

110 Sandberg and Nelson (2020) propose a risk chain model of biorisk to identify what kinds of actors pose the greatest risk. They suggest that in the near future we may be

community? What domestic criminal and civil laws exist to deter and prevent deployment of a biological weapon, and how could they be adapted to better address threats from synthetic biology? What can be learned from deterrence approaches in political science (Knopf, 2010)?[111] How well do traditional legal mechanisms effectively reach this growing set of actors (for example, related to attribution, information hazards, dual-use concerns, and restrictions and monitoring, discussed as separate research projects)? Given the current limitations of the international legal framework to address wrongful acts by non-state actors and biorisks in general,[112] how can international institutions or instruments, such as the Biological Weapons Convention, be strengthened (Means, 2019; Scrivner, 2018; Wilson, 2013)? What new institutions or instruments are desirable?

Similarly, motivations and corresponding sources of harm can vary widely.[113] Existential catastrophe could result from pandemic pathogens (known and recreated, novel, or modified to be more dangerous), widespread eradication of food sources, modified or novel organisms with broad capacity for harm (Schoch-Spana et al., 2017), or other threats that lead to risk factors such as global conflict (Section 3.2.1). There could be erosion of norms against biowarfare that would otherwise provide deterrence, through state dynamics or non-state actions. For example, smaller, targeted biological attacks could become commonplace, similar to cyber

more concerned about highly skilled researchers or other "insider" threats, while less sophisticated actors could pose a similar threat over time, as synthetic biology becomes more accessible through less expensive and easier to use tools and methods.

111 Deterrence may also come from other sources, such as availability and use of a vaccine and other countermeasures. Kosal (2014) argues that improving public health infrastructure could serve as a deterrent to misuse. These are discussed as tools for responding to an event in Section 5.2.4.

112 For example, the Biological Weapons Convention allows ample room for argument that particular research or biological agents have a peaceful purpose, and no mandatory verification or enforcement mechanisms exist. There are confidence-building measures—annual declarations of critical information on research, development, and more—which were introduced "in order to prevent or reduce the occurrence of ambiguities, doubts and suspicions and in order to improve international co-operation in the field of peaceful biological ambiguities" (United Nations Office for Disarmament Affairs, 2015); however, there are few, if any, consequences for failing to participate (Chevrier & Hunger, 2000, pp. 31–32). By comparison, the Chemical Weapons Convention (CWC) allows for strict verification of compliance following mandatory destruction of all declared chemical weapons and production sites, as well as possible "challenge inspections." However, the CWC has a similar issue with dual-use, and "chemical weapon" is defined by intended purpose rather than lethality or quantity.

113 Possible motivations could be political, economic, or sociocultural, perhaps to seek attention, make a statement, blackmail, incapacitate, destabilize, retribute, or deter (Gandhi et al., 2011; Revill, 2017, Figure 2 at pp. 630–631).

attacks with economic motivations.[114] How can the law adapt to the changing risk landscape? Would different legal mechanisms be appropriate to deter release from different motivations, and are any of these motivations more concerning or likely to pose existential risk? Is there a risk of norms against biowarfare being eroded (Ilchman & Revill, 2014), and if so, how can the law promote "biopeace"? How could this look different for international and national law?

5.1.2 Reducing Accident Risks (Biosafety)

"Accident risks" here are defined as any unintentional release of a harmful biological agent.[115] Biosafety regulations and guidelines apply to research involving infectious agents, toxins, and other biological hazards, aiming to safeguard against accidental release, ensure reporting and transparency about accidents, and provide oversight and monitoring.[116] However, some have argued that even maximum containment labs are prone to error and thus inadequate for potential pandemic pathogens (Klotz, 2019). What kind of containment, reporting, and transparency mechanisms would be more effective? What could be learned from accident reporting in other industries, such as aviation (Gronvall, 2015, p. 6), or high reliability organizations (Roberts & Bea, 2001)? Do existing criminal law provisions penalize the (concrete or abstract) increase of existential accident risks (e.g., Section 221 of the German Criminal Code; see also Duff & Marshall, 2015; Simester & von Hirsh, 2009), discussed more in Section 6.1.9? What other legal mechanisms are conceivable to reduce accident risks, such as deterrence via civil liability?

While existential risk from accidents was once limited to academic and commercial labs, it is increasingly within the reach of other groups and individuals.

---

114 As synthetic biology and biological agents are used for production in materials science and other industries, those same industries will also become susceptible to biowarfare.

115 Compare to accident risks in artificial intelligence, which encompass "any unintended and harmful behavior of an AI system" (Section 4.1.1). In the discussion of synthetic biology, accident risk focuses on the specific risk of unintentional release, while unintentional consequences are discussed separately. For an informal discussion of historical accidental release of pandemic pathogens, see Shulman (2020).

116 While the term "biosafety" has several accepted definitions (Beeckman & Rüdelsheim, 2020), here we use it to refer specifically to principles and practices to prevent unintentional release or exposure. Biosafety guidelines commonly specify different levels of biocontainment precautions required to isolate dangerous biological agents in a facility, referred to as biosafety level (BSL), containment level (CL), or pathogen/protection level (P), with BSL-1/CL1/P1 as the lowest and BSL-4/CL4/P4 as the highest. In the United States, the Centers for Disease Control and Prevention specify these levels. The same levels are defined in the European Union Directive 2000/54/EC, Biological Agents at Work, the Canadian Biosafety Standards and Guidelines, and elsewhere (National Academy of Sciences & National Research Council, 2012, Chapter 4 & Appendix E).

Synthetic biology no longer requires years of training and experience in laboratories, where biosafety and containment protocols are accompanied by certification programs and institutional oversight. A scientist could theoretically find themselves in safety situations that exceed their biosafety experience. Powerful equipment and technologies outside of a lab may go without regular maintenance or checks and result in bio-errors. In the context of accidents, existential risk seems most likely from release of a potential pandemic pathogen. How can the law reduce existential risk from accidents outside of traditional laboratories? What role should professional self-regulation, best practices and norms (Open Philanthropy, 2017), and other forms of soft law play?

Comparative law may offer insights on potential gaps and more effective measures, yet little research exists comparing biosafety governance in different countries, let alone the relative effectiveness of different strategies. What laws and regulations exist in different countries to minimize accident risks (Beeckman & Rüdelsheim, 2020, Appendix 1; National Academy of Sciences & National Research Council, 2012, Chapter 4 & Appendix E; Osman, 2018; Van Houten & Fleming, 1993)? To what extent have they been implemented in practice?[117] How might their effectiveness be measured, and what uncertainties exist in such an analysis? What do they reflect about biosafety norms? How have different nations attempted to regulate the DIY bio community, and with what result?

5.1.3 Restrictions and Monitoring Measures

Laws that impose lab safety requirements or place other limits on research, use, or access to materials and equipment reduce existential risk by making it more difficult to develop, produce, or accidentally release the most harmful biological agents. The effectiveness of those laws depends on the ability to verify and enforce compliance. However, biological weapons, including those made with synthetic biology, have characteristics that make verification and enforcement technically difficult, compared to nuclear and chemical weapons (Bakerlee et al., 2020; Bressler & Bakerlee, 2018). Biological weapons require fewer resources and are relatively easy to develop and manufacture in secret, due to the multiple-use nature of materials, equipment, and techniques used.[118]

---

[117] According to Gronvall (2015), "There is now adequate guidance for laboratories to develop oversight systems to catch and contain accidents, but not all research institutions adhere to such guidance, require adequate training, or have sufficient resources to dedicate to biosafety. There is also great variability from one research institution to another, even within a nation."

[118] "The knowledge, materials, and technologies needed to make and use a biological weapon are readily accessible around the world." Gronvall (2017).

Consider that nuclear weapons require highly enriched uranium, which emits readily detectable radiation, as well as specific equipment and infrastructure that is expensive, technologically advanced, large and difficult to hide, and has few other uses. In contrast, synthetic biology has no need for large facilities and uses materials and equipment that are widely used for a variety of research projects, without a clear indicator of malicious intent. Biological materials are widely available in labs and nature, and it is increasingly possible to synthesize materials and organisms *de novo*, allowing actors to circumvent screening requirements[119] and avoid attribution (e.g., Gronvall, 2016, pp. 36–41; Gronvall et al., 2009, p. 434).

In international law, the Biological Weapons Convention lacks effective monitoring and enforcement mechanisms (Means, 2019; Scrivner, 2018) and faces financial and political challenges.[120] What legal mechanisms have been used or proposed for the monitoring and enforcement of legal instruments, for example through verification, transparency, confidence-building measures, and other measures short of verification (Lentzos, 2019)? What can be learned from existing compliance and enforcement protocols for other weapons and controlled agents (Becker et al., 2005)? What measures are most effective to prevent proliferation when considering existential risk reduction, rather than considering the ability to strictly verify compliance?

In national law, what legal mechanisms might exist, such as screening and restrictions on providing dual-use technology and materials (Garfinkel, 2007; Kobokovich et al., 2019)? What might be effective for different points of intervention (for example, equipment, labs, vendors, institutional researchers, DIY bio community)? Is there a need for stronger forms of supervision (Bostrom, 2019)? If so, would they violate civil rights and liberties? What limits should exist on monitoring and surveillance, such as to prevent abuse or avoid an attractor state or lock-in to a totalitarian state? Do specific synthetic biology applications have adequate oversight (Gronvall, 2015, p. 8)? More broadly, how can oversight mechanisms adapt as circumstances change, such as with emerging technology or changing risks? What role could soft law, such as other guidance and norms, have in a monitoring regime at an international (Cameron et al., 2020) or national level?

---

119 Early proposals and guidance sought to address concerns that pathogen or toxin DNA could be manipulated or created through the use of nucleic acid synthesis technologies by requiring commercial firms to screen purchases for synthetic DNA (e.g., Garfinkel et al., 2007; U.S. Department of Health and Human Services, 2020). However, changes to gene synthesis technologies and market conditions have reduced the efficacy of these biosecurity protections (Kobokovich et al., 2019), a trend likely to continue as technology develops.

120 A joint NGO statement in 2018 described the Convention as "in a precarious state," with financial debts from certain state parties putting its operation at risk (Center for Global Health Science and Security et al., 2018).

5.1.4 Attribution

Attribution is the ability to identify or rule out the source of a biological threat. Attribution offers three main security benefits, which can reduce existential risk: (a) informing response efforts and mitigating consequences by providing information about the motive of the actor and capabilities of the biological agent, (b) identifying responsible parties for appropriate legal recourse, and (c) deterring reckless accident and misuse, and preventing future misuse by the same actors, if perpetrators are held accountable[121] (Lewis et al., 2020). In the context of synthetic biology, attribution involves determining whether a biological agent involved has been genetically engineered and, if so, where it was engineered, by whom, and why.

Attribution of synthetic biology agents poses unique technical challenges. Biological agents may be developed and deployed in a clandestine manner. Once released, they may propagate, replicate, and mutate in unpredictable ways, making it more difficult to identify the actor or location of release.[122] Technical forensics may aid in attribution,[123] but are not as reliable or complete as for nuclear and chemical weapons. As a result, attribution of synthetic biology agents may depend on non-technical indicators (for example, location, victims, epidemiological features) and intelligence (for example, human sources, communications, surveillance and monitoring data). How can the law ensure that several sources of information are available to support or supplement technical measures (for example, legal ability to collect samples, gather intelligence)? How can attribution methods meet standards for admissibility as evidence under national law or at an international tribunal (Bidwell & Bhatt, 2016, pp. 18–20)?

Development of attribution measures may have the unintended effect of increasing certain risks. First, the possibility of being found culpable may motivate concealment of misuse or accident in a way that could create or aggravate risk

---

121 Attribution is only meaningful if it leads to some form of legal recourse, as described above for accident and misuse. Attribution is of limited value if an actor intends to claim responsibility or can avoid consequences for misuse or accidental release of a biological agent.

122 Compare to chemical and nuclear weapons, which can generally be traced (cf. footnotes 119–120 and accompanying text, discussing technical challenges in monitoring the development and production of biological weapons compared to chemical and nuclear weapons).

123 Attribution tools for synthetic biology include, for example, advanced sequencing to rapidly characterize an agent (NASEM, 2018a, Box 8-2 and accompanying text), forensics to detect engineering and identify the engineer (Lewis et al., 2020; Scoles, 2020; see also IARPA, 2020; NASEM, 2017a), machine-learning tools to predict lab-of-origin, nation-of-origin, and ancestor lab (Alley et al., 2020), and microbial forensics (National Research Council, 2014).

(Cotton-Barratt et al., 2020, p. 274).[124] Second, tools and techniques used for attribution are dual-use, meaning they might also be used to evade attribution. How can the law minimize these risks? What role can the law play in balancing the benefits of developing attribution measures with the risks from their dual-use nature?

5.1.5 Dual-Use Concerns

"Dual-use" refers to something that can be used for beneficial purposes or to cause harm.[125] The dual-use nature of synthetic biology poses an existential risk, from misuse as well as accident, as research to advance beneficial applications may have harmful applications or present other risks. Legal instruments create prohibitions based on these dichotomies (Millett, 2017), yet much research and technology exists along some spectrum of dual use; even extremely dangerous biotechnology has a plausible argument for how it could have a defensive or peaceful use, or may itself create the need for research on a countermeasure. Notably, dual-use concerns have been raised by gain-of-function research, in which a biological entity is given a new property.[126] What types of institutions and legal mechanisms have been used to reduce existential risk from dual-use concerns throughout the research life cycle—such as prohibitions on certain types of research or involving certain materials, limiting access to materials and equipment, export controls (Kanetake, 2018), intellectual property restrictions, oversight committees at different stages (NASEM, 2018b, pp. 43–58 & Table 3-1; Resnik, 2013), and advisory boards such as the National Science Advisory Board for Biosecurity (NASEM, 2017b, pp. 31–38)? How could international instruments or institutions be strengthened or created to address dual-use concerns (Millett, 2017; NASEM, 2017b, pp. 38–44)? What can we learn from existing regulations (Lev, 2019)? What other mechanisms are conceivable (Marcello & Effy, 2018)? What limits do constitutional law and other instruments or rights place on mechanisms to control research and development (Ram, 2017; Santosuosso et al., 2007)? What role can norms, codes of ethics, and other soft law play (NASEM, 2018b, pp. 58–78 & Table 3-2)?

---

[124] For case studies of how this incentive can weaken prevention and response, see Chernov and Sornette (2016).

[125] Dichotomies of dual use have been conceptualized as: (a) war or peace, (b) good or evil, (c) offense or defense, (Evans & Commins, 2017), and (d) military or civilian (Mahfoud, et al., 2018). For an informal discussion of understandings of dual-use, see Weiss Evans (2018).

[126] From a scientific perspective, not all gain-of-function research is concerning, such as research to confer pest resistance to crops. However, the term "gain-of-function" often refers specifically to gain-of-function research *of concern*, in the same way that "dual-use" often refers specifically to dual-use research, technology, or materials *of concern*.

By clearly identifying what is and is not prohibited, the law could set clear expectations and support decisive action. In an international agreement, clear definitions could also reduce doubt, suspicion, and proliferation throughout other countries seeking to protect themselves, and thereby reduce overall biorisk (cf. Enemark, 2017; Joint NGO statement, 2018). However, it is also important that a dual-use framework remain adaptable to changing risk considerations. In what ways can the law create bright-line rules to identify dual-use research, materials, and technology of concern? What about bright and fuzzy lines? What kind of framework or delineations would be useful (for example, categories for what is permitted, prohibited, or permitted with special oversight or regulatory requirements)? Given that the weight of considerations may change over time as new defense- and offense-enabling technologies come into play (cf. Lewis, 2019; NASEM, 2017a), what kind of process would be appropriate to (a) assign categories and (b) update these assignments with some frequency (Dubov, 2014, p. 251; Palmer, 2020)? How would this interact with legal mechanisms for addressing information hazards? What can we learn from other fields of law?

5.1.6 Information Hazards

Biorisks arise not only from biological materials, but also from biological information; information can also be dual use. "Information hazards" are risks that arise from dissemination or potential dissemination of true information that may cause harm or enable some agent to cause harm (Bostrom, 2011b). If published, they may give ideas or implementation details to those who would misuse or carelessly use it (Crawford et al., 2019). The dual-use nature of much biological information makes it difficult to draw clear lines around what information is a hazard or what scientific research could produce hazardous information (Lewis et al., 2019). How can the law anticipate and manage potential information hazards (Lewis, 2018b; Lewis et al., 2019, pp. 979–980)? What can be learned from discussions on broader dual-use concerns or on information hazards in other fields? What legal mechanisms or areas of law have been used or are conceivable to address information hazards—such as export controls (Hindin et al., 2017; NASEM, 2017b, pp. 47–50), administrative law, security classification, or intellectual property law? How could the regulation of such information adapt to the changing risks over time? To what extent is restricting scientific knowledge consistent with applicable constitutional law (Ram, 2017)? What role should professional self-regulation, journal policies (Casadevall et al., 2013), best practices and norms, and other forms of soft law play?

5.1.7 Reducing Unintended Consequences

Emerging synthetic biology technologies could pose risks that are unknown or difficult to anticipate with specificity at the time of deployment. Thus, even intentional release of organisms could carry a risk of unintended harmful consequences.[127] While the nature of some risks may be known, there could still be uncertainty about its likelihood and specific details. How can the law reduce the risk of harmful and unintended effects stemming from synthetic biology? What kind of analysis is appropriate to assess the risks and benefits (Section 6.3.1)? Is there a need for reporting, registration, other documentation of certain types of information, required containment and response strategies, ongoing monitoring following release, or liability schemes (e.g., Warmbrod et al., 2020)? If so, how can this be done effectively?

Gene drives present a specific need to reduce unintended consequences. A gene drive is a type of genetic element that improves its own chances of inheritance in future generations. Through genetic engineering, gene drive systems can be used to suppress a population (for example, disease vectors, plant pests) or alter most of a population to express a desired trait (for example, to increase traits that correspond with well-being or survival of desired species, to increase productivity of resources that are heavily harvested). Due to the nature of gene drives, they present a greater risk of competing with native species and acting like an invasive species, leading to greater concern for potential movement across political boundaries. If multiple gene drives target the same organism (or less likely, the same sequence), there could also be unexpected and unintended interactions (Warmbrod et al., 2020, p. 20 & Appendix 2). How can the law reduce risks from environmental release and transboundary movement of organisms with gene drives (Kuzma & Rawls, 2016; Warmbrod et al., 2020), at a national and international level? What national and international laws exist and might address release of organisms with gene drives (e.g., NASEM, 2016a, Chapter 8; Rabitz, 2019)? What other biosafety, risk assessment, and regulatory measures or legal institutions could address gene drive research and reduce risk of and mitigate unintended consequences (Kofler et al., 2018; Warmbrod et al., 2020)? What factors should be considered, such as persistence and reversibility (Eckerström Liedholm, 2019), and specific technical solutions to meet them, such as a self-extinguishing daisy-drive to make untested gene drives less persistent, or ensuring reversibility with a tested reversal drive

[127] For example, (a) modified microbes could have allergenic properties, transfer antibiotic resistance into a harmful strain of bacteria, or cause a microbial strain to become pathogenic, and (b) environmental release could have unforeseen consequences on the balance of functioning ecosystems, lead to competition with native species, or result in horizontal gene transfer (i.e., to non-target organisms) (e.g., Hewett et al., 2016).

(Warmbrod et al., 2020; see also Backus & Delborne, 2019)?[128] How can the law facilitate coordination and communication between researchers and stakeholders? What legal instruments exist that already apply? What areas are unsettled?

5.1.8 Flexible and Clear Regulatory Approach

Specific language in regulation of technology can limit its applicability to that which is known now. For example, list-based approaches that create bright lines allow emerging developments to escape regulation (Carter & Friedman, 2015, pp. 8-9 and throughout). What alternatives exist to list-based approaches,[129] which might create a more flexible safety net (Casadevall & Relman, 2010; DiEuliis et al., 2017; Lewis, 2020; Lewis et al., 2019; NASEM 2018a, Chapter 8)? What can we learn from related research on flexible constitutions (Section 6.1.3)?

However, ambiguity may limit enforceability, or even sow doubt and encourage proliferation in an international context (cf. Enemark, 2017). For example, the Biological Weapons Convention describes "microbial or other biological agents, or toxins" with no "protective" purpose, providing considerable room for argument. Especially for international agreements, how can a legal instrument ensure sufficient clarity to reduce doubt and corresponding defensive proliferation, while also allowing adaptability? How might these instruments and institutions be designed to facilitate easier consensus around updating provisions or interpretations?

EXISTING ACADEMIC LITERATURE

Bakerlee, C., Guerra, S., Parthemore, C. Soghoian, D., & Swett, J. (2020). *Common misconceptions about biological weapons*. Council on Strategic Risks. https://councilonstrategicrisks.org/2020/12/07/briefer-common-misconceptions-about-biological-weapons/

Becker, U., Müller, H., & Wunderlich, C. (2005). While waiting for the protocol. *The Nonproliferation Review*, *12*(3), 541–572. https://doi.org/10.1080/10736700600601194

[128] Reversal drives allow researchers to contain the damage and manage unforeseen consequences from release of an organism with a gene drive. By default a gene drive is persistent, requiring only a single process; however, a daisy-drive is self-extinguishing (Noble et al., 2019), providing a way to reduce geographic spread and conduct more limited field trials by limiting the number of generations it can spread (Eckerström Liedholm, 2019).

[129] List-based approaches include the Select Agent Regulations set forth by the U.S. Department of Health and Human Services and U.S. Department of Agriculture, which review and republish the lists at least every other year (Centers for Disease Control, 2020), and the seven experiments of concern (National Research Council, 2004; Rapport, 2014).

Beeckman, D. S. A., & Rüdelsheim, P. (2020). Biosafety and biosecurity in containment: A regulatory overview. *Frontiers in Bioengineering and Biotechnology*, *8*(650), 1–7. https://doi.org/10.3389/fbioe.2020.00650

Bidwell, C. A., & Bhatt, K. (2016, February). *Use of attribution and forensic science in addressing biological weapon threats: A multi-faceted study*. Federation of American Scientists. https://fas.org/pub-reports/biological-weapons-and-forensic-science/

Bostrom, N. (2011b). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, *10*, 44–79. https://nickbostrom.com/information-hazards.pdf

Carter, S. R., & Friedman, R. M. (2015, October). *DNA synthesis and biosecurity: Lessons learned and options for the future*. J. Craig Venter Institute. https://www.jcvi.org/research/dna-synthesis-and-biosecurity-lessons-learned-and-options-future

Casadevall, A., & Relman, D. (2010). Microbial threat lists: obstacles in the quest for biosecurity?. *Nature Reviews Microbiology*, *8*(2), 149-54. https://doi.org/10.1038/nrmicro2299

Cotton-Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: Prevention, response, resilience, and why they all matter. *Global Policy*, *11*(3), 271–282. https://doi.org/10.1111/1758-5899.12786

Dubov, A. (2014). The concept of governance in dual-use research. *Medicine, Health Care and Philosophy*, *17*, 447–457. https://doi.org/10.1007/s11019-013-9542-9

Enemark, C. (2017). *Biosecurity dilemmas*. Washington, DC: Georgetown University Press. http://press.georgetown.edu/book/georgetown/biosecurity-dilemmas

Gronvall, G. K. (2015, February). *Mitigating the risks of synthetic biology*. Council on Foreign Relations: Center for Preventive Action. https://www.jstor.org/stable/resrep24166

Gronvall, G. K. (2016). *Synthetic biology: Safety, security, and promise*. Baltimore, MD: CreateSpace Independent Publishing Platform.

Gronvall, G. K., Bouri, N., Rambhia, K. J., Franco., C., & Watson, M. (2009). Prevention of biothreats: A look ahead. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 7(4), 433–442. https://doi.org/10.1089/bsp.2009.1112

Ilchmann, K., & Revill, J. Chemical and biological weapons in the 'New Wars'. *Science and Engineering Ethics*, *20*, 753–767 (2014). https://doi.org/10.1007/s11948-013-9479-7

Kobokovich, A., West, R., Montague, M., Inglesby, T., & Gronvall, G. K. (2019). Strengthening security for gene synthesis: Recommendations for governance. *Health Security*, *17*(6), 419–429. http://doi.org/10.1089/hs.2019.0110

Kofler, N., Collins, J. P., Kuzma, J., Marris, E., Esvelt, K., Nelson, M. P., Newhouse, A., Rothschild, L. J., Vigliotti, V. S., Semenov, M., Jacobsen, R., Dahlman, J. E., Prince, S., Caccone, A., Brown, T., Schmitz, O. J. (2018, November 2). Editing nature: Local roots of global governance. *Science*, *362*(6414), 527–529. https://doi.org/10.1126/science.aat4612

Lentzos, F. (2019). Compliance and enforcement in the biological weapons regime. United Nations Institute for Disarmament Research. https://www.unidir.org/sites/default/files/2020-02/compliance-bio-weapons.pdf

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

Lewis, G., Jordan, J. L., Relman, D. A., Koblentz, G. D., Leung, J., Dafoe, A., Nelson, C., Epstein, G. L., Katz, R., Montague, M., Alley, E. C., Filone, C. M., Luby, S., Churche, G. M., Millett, P., Esvelt, K. M., Cameron, E. E., Inglesby, T. V. (2020). The biosecurity benefits of genetic engineering attribution. *Nature Communications*, *11*(6294). https://doi.org/10.1038/s41467-020-19149-2

Marcello, I., & Effy, V. (2018). Dual use in the 21st century: emerging risks and global governance. *Swiss Medical Weekly*, *148*(14688). https://doi.org/10.4414/smw.2018.14688

Millett. P. D. (2017, January 17). Gaps in the international governance of dual-use research of concern. In National Academies of Sciences, Engineering, and Medicine, *Dual use research of concern in the life sciences*. DC: The National Academies Press. https://doi.org/10.17226/24761 (under the Resources tab)

National Academies of Sciences, Engineering, and Medicine (2018a). *Biodefense in the age of synthetic biology*. The National Academies Press. https://doi.org/10.17226/24890

National Academies of Sciences, Engineering, and Medicine (2018b). *Governance of dual use research in the life sciences: Advancing global consensus on research oversight: proceedings of a workshop*. The National Academies Press. https://doi.org/10.17226/25154

National Academies of Sciences, Engineering, and Medicine (2017a). *A Proposed Framework for Identifying Potential Biodefense Vulnerabilities Posed by Synthetic Biology: Interim Report*. The National Academies Press. https://doi.org/10.17226/24832

National Academies of Sciences, Engineering, and Medicine (2017b). *Dual use research of concern in the life sciences: Current issues and controversies*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24761

National Academies of Sciences, Engineering, and Medicine (2016a, July 28). *Gene drives on the horizon: Advancing science, navigating uncertainty, and aligning research with public values*. The National Academies Press. https://doi.org/10.17226/23405

National Academy of Sciences and National Research Council (2012). *Biosecurity challenges of the global expansion of high-containment biological laboratories: Summary of a workshop*. The National Academies Press. https://doi.org/10.17226/13315

Nouri, A., & Chyba, C. F. (2008). Biotechnology and biosecurity. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks*. Oxford University Press.

Palmer, M. J. (2020). Learning to deal with dual use. *Science*, *367*(6482), 1057. https://doi.org/10.1126/science.abb1466

Rabitz, F. (2019). Gene drives and the international biodiversity regime. *Review of European, Comparative and International Environmental Law*, *28*(3), 339–348. https://doi.org/10.1111/reel.12289

Ram, N. (2017). Science as speech. *Iowa Law Review*, *103*(3), 1187–1238. https://ilr.law.uiowa.edu/print/volume-102-issue-3/science-as-speech/

Resnik, D. B. (2013). Scientific control over dual-use research: prospects for self-regulation. In B. Rappert, & M. J. Selgelid (Eds.), *On the dual uses of science and ethics. Principles, practices and prospects* (pp. 237–254). Canberra: Australian National University E-Press.

Sandberg, A., & Nelson, C. (2020, June 10). Who should we fear more: Biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk. *Health Security*, *18*(3), 155–163. https://doi.org/10.1089/hs.2019.0115

Santosuosso, A., Sellaroli, V., & Fabio, E. (2007). What constitutional protection for freedom of scientific research? *Journal of Medical Ethics*, *33*(6), 342–344. http://dx.doi.org/10.1136/jme.2007.020594

Schoch-Spana, M., Cicero, A., Adalja, A., Gronvall, G., Kirk Sell, T., Meyer, D., Nuzzo, J. B., Ravi, S., Shearer, M. P., Toner, E., Watson, C., Watson, M., & Inglesby, T. (2017). Global catastrophic biological risks: Toward a working definition. *Health Security*, *15*(4), 323–328. https://doi.org/10.1089/hs.2017.0038

Scrivner, S. (2018). Regulations and resolutions: Does the BWC prevent terrorists from accessing bioweapons? *Journal of Biosecurity, Biosafety, and Biodefense Law*, *9*(1), 1–5. https://doi.org/10.1515/jbbbl-2018-0006

Warmbrod, K. L., Kobokovich, A., West, R., Ray, G., Trotochaud, M., & Montague, M. (2020, May 18). *Gene drives: Pursuing opportunities, minimizing risk*. Johns Hopkins Bloomberg School of Public Health, Center for Health Security. https://www.centerforhealthsecurity.org/our-work/publications/gene-drives-pursuing-opportunities-minimizing-risk

EXISTING INFORMAL DISCUSSION

Berger, A. (2014, June 26). *Potential global catastrophic risk focus areas*. Open Philanthropy. https://www.openphilanthropy.org/blog/potential-global-catastrophic-risk-focus-areas

Bressler, D., & Bakerlee, C. (2018, December 6). *"Designer bugs": How the next pandemic might come from a lab*. Vox. https://www.vox.com/future-perfect/2018/12/6/18127430/superbugs-biotech-pathogens-biorisk-pandemic

Centre for the Study of Existential Risk. *Global catastrophic biological risks*. University of Cambridge. https:// www.cser.ac.uk/research/global-catastrophic-biological-risks

Crawford, M., Adamson, F., & Ladish, J. (2019, September 16). *Bioinfohazards* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/ixeo9swGQTbYtLhji/bioinfohazards-1

Klotz, L. (2019, February 25). *Human error in high-biocontainment labs: a likely pandemic threat*. Bulletin of the Atomic Scientists. https://thebulletin.org/2019/02/human-error-in-high-biocontainment-labs-a-likely-pandemic-threat/

Lewis, G. (2020, March). *Reducing global catastrophic biological risks*. 80,000 Hours. https://80000hours.org/problem-profiles/global-catastrophic-biological-risks/

Lewis, G. (2018b, February 19). *Horsepox synthesis: A case of the unilateralist's curse?* Bulletin of the Atom Scientists. https://thebulletin.org/2018/02/horsepox-synthesis-a-case-of-the-unilateralists-curse/

5.2 Coordination and Response

Many have recognized the need for global cooperation in order to avoid existential risk (see, e.g., Bostrom, 2013; Farquhar et al., 2017, p. 6). This holds true for biological risks, where one nation can have a global impact, and often a multilateral approach is most effective (Heyman et al., 2009). Infectious disease, organisms,

and knowledge are not confined to national borders; potential pandemic pathogens can spread with increasing ease due to globalization and air travel, and organisms with gene drives may travel across political boundaries. To respond effectively, there must be a shared and cooperative approach for the detection and mitigation of threats to global health. Nations must also coordinate local response and manage sharing of information across local and national boundaries.

Research to improve coordination and response is dual purpose, as many legal and technical measures to detect and respond to anthropogenic biological risks, such as robust surveillance systems, availability of medical countermeasures, and surge capacity for healthcare systems, are also relevant to natural pandemics (NASEM, 2018a, Chapter 8). The following research projects detail mechanisms through which the law could reduce existential risk by improving global and local coordination and response.

RESEARCH PROJECTS

5.2.1 Global Cooperation

While many actors can help address global catastrophic risk and existential risk, the international community will probably need to play a major role (Cotton-Barratt et al., 2016, p. 88). Promising legal research on global cooperation and response could first survey the landscape and identify areas for change. What international legal frameworks are relevant to synthetic biology (Keiper & Atanassova, 2020; Lai et al., 2019, Table 1), and what protocols or mechanisms do they have for ongoing review and changes? How might these mechanisms be strengthened?

It would also be useful to learn how international bodies could more easily reach consensus. Future implications of synthetic biology may be difficult to predict and warrant an adaptable method of governance (Zhang, 2011), and efforts to adapt or strengthen existing instruments have faced different limitations and challenges. What meta-level process could be used to reach consensus on topics such as dual-use, information hazards, and emerging technology risks? What flexible and evolving art of governance would facilitate effective interactions among current and emerging actors, with representation by various stakeholders? What would cultivate accountability, mutual trust, and responsiveness to emerging technologies and concerns? What role could an institution or protocols within an instrument play? What can we learn from more general research on mechanisms of cooperation and world governance (Section 6.1.2)?

5.2.2 Global Pandemic Response

A primary concern is international detection and response to a potential pandemic. An epidemic, which is an outbreak of disease affecting many people within a region, if not contained can become a pandemic, which is spread over a wider geographic area (usually multiple countries or continents) and affects a high proportion of the population (Merriam-Webster). There is a need for collective preparedness, as risky governance by one nation could endanger others and lead to global catastrophic or existential risk. Given the risk of a natural or engineered pandemic,[130] it seems worthwhile to investigate the specific question of pandemic detection and response. What can we learn from how nations and global institutions have responded to epidemics and pandemics in the past (e.g., Sirleaf, 2018a)? What legal institutions or tools can help with rapid anticipation, prevention, and response to outbreaks (Farquhar et al., 2017, Section 2.2; Sirleaf, 2018b)? How could existing institutions or instruments, such as the International Health Regulations,[131] be adapted to better address this need?

5.2.3 Pandemic Finance

Preventing and managing the spread of an epidemic requires both a source of funds and effective mobilization of these funds for response. Several funding sources exist but are problematic for responding to a potential pandemic; funds may be preallocated, distributed too slowly to prevent spread, dependent on private giving, take the form of undesirable loans (Bruns, 2019), or, as in the case of the World Bank Group's Pandemic Emergency Financing Facility, discontinued (Hodgson, 2020). What institution or legal mechanism could facilitate financing pandemic response and management, ensuring that funds are available, allocated to pandemic response, and distributed effectively? What kind of trigger will ensure that money and resources are delivered in a timely manner, to catch a potential pandemic as early as possible (see Meenan, 2020; NASEM, 2016b, ch. 6)? What kind of insurance (Taylor, 2008; Cotton-Barratt, 2014), reinsurance (Anthony & Neill, 2020),

130 See above, footnote 109 and accompanying text.

131 The International Health Regulations (IHR) were adopted by the World Health Assembly in 1969 and last revised in 2005, aim "to prevent, protect against, control and provide a public health response to the international spread of disease in ways that are commensurate with and restricted to public health risks, and which avoid unnecessary interference with international traffic and trade" (Article 2). They require members to assess events within their respective territories and use directives set forth in the IHR, including notice of initial assessment; public health information; measures taken to respond; and ongoing information regarding studies, cases and deaths, and spread of the disease (Article 6).

financial institution, capital market instrument, or other instruments are conceivable, and how might they interact (Farquhar et al., 2019; NASEM, 2016b)? What are their advantages and disadvantages? What can we learn from their usage in other fields?

5.2.4 National Public Health Preparedness

In an ideal response to a potential pandemic or other public health emergency, a nation detects the threat early and responds appropriately. Responsiveness hinges on several factors, including coordination among government agencies, officials, and non-government actors; clear roles and responsibilities; preparedness testing; surveillance, monitoring, and reporting capabilities for early detection (for example, epidemiological methods of identifying victims, agents, and modes of transmission); countermeasures and a robust supply chain for quick response; mitigation strategies, emergency response, availability of supportive health care facilities, and effective procedures for isolation and quarantine; and legal ability to enact and enforce pharmaceutical and nonpharmaceutical interventions (see Avin et al., 2018, Figure 3, p. 5; Khan, 2018; Kun, 2014; NASEM, 2017a, p. 34; NASEM, 2020a; Nelson et al., 2007). What institutions, framework, or infrastructure would allow for a quick and effective response to a biological threat? What are the barriers? What powers are useful or necessary for oversight, monitoring, and response? Should any of these be limited for use in certain circumstances, and if so, which ones and how? To what extent are they consistent with existing law?[132] Presented as a separate research project is the question of coordination among different actors.

5.2.5 National Coordination

Nations often rely on several actors to prepare for biorisk, detect a threat early, and respond appropriately. Coordination is a key factor, as detection and response

[132] More specifically, near-term research could address specific legal mechanisms or powers, bridging the near- and long-term: What specific legal mechanisms could be used to implement public health interventions as preventive or responsive measures (for example, vaccines, mask mandates, travel restrictions for individuals who are ill or traveling from a suspect country, quarantine, or isolation mandates, air filtration requirements for businesses remaining open during a pandemic, measures to prevent spread of misinformation)? What exemptions are or would be permitted under existing laws, and what is the impact on biorisk? To what extent are biomonitoring and contact tracing (for example, metadata on hospital visitation and symptoms, broader network effects bigger than individual level of contact tracing) consistent with applicable privacy laws?

could involve federal or local agencies, other government bodies, and the private sector.[133] Government actors may have overlapping responsibilities in biodefense efforts, and inadequate planning and coordination can increase the probability of a given risk reaching catastrophic levels. What are the legal barriers to national coordination, such as lack of clear jurisdiction or responsibility (cf. Kvinta, 2011) or lack of harmonized state or local laws? How could they be overcome? How might it look for a centralized body or command structure to take force during a pandemic or other bio-threats? What would be the limits on such a body? Would this look different for different nations, and if so, how? Given existing structures of governance, what approaches could optimally increase coordination in the near- and long-term?

These questions could be addressed through a broader comparative legal analysis, to examine what legal mechanisms for responding to biorisk have been effective in different contexts, and how. What are the characteristics of governments, institutions, and mechanisms that correspond to different outcomes? Do early and effective detection and response correspond to particular decision-making processes, emergency powers, clear structures for coordination, adaptability in an existing regulatory regime, or other factors? How does it vary by the type or scope of the threat?

EXISTING ACADEMIC LITERATURE

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Cotton-Barratt, O., Farquhar, S., Halstead, J., Schubert, S., & Snyder-Beattie, A. (2016). *Global catastrophic risks 2016*. Global Challenges Foundation. https://globalchallenges.org/wp-content/uploads/2019/07/Global-Catastrophic-Risk-Annual-Report-2016.pdf

Farquhar, S., Cotton-Barratt, O., & Snyder-Beattie, A. (2017). Pricing externalities to balance public risks and benefits of research. *Health Security*, *15*(4), 401–408. https://doi.org/10.1089/hs.2016.0118

Farquhar, S., Halstead, J., Cotton-Barratt, O., Schubert, S., Belfield, H., & Snyder-Beattie, A. (2017). *Existential risk: diplomacy and governance*. Global Priorities Project. https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

[133] In the United States several iterations of biodefense strategies accompany a large biodefense budget. The 2018 National Biodefense Strategy is currently in force, preceded by the Obama administration's 2009 National Strategy for Countering Biological Threats and the 2012 National Strategy for Biosurveillance, and the George W. Bush administration's 2004 Homeland Security Presidential Directive-10, which followed the 2001 anthrax attacks. It recognizes the importance of "multi-sectoral cooperation," through engagement and cooperation across all levels of government and partnership with non-governmental organizations and the private sector (p. 4).

Heyman, D., Epstein, G. L., & Moodie, M. (2009, December). *The Global Forum on Biorisks: Toward effective management and governance of biological risks*. Center for Strategic and International Studies. https://fas.org/programs/bio/resource/documents/The%20Global%20Forum%20on%20Biorisks.pdf

Kvinta, B. (2011). Quarantine powers, biodefense, and Andrew Speaker. *Journal of Biosecurity, Biosafety and Biodefense Law*, *1*(1), 1–17. https://doi.org/10.2202/2154-3186.1002

Lai, H-E., Canavan, C., Cameron, L., Moore, S., Danchenko, M., Kuiken, T., Sekeyová, Z., & Freemont, P. S. (2019). Synthetic biology and the United Nations. *Trends in Biotechnology*, *37*(11), 1146–1151. https://doi.org/10.1016/j.tibtech.2019.05.011

Larsen, R., Boddie, C., Watson, M., Gronvall, G. K., Toner, E., Nuzzo, J., Cicero, A., & Inglesby, T. (2015, July). *Jump start: Accelerating government response to a national biological crisis*. Johns Hopkins Center for Health Security. https://www.centerforhealthsecurity.org/our-work/2015%20Jump%20Start/Jump%20Start

National Academies of Sciences, Engineering, and Medicine. (2016b). *Global health risk framework: Pandemic financing: Workshop summary*. The National Academies Press. https://doi.org/10.17226/21855

National Academies of Sciences, Engineering, and Medicine. (2018a). *Biodefense in the age of synthetic biology*. The National Academies Press. https://doi.org/10.17226/24890

Sirleaf, M. (2018a). Ebola does not fall from the sky: Structural violence & international responsibility. *Vanderbilt Journal of Transnational Law*, *51*(2), 477–554.

Sirleaf, M. (2018b). Responsibility for epidemics. *Texas Law Review*, *97*(2), 285–351. https://texaslawreview.org/responsibility-for-epidemics/

National Research Council (2004). *Biotechnology research in an age of terrorism*. The National Academies Press. https://doi.org/10.17226/10827

Taylor, P. (2008) Catastrophes and insurance. In N. Bostrom & M. M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 164–183). Oxford University Press.

Zhang, J., Marris, C., & Rose, N. (2011, May). *The transnational governance of synthetic biology: Scientific uncertainty, cross-borderness and the 'art' of governance*. London: BIOS (Centre for the Study of Bioscience, Biomedicine, Biotechnology and Society). http://openaccess.city.ac.uk/16098/

EXISTING INFORMAL DISCUSSION

Anthony, G., & Neill, S. (2020 Jun. 5). *The International Underwriting Association backs proposals for "Pandemic Re"*. National Law Review. https://www.natlawreview.com/article/international-underwriting-association-backs-proposals-pandemic-re

Bruns, R. (2019). *Finance in a pandemic*. Event 201: A Global Pandemic Exercise. https://www.centerforhealthsecurity.org/event201/event201-resources/finance-fact-sheet-191009.pdf

Meenan, C. (2020 May 19). *The future of pandemic financing: Trigger design and 2020 hindsight*. Centre for Disaster Protection. https://www.disasterprotection.org/latest-news/the-future-of-pandemic-financing-trigger-design-and-2020-hindsight

5.3 Sharing the Benefits of Synthetic Biology

Similar to AI, synthetic biology could create vast potential for advancement and wealth across many industries and groups. Development could be directed and captured by a small set of actors, concentrating wealth and allocating benefits and risks to favor certain populations. If this distribution is suboptimal, humanity could permanently lose great potential (p-risk) or allow great suffering (s-risk) (Section 3.2.1). Therefore, it seems promising to investigate what legal mechanisms could be used to distribute benefits and risks, as well as how they ought to be distributed.

RESEARCH PROJECTS

5.3.1 Steering Research and Development

It may be possible and desirable to shape the direction of research and development to address near- and long-term global priorities. Synthetic biology is well-suited to address other cause areas. Climate change could be mitigated with biofuels, carbon capture, and sustainable production,[134] or global health and development aided through improved access to food,[135] clean water,[136] and healthcare.[137] Given the promise of synthetic biology, suboptimal development could represent permanent loss of great potential, constituting a p-risk. What legal tools could help steer such technological progress? How could intellectual property law, economic development law such as taxes and subsidies (cf. Posner, 2008), trade law, and other legal fields influence development of the synthetic biology market? What can we learn from other industries?

---

134 Climate change issues could be mitigated by carbon capture by bioengineered plants (DeLisi, 2019), biofuels and biorefinery for alternative energy, optimizing carbon conversation or recapturing carbon in synthetic biology processes (François et al., 2020), more sustainable production methods (Le Feuvre & Scrutton, 2018), and engineering crops to withstand climate warming (Quint, et al. 2016).

135 Access to food could be improved with increased yield, nutrition, and sustainability of crops and other agricultural products (Roell & Zurbriggen, 2020; Wurtzel et al., 2019), quality monitoring, processing, and storage (Aguilar et al., 2019; Tyagi et al., 2016).

136 Synthetic biology has broad bioremediation applications, including microbial and plant-based solutions for cleaning up air, water, and soil pollution (Rylott & Bruce, 2020).

137 Rooke (2013).

5.3.2 Access and Benefits-Sharing

As with other advancements, the allocation of potential benefits from synthetic biology could favor wealthier countries by default for at least two reasons: (a) firms are more likely to develop drugs and other products that will principally benefit those who can afford them, and (b) synthetic biology is complex and often capital intensive, meaning investors and workers in already-wealthy countries are more likely to capture the benefits to sellers, including intellectual property (Hollis, 2013). Some mechanisms exist for limited benefits-sharing; notably, the Convention on Biological Diversity and Nagoya Protocol on Access and Benefit-sharing aim, in part, to share benefits arising from genetic resources based on the geographic source, with varied national implementation of provider and user measures[138] (Sirakaya, 2019). However, there is no consensus on whether digital sequence information is within their scope, leading to ongoing discussion (see Ad Hoc Technical Expert Group on Synthetic Biology, 2015, para. 31; Bagley & Rai, 2013; Laird & Wynberg, 2018).[139] DIY bio, open access publishing, and "open source" biology could increase accessibility in low-income areas, but the wealthiest would still have earliest access and lesser risk from inadequate tools or expertise, such as for material storage or quality control (e.g., Foster, 2016). Other proposals and approaches for access and benefits-sharing include differential pricing, voluntary licensing models (Palfrey, 2017), compulsory licenses, payment mechanisms based on health impact (Hollis, 2013; WHO, 2013), allocation based on health access and risk factors,[140] and establishing rights and systems for accountability (Friedman & Gostin, 2015; Gostin & Friedman, 2020).

What institutions or legal instruments could equitably distribute wealth and resources produced by synthetic biology? Would their form vary geographically, at the national and international level, by nation, or by technology, and if so, how? How can they account for future development across all sectors, emergence of new technologies and resources, and means of bypassing such measures (United

---

[138] These provider and user measures enable enforcement of access and benefits-sharing requirements, often formalized in an agreement between the provider and user. Provider measures are established by a source country to ensure that its genetic resources are accessed based on mutually agreed-upon terms and with prior informed consent. User measures ensure that genetic resources are accessed according to these measures, for example through reporting requirements and compliance checkpoints.

[139] Several reports and decisions adopted by the Conference of the Parties to the Convention on Biological Diversity specifically discus synthetic biology, including Report of the Eleventh Meeting (2012 Dec. 5), Decision XII//24 (2014 Oct. 17), Decision XIII/17 (2016 Dec. 16), and Decision 14/19 (2018 Nov. 30).

[140] Most recently this type of framework was developed to plan for equitable vaccine allocation for COVID-19 (NASEM, 2020b, 2020c).

Nations Conference on Trade and Development, 2019 throughout & at p. 20)? What factors should be considered in distribution? How should they address changing circumstances over time? To what extent do DIY bio and open access distribute benefits optimally, weighed against the risks and distribution of risks, and what role might they play in an access and benefits-sharing regime?

5.3.3 Intellectual Property Regime

Intellectual property regimes may be important for synthetic biology, although in different ways than for AI (Section 4.3.5). In synthetic biology, the most relevant type of intellectual property is patents, with others used less frequently. Thus, patent law regimes in particular could help guide research and development toward desirable outcomes—influencing the rate of innovation, research directions, and magnitude and distribution of benefits (König et al., 2015). What intellectual property mechanisms have been used to steer innovation and public access in the past, and what were the consequences? For example, a patent law regime could permit compulsory licenses (Shore, 2020; but see Sirleaf, 2018b, p. 347, footnotes 346–347 and accompanying text), change patent eligibility for specific subject matter, tighten requirements for patentability, change exclusivity periods, or provide non-patent incentives.[141] What other mechanisms are conceivable (Douglas & Stemerding, 2014, Table 5 & pp. 14-15; Miguel Beriain, I. d., 2014)? Could human rights provide a basis for intellectual property law reform (Hale, 2018)? How might the law interact with soft governance and norms, for example around open source biology?

5.3.4 Distribution of Risks

Some risks from synthetic biology may be directed to certain populations or geographical locations, while universal risks may be readily avoided and mitigated locally by those with resources. Synthetic biology could replace the means of livelihood for people in developing countries (Kaebnick et al., 2014) or result in release of genetically engineered organisms that less wealthy countries do not have the resources to protect against (Hollis, 2013). This could have cascading effects, making it a risk factor. Clinical trials and experimental testing present varying and potentially great risks to humans and the environment, giving rise to questions of

[141] For example, the United States Orphan Drug Act of 1983 promotes development of treatments for rare diseases by offering incentives such as extended market exclusivity, reduced fees, and substantial tax credits for research and development. Others have adopted similar legislation, including Japan in 1993 and the European Union in 2000.

protection, informed consent, liability, and compensation.[142] What institutions or legal frameworks could equitably spread the distribution of risk? To what extent could and should they involve allocation of resources to protect against risk or liability and compensation schemes? Would their form vary geographically, or at the national and international level, and if so, how? What ethical criteria should research and clinical trials meet, and how can that change with circumstances? What questions should be answered in deciding whether to have human challenge trials, or other potentially great risks, during an emergency? Should requirements for informed consent (Kuiken, 2020, pp. 286–287; Sommers, 2020), compensation, and liability change during an emergency, and if so, how? Are there other great benefits or risks or extenuating circumstances that may warrant a different framework?

5.3.5 Human Enhancement and Beings Other than Humans

How should the law handle beings other than those we know today? With advancements in synthetic biology may come human enhancement beyond our limits today (Al-Rodhan, 2020; Gaspar et al., 2019; Masci, 2016), synthetic organisms with sentience, and animals that have been modified to have more human characteristics or contain human tissue, including brain tissue in the case of human-animal neurological chimeras (Crane et al., 2019; Kwisda et al., 2020; NASEM, 2020; Porsdam Mann et al., 2019). What can we learn from existing and proposed frameworks for legal personhood, citizenship, and rights and duties of humans and non-humans (Kurki, 2017)? Are these frameworks adequate for addressing potential ethical, legal, and societal issues that could arise with modified or synthetic beings (Emanuel et al., 2019, p.12–14; Wittes & Chong, 2014)? If not, what new or adapted framework could address these possibilities? What are the downstream legal and ethical implications of such a framework?

Given the vast potential of synthetic biology to positively (or negatively) shape the far future, how can the law consider animals and beings other than humans in

[142] Testing of particular concern may include (a) population testing, which presents a great burden in obtaining the informed consent of all potential participants and may not be as effective if the population is aware of being studied (DuBois, 2011; LaFreniere, 2019; Sutton, 2005) and (b) human challenge trials, in which participants are intentionally challenged with an infectious disease organism, for diseases that have high levels of morbidity and/or are poorly understood (Kolber, 2020). For a discussion of liability and compensation plans in the United States and possible alternatives, see Chapman et al., 2020 and Thomas, 2011. The World Health Organization (WHO), Expert Committee on Biological Standardization has published reports on regulatory considerations for human challenge trials (WHO, 2016; WHO, 2017), and the Working Group for Guidance on Human Challenge Studies in COVID-19 has published key criteria for ethical acceptability of such trials for COVID-19 (WHO, 2020).

distributing the benefits and risks it entails? Measures to prevent, detect, and respond to risk are attuned to humanity, while failing to address the welfare of vast numbers of animals. This oversight allows suffering and existential risks for non-human species.[143] How could legal mechanisms or proposals from other research questions in this Section be adapted to address these risks? Is an entirely separate institute or legal instrument warranted?

What can we learn from the broader discussions on non-human sentience in animal law (Section 9.2), artificial intelligence (Section 4.2.2), extraterrestrial intelligence (Section 8.2.3), sentience-sensitive institutions (Section 6.1.10), and moral circle expansion in judicial decision-making (Section 6.2.4)?

EXISTING ACADEMIC LITERATURE

Badley, M. A., & Rai, A. K. (2013). *The Nagoya Protocol and synthetic biology research: A look at the potential impacts*. Woodrow Wilson International Center for Scholars. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=5916&context=faculty_scholarship

Douglas, C. M., & Stemerding, D. (2014). Challenges for the European governance of synthetic biology for human health. *Life Sciences, Society and Policy*, *10*(6), 1–18. https://doi.org/10.1186/s40504-014-0006-7

Emanuel, P., Walper, S., DiEuliis, D., Klein, N., Petro, J. B., & Giordano, J. (2019, October). *CCDC CBC-TR-1599, Cyborg Soldier 2050: Human/Machine Fusion and the Implications for the Future of the DOD* (CCDC CBC-TR-1599). U.S. Army Combat Capabilities Development Command, Chemical Biological Center. https://community.apan.org/wg/tradoc-g2/mad-scientist/m/articles-of-interest/300458

Friedman, E. A., & Gostin, L. O. (2015). Imagining global health with justice: In defense of the right to health. *Health Care Analysis*, *23*(4), 308–329. https://doi.org/10.1007/s10728-015-0307-x

Gaspar, R., Rohde, P., & Giger, J. (2019). Unconventional settings and uses of human enhancement technologies: A non-systematic review of public and experts' views on self-enhancement and DIY biology/biohacking risks. *Human Behavior and Emerging Technologies*, *1*(4), 295–305. https://doi.org/10.1002/hbe2.175

Gostin, L. O., & Friedman, E. A. (2020). Imagining global health with justice: Transformative ideas for health and well-being while leaving no one behind. *Georgetown Law Journal*, *108*(6), 1535–1606. https://www.law.georgetown.edu/georgetown-law-journal/wp-

[143] Measures to prevent, detect, and respond to potential pandemics and other existential risks for animals may also have benefits for humanity, although it is uncertain whether their absence constitutes a risk factor. For example, if a virus is transmissible between humans and animals, the ability to detect it in animals and respond could limit its spread before it reaches humans, or it could prevent a human virus from mutating in animals and later causing a repeat outbreak among humans. For informal discussion, see McKenna, 2020, Briggs, 2020, and Calma, 2020 ("Animal health and human health are 'tightly interconnected'").

content/uploads/sites/26/2020/06/Gostin-Friedman_Imagining-Global-Health-with-Justice-Transformative-Ideas-for-Health-and-Well-Being-While-Leaving-No-One-Behind.pdf

Hale, Z. A. (2018). *Patently unfair: The tensions between human rights and intellectual property protection*. The Arkansas Journal of Social Change and Public Service. https://ualr.edu/socialchange/2018/04/04/patently-unfair/

Hollis, A. (2013). Synthetic biology: Ensuring the greatest global value. *Systems and Synthetic Biology*, *7*, 99–105. https://doi.org/10.1007/s11693-013-9115-5

Kaebnick, G. E., Gusmano, M. K., & Murray, T. H. (2014). The ethics of synthetic biology: Next steps and prior questions. *Synthetic Future*, *44*(S5), S4–S26. https://doi.org/10.1002/hast.392

König, H., Dorado-Morales, P., & Porcar, M. (2015). Responsibility and intellectual property in synthetic biology: A proposal for using Responsible Research and Innovation as a basic framework for intellectual property decisions in synthetic biology. *EMBO reports*, *16*(9), 1055–1059. https://doi.org/10.15252/embr.201541048

Kuiken, T. (2020) Biology without borders: Need for collective governance? In B. D. Trump, C. L. Cummings, J. Kuzma, & I. Linkov (Eds.), *Synthetic biology 2020: Frontiers in risk analysis and governance* (pp. 269–295). Springer, Cham. https://dx.doi.org/10.1007/978-3-030-27264-7_12

Kurki, V. (2017). Why things can hold rights: Reconceptualizing the legal person. In V. Kurki, & T. Pietrzykowski (Eds.), *Legal personhood: Animals, artificial intelligence and the unborn* (pp. 69–89). Springer, Cham. https://doi.org/10.1007/978-3-319-53462-6

Kuzma, J., & Rawls, L. (2016). Engineering in the wild: Gene drives and intergenerational equity, *Jurimetrics Journal*, *56*, 279–296. https://research.ncsu.edu/ges/files/2014/02/engineering_the_wild.authcheckdam.pdf

Kwisda, K., White, L., & Hübner, D. (2020). Ethical arguments concerning human-animal chimera research: A systematic review. *BMC Medical Ethics*, *21*. https://dx.doi.org/10.1186/s12910-020-00465-7

Laird S. A., & Wynberg R. P. (2018, January 10). *A fact finding and scoping study on digital sequence information on genetic resources in the context of the convention on biological diversity and Nagoya Protocol* (CBD/DSI/AHTEG/2018/1/3). Convention on Biological Diversity, Ad Hoc Technical Expert Group on Synthetic Biology. https://www.cbd.int/doc/c/e95a/4ddd/4baea2ec772be28edcd10358/dsi-ahteg-2018-01-03-en.pdf

Miguel Beriain, I. d. (2014). Synthetic biology and IP rights: Looking for an adequate balance between private ownership and public interest. In J. Boldt (Ed.), *Synthetic biology: Metaphors, worldviews, ethics, and law*. Springer VS. http://doi.org/10.1007/978-3-658-10988-2

Nielsen, M. E. J., Kongsholm, N. C. H., & Schovsbo, J. (2019). Property and human genetic information. *Journal of Community Genetics*, *10*, 95–107. https://doi.org/10.1007/s12687-018-0366-4

Palfrey, Q. A. (2017). Expanding access to medicines and promoting innovation: A practical approach. Georgetown Journal on Poverty Law and Policy, 24(2), 161–203.

Porsdam Mann, S., Sun, R., & Hermerén, G. (2019) A framework for the ethical assessment of chimeric animal research involving human neural tissue. *BMC Medical Ethics*, *20*. https://doi.org/10.1186/s12910-019-0345-2

Posner, R. A. (2008) Public policy towards catastrophe. In N. Bostrom & M. M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 184–201). Oxford University Press.

Sirakaya, A. (2019). Balanced options for access and benefit-sharing: Stakeholder insights on provider country legislation. *Frontiers in Plant Science*, *10*, 1175. https://doi.org/10.3389/fpls.2019.01175

Sirleaf, M. (2018b). Responsibility for epidemics. *Texas Law Review*, *97*(2), 285–351. https://texaslawreview.org/responsibility-for-epidemics/

Sommers, R. (2020). Commonsense consent. Yale Law Journal, 129(8), 2232–2324. https://www.yalelawjournal.org/article/commonsense-consent

United Nations Conference on Trade and Development (2019). *Synthetic biology and its potential implications for biotrade and access and benefit-sharing* (UNCTAD/DITC/TED/INF/2019/12). https://unctad.org/system/files/official-document/ditctedinf2019d12_en.pdf

Wittes, B., & Chong, J. (2014, September). *Our cyborg future: Law and policy implications*. Brookings Institution. https://www.brookings.edu/research/our-cyborg-future-law-and-policy-implications/

World Health Organization, World Intellectual Property Organization, & World Trade Organization (2013). *Promoting access to medical technologies and innovation: Intersections between public health, intellectual property and trade*. https://www.who.int/phi/promoting_access_medical_innovation/en/

EXISTING INFORMAL DISCUSSION

Al-Rodhan, N. (2020, June 29). *A neurophilosophy of two technological game-changers: Synthetic biology & superintelligence*. Blog of the American Philosophical Association. https://blog.apaonline.org/2020/06/29/a-neurophilosophy-of-two-technological-game-changers-synthetic-biology-superintelligence/

Cotton-Barratt, C. (2014, October 1). *Effective policy? Requiring liability insurance for dual-use research* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/zvRerivrWdZ5J5rD9/effective-policy-requiring-liability-insurance-for-dual-use

Masci, D. (2016, July 26). *Human enhancement: The scientific and ethical dimensions of striving for perfection*. Pew Research Center. https://www.pewresearch.org/science/2016/07/26/human-enhancement-the-scientific-and-ethical-dimensions-of-striving-for-perfection/

National Academies of Sciences, Engineering, and Medicine (2020). *Ethical, legal, and regulatory issues associated with neural chimeras and organoids*. https://www.nationalacademies.org/our-work/ethical-legal-and-regulatory-issues-associated-with-neural-chimeras-and-organoids