Written Evidence Submitted by
Nick Bostrom, Haydn Belfield and Sam Hilton, University of Oxford's Future of Humanity Institute and University of Cambridge's Centre for the Study of Existential Risk
(RFA0061)

1. Executive Summary

1.1 We are researchers at the University of Oxford's Future of Humanity Institute and the University of Cambridge's Centre for the Study of Existential Risk. This submission draws especially on our expertise on low-probability, highly destructive risks and the development of R&D funding agendas to address these risks.

1.2 In this response we particularly focus on the following three of the Committee's questions:

- What should the focus be of the new research funding agency and how should it be structured?
- What funding should ARPA receive, and how should it distribute this funding to maximise effectiveness?
- What can be learned from ARPA equivalents in other countries?

1.3 We make the following recommendations:

- Focus areas
    - Biological risks, risks from AI, and forecasting risks should be focus areas
- Structure and culture
    - Establish a mechanism to manage Ethical, Legal and Societal Issues (ELSI)
    - Have long-term funding commitments and political support
    - Focus on transformation
    - Keep as flat a hierarchy as possible
    - Have narrow projects with specific aims - aim for defined outcomes and practical applications
    - Have a commitment to seeing successful projects through to technology transfer, with well-established and well-funded pipelines to do so
    - Give expert PMs wide latitude to take risks when working on a programme
    - Scientific, Engineering and Technical Assistance (SETA) staff should not be term-limited
- Mechanisms to Identify and Start Projects
    - Conduct 'pre-mortems' - including the question 'might you regret developing this technology?'
    - Have Project Managers (PMs) do a 'New Start Pitch' for new program ideas
    - Consider adopting criteria akin to the Heilmeier Catechisms when choosing programmes, in order to maximise their value

- Testing and Evaluation
    - Consider using independent third-parties to perform evaluations and having a Chief of Testing & Evaluation
    - Empower the Agency Director to quickly end programmes that are not succeeding
- Recommended tools
    - Use prize-challenges to drive innovation
    - Set up civilian 'mini-Jasons'
- Hiring
    - When making hiring decisions, recommended general requirements are expertise, good strategic judgement about focus, open-mindedness and the ability to get things done.
    - When interviewing, have prospective PMs present their project proposal and defend it
    - Engage with experts, particularly for the hiring process

1.4 We go into more detail on these recommendations below, and would be happy to discuss them with the Committee.

2. Suggested Focus Areas for UK ARPA

High-impact risk research and global leadership in emerging technologies

2.1 The government's new Research and Development Roadmap highlights that UK ARPA "will target areas where the UK can gain competitive advantage and lead the world in the creation of new technologies". But in the wake of COVID-19, there will also be an unusually tractable window for focusing research resources on highly destructive risks, especially those from emerging technologies. Including high-impact risk research in UK ARPA's mandate would be a double win. It would transform the UK's ability to deal with extreme risks, whilst further positioning the UK as a responsible pioneer in world-leading technologies.

2.2 The UK is a world-leader in biotechnology and AI. In these important but still emerging sectors, some projects will succeed wildly and many will not work out. Academic researchers, start-ups and large company R&D teams are all making innovative, exploratory - and therefore necessarily risky - bets on new technologies. These sectors are thus an excellent match for a high-risk/high-reward organisation such as UK ARPA. UK ARPA can support these industries, and also make sure that this innovation promotes, and does not harm, UK national security.

2.3 COVID-19 has brutally demonstrated how underprepared the UK is for low-probability, highly destructive risks. We need persistent innovation to better predict and prepare for these risks, which are amongst the most dangerous we face as a nation. The risks we face are novel, and some of the mitigation approaches we take will not work - this is again an excellent fit for UK ARPA.

2.4 Recommended focus areas include research into biosecurity risks, artificial intelligence risks, and better forecasting methods.

Biosecurity

2.5 Our current suite of interventions to a novel biological threat can either be rapidly deployed (e.g. non-pharmaceutical interventions) or can be highly effective (e.g. vaccines), but not both. Innovative technologies both now and in the future can help close this gap, and should be urgently prioritised for development.

2.6 The UK has a world-leading biotechnology sector, and UK ARPA could contribute significantly to technological development in this area. This could include authoring a roadmap for 21st century biosecurity, scoping out new biodefense applications of horizon technologies, and outlining how opportunities can be brought to technical readiness (e.g. the use of metagenomic sequencing to pathogen blind diagnostics and comprehensive environmental biosurveillance).

2.7 Specific research ideas are included below:

- Environmental pathogen biosurveillance:

    - Programmes to bring this to fruition could focus on Reagent-free diagnostics, perhaps with a specific focus on PCR, DNA sequencing, and CRISPR diagnostics;
    - Next-Generation Metagenomics (pathogen agnostic infectious disease metagenomics);
    - More robust forms of sequencing that can handle complex analytes;
    - Development and miniaturisation of sample acquisition and preparation technologies.
- Safe synthetic biology (e.g. programmes with goals similar to DARPA's Safe Genes);
- Non-pharmaceutical medical countermeasure R&D for large or fast-moving pandemics;
- Non-Invasive pathogen agnostic infection detection methods;
- Rapid scale-up of therapeutics, such as monoclonal antibodies and small-molecule antivirals;
- Safe Bio Spaces - with a focus on BSL-4 security labs to make them safer;
- Technologies for securing physical spaces and preventing transmission (e.g. safe planes and trains);
- Developing better tools for DNA synthesis screening (the technical aims should include accuracy under 200 base pairs and prediction of sequence/pathogen from oligonucleotides, with the stretch goal being prediction of pathogenicity in novel pathogens from sequence data alone);
- Systems engineering to better integrate processes for continual biosecurity; and
- Innovative and improved face masks, ventilation / air filtration, UV sterilization, food sterilization, etc.

Artificial intelligence

2.8 The UK has a world-leading AI sector, and UK ARPA could contribute significantly to technological development in this area. Most AI-related research and training should have a broad scope (since AI will affect so many processes and domains that a breadth of expertise will be essential).

2.9 But specific AI research ideas which may be suitable for UK ARPA are included below:

- AI for Good, or more narrowly AI for the SDGs: developing AI systems that serve a societal goal, for example better weather prediction, better grid resilience, better crop disease diagnosis, etc.
- Resilience to adversarial examples - inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake;
- Methods for Test & Evaluation (T&E), Verification & Validation (V&V), and assurance of deep learning systems (further details immediately below);
- Formal verification research (further details immediately below);
- Forecasting the progress and consequences of AI;
- Countermeasures to malicious uses of AI (e.g. research that helps identify videos as deep fakes);
- AI ethics (fairness, accountability, explainability, transparency and bias); and

- Ensuring security of hardware supply (i.e. secure access to capable AI chips to support many applications);
- More efficient and secure hardware design.

Fund research of methods for Test & Evaluation, Verification & Validation, and assurance of deep learning systems

2.10 Deep learning methods are rapidly providing new capabilities across a range of sectors, and the UK is a leading contributor to those advances, in academia and business, for example at Google DeepMind in London.

2.11 However it is often the case that deep learning methods cannot be deployed in safety-critical applications and defense applications for which high assurance of correct system behaviour is required. This is because traditional T&E and V&V methods for gaining such high assurance typically cannot be applied to deep learning systems.

2.12 More research funding is needed to remedy this, before deep learning methods are deployed in these application areas. An equivalent project in the United States is DARPA XAI.

Formal Verification research

2.13 Formal Verification (FV) research involves the use of mathematical methods to offer formal proofs that a system will operate as intended. These will be important for deploying these systems in safety-critical application areas. There is likely to be a gap in converting formal verification research into practice.

Forecasting

2.14 Research into:

- Improving the accuracy of long-term forecasts;
- Improving forecasting techniques, for example using quantified falsifiable predictions or creating large and effective public prediction markets;
- Full inference cycle tournaments, as proposed by Philip Tetlock.

2.15 In terms of current good practice, the Office of Budget Responsibility produces publicly available fiscal and economic forecasts, and annual reviews (in a report to Parliament) of how well their forecasts matched reality and how they can improve. Such techniques could be used to improve how the UK predicts the probabilities of future disasters; this could in part be similar to the work done in the US by IARPA's intelligence community prediction market.

Other possible areas for ARPA

2.16 Research into:

- Structured transparency - developing assorted surveillance and security technologies combined with consideration on systems of deployment that would consider factors such including privacy, ethics, cost, and robustness to adversaries;
- Cognitive enhancement research and development, including both biological and non-biological;
- Prediction and mitigation strategies for other neglected low probability but potentially highly destructive risks, such as:
    - Nuclear winter;
    - Extreme climate change scenarios,
    - Asteroid impacts;
    - Supervolcanoes;
    - Solar flares.

3. The Structure and Culture of UK ARPA

Establish a mechanism to manage Ethical, Legal and Societal Issues (ELSI)

3.1 This mechanism should address the full implications of novel research on society. DARPA's Living Foundries programme has an ELSI advisory board. Whilst there are several possible types of mechanism (either internal or external, formal or informal), UK ARPA's chosen mechanism should:

- Review all incoming programme proposals from programme managers and flag potential areas of concern in advance;
- Track research as it is conducted and flag emerging issues;
- Assess how results should be released and published;
- Assess potential applications of research;
- Engage openly with individual programme managers and, where possible, be able to discuss concerns with chosen external stakeholders.

3.2 Lessons should be learned from the 'freewheeling' 1960s early-ARPA institutions, and in particular the harms that stem from Programme Officers acting unilaterally (for example creating chemical defoliants like Agent Orange at the beginning of the Vietnam War (Stellman, 2018)). In the context of AI safety and biosecurity, these are highly serious considerations.

Have long-term funding commitments and political support

3.3 UK ARPA should be an independent organisation with the ability to set its own goals within its mandates. A commitment to fund, or having funding allocated, for at least 10 years would give UK ARPA a chance to have some of its technology bets mature to the point where an assessment could be made of their success.

3.4 UK ARPA should not feel short-term pressures to deliver. If it does, it may be incentivised to take on less ambitious projects that do not have a high return on investment.

ARPAs in the US run on huge budgets - this allows the inevitable project failures to be absorbed by project successes.

Focus on transformation

3.5 UK ARPA should focus on transformative technologies, rather than incremental and near-to-market innovation.

Keep as flat a hierarchy as possible

3.6 This will allow UK ARPA to attract the world's top talent and encourage programme managers (PMs) to be relentlessly mission-focused, with as few distractions and interference as possible.

Have narrow projects with specific aims - aim for defined outcomes and practical applications

3.7 This type of research can be classified as 'use-inspired basic research', as per Pasteur's Quadrant (Stokes, 1996). As opposed to pure basic research or pure applied research, use-inspired basic research is at the intersection between fundamental understanding of scientific problems, and immediate use-value for society.

Have a commitment to seeing successful projects through to technology transfer, with well-established and well-funded pipelines to do so

3.8 For example, the US Department of Defense provides DARPA with pipelines to technology transfer. In the early stages, UK ARPA should work to identify these pipelines. It should consider hiring someone dedicated to ensuring effective technology transfer by making sure research is able to be applied and adopted by its secondary and ultimate users. Without this, optimisation problems can arise (e.g. researchers may be incentivised to optimise for short-term success and forget long-term applicability and generalisability).

3.9 The Advanced Research Projects Agency-Energy (ARPA-E) advances high-impact energy technologies to transition them to commercial viability. Like ARPA-E, UK-ARPA is likely to have several different customers and to be more resource-constrained than DARPA. As such, a UK-ARPA might want to consider transitioning technology at a higher Technology Readiness Level to other organisations, as well as having partnerships with other organisations (e.g. in US, Canada, Australia, New Zealand).

Give expert PMs wide latitude to take risks when working on a programme

3.10 We recommend following the example of DARPA by:

- Making PMs singularly responsible for funding decisions;
- Cycling them out in three to five years, unless they are promoted;
- Making promotion difficult (to attract people from academia or industry who come into government for a few years to make a breakthrough, and then move on);
- Giving PMs authority to start new programmes quickly and a high degree of autonomy once their programme has been approved;

- Having highly flexible contracting and hiring capabilities (other ARPAs have used gamification, prize-challenges, and Other Transaction Agreements to pay teams of experts from outside the public sector).
- Setting term limits for programme managers (PMs): These could start on a two year contract which can be renewed, meaning that their average tenure is four to six years. This allows churn, and prevents stagnation of ideas.

Scientific, Engineering and Technical Assistance (SETA) staff should not be term-limited

3.11 They are the institutional memory of an ARPA and provide historical context to PMs.

4. Mechanisms to Identify and Start Projects

Before projects get started

4.1 Conduct a ‘pre-mortem’ - ask the question: ‘might you regret developing this technology - for instance, is it worse for this technology to be misused by bad actors, or in the hands of competitors, than it is beneficial to be in your own hands?’ It is also worth considering whether others, such as other intelligence agencies, could misinterpret the intent behind a project.

4.2 Solicit written feedback from all senior staff, outside advisors, and transition partners before starting a program.

Before a project is pitched

4.3 Peers should ‘red team’ concerns to pressure test proposals before they get pitched to seniors. It is helpful to include ‘transition partners’ in this: the team who would on-board the technology.

During a pitch

4.4 IARPA ensure that Project Managers (PMs) do a ‘New Start Pitch’ for new program ideas:

- These are about three hours long for existing PM members of staff (two hours of presentation with frequent interruptions, followed by an hour of detailed discussion). They tend to be similar in length and scrutiny to a dissertation proposal.
- For prospective PMs who are not on staff, these pitches tend to be much shorter, so as not to put off good candidates with demanding existing jobs - they tend to present about 20 slides and take questions.
- Questions tend to focus on the Heilmeier Catechisms (see below);
- They typically take a few months of full-time work to prepare for;
- PMs are usually provided with support from agency staff to help with background research;

- The assessment of the pitch includes a security assessment of the idea, as well as a civil liberties and privacy protection assessment; and
- It also includes a risk assessment (see Danzig, p22, for a great list of risk assessment questions used by IARPA).

4.5 Consider adopting criteria akin to the Heilmeier Catechisms when choosing programmes, in order to maximise their value. Heilmeier, a former DARPA Director, crafted a set of questions to help Agency officials think through and evaluate proposed research programmes:

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares? If you are successful, what difference will it make?
- What are the risks?
- How much will it cost?
- How long will it take?
- What are the midterm and final 'exams' to verify success?

Before getting started

4.6 One US expert recommended sending notes to US ARPA and equivalent communities saying "we are planning a project on X - is any reason we shouldn't do this?" This may prevent duplication of work, unknown risks etc.

4.7 If it looks like you plan to go forward with a particular project, then in most circumstances it will be worth releasing a broad public announcement that a project will commence.

5. Testing and evaluation (including discontinuing projects)

Ensure that Testing and Evaluation is a central part of UK ARPA

5.1 Typically, 25% of the programme budget should be spent on measuring success (Bonvillian, 2019, p.442). When using prize-challenges (see below) IARPA spends around 50% of the programme budget on evaluation.

5.2 A testing plan should be a core part of a program pitch put together by PMs.

5.3 IARPA uses independent third-parties to perform evaluations. These are usually government labs, Federally Funded Research and Development Centers (FFRDCs), or University Affiliated Research Center (UARCs).

5.4 IARPA has a Chief of Testing & Evaluation, with expertise in experimental design and statistical inference.

Empower the Agency Director to quickly end programmes that are not succeeding

5.5 This will require a competitive programme structure, in which multiple teams are funded in parallel. IARPA predominantly uses a tournament model – funding teams in parallel that pursue a common set of objectives, whether through grants, contracts, or prize-challenges.

5.6 Every six months there should be a full Program Management Review (PMR) to decide whether to continue or discontinue the program. PMRs should cover technical results, program management, and finances. They should follow a similar structure to a New Start Pitch, with the same stakeholders engaged.

5.7 Failures that are due to technical ambition should be celebrated; a significant percentage of programs should fail otherwise the problems being worked on are too easy (IARPA aims for about 50% technical failure rate).

5.8 It is important that PMs feel motivated to be honest about about how a program is going, as the effectiveness of UK ARPA relies on lots of projects being cut.

6. Recommended tools for UK ARPA

Use prize-challenges to drive innovation

6.1 By offering a reward upon completion of a specific objective, prize-challenges would enable UK ARPA to be more cost effective. There are several reasons for this:

- UK ARPA would only commit a large budget once an idea had proven some success;
- Prize challenges are about 10 times more cost-effective than traditional projects. There are much lower overhead costs compared to the admin involved in contracts and procurements;
- They tend to prompt a large degree of spending on research by the contestants - either because competitors overestimate the probability of winning, or because they place a significant value on the reputational reward for winning or being shortlisted. This is especially the case if a project is obviously beneficial to society, with which teams want to be associated.

6.2 Prize-challenges also increase the number of minds tackling a particular problem without having to predict which team or approach is most likely to succeed. Similarly, they are a means of identifying talented individuals and teams, who can be seconded for future programs.

6.3 A prize-challenge should not be an end in itself, but one means within a broader strategy for spurring change. When considering prizes, agencies should select the type of prize best able to accomplish the broader aim.

6.4 For example, the US National Cancer Institute (NCI) launched the Breast Cancer Startup Challenge (BCSC), aiming to accelerate the development and commercialisation of emerging breast cancer technology. Because of long, complex developmental timelines associated with biomedical technologies, the agency needed to find new ways to innovate to achieve its mission. Creating BCSC addressed that: it provided an additional avenue for finding partners, and advanced development and commercialisation of nine breast-cancer-related technologies. A UK example is the Longitude Prize. For more information, other examples, and advice on how to incorporate prize-challenges into government work, see this US government toolkit.

Set up civilian 'mini-Jasons'

6.5 In the US, JASON is a Pentagon-supported independent group of top-rank academic scientists. They spend over a month together during the summer, with other meetings during the year, where they work together on addressing big innovation challenges.

6.6 JASON is funded by the Defence Department and most of its work is military-focused, but UK mini-JASONs could focus on civilian areas such as green (e.g. carbon capture and storage) technology.

6.7 This model allows sufficient time for scientists to actually work through problems together rather than just exchange ideas (which is all that is possible in most normal-length meetings). The challenge is to assemble a group of really top-ranked scientists who enjoy cross-disciplinary discourse and brainstorming ideas.

6.8 JASON chooses its own new members, and many have remained active for several decades. As with JASON, the success of mini-JASONs are dependent on the members committing a substantial amount of time together, and working on the kind of problems that play to their strengths.

7. Hiring for UK ARPA

When making hiring decisions

7.1 Recommended general requirements are:

- Entrepreneurial creativity, to a degree that is rare relative to existing staff in that domain (anecdotally, we have heard stories of ARPA-type agencies traditionally paying too much attention to PMs with academic credentials, and not enough on entrepreneurs who tend to have more experience in and comfort with dropping ideas that are not working);
- Have a high tolerance for unconventional personalities and ways of thinking;
- Ideally, attracted to long-term thinking;
- Total expertise in the field (getting PMs who have experience being the ultimate customer of their proposed research area increases the chance the research will be use-inspired and applicable);
- Good strategic judgement about what to focus on;

- Being open-minded, and able to confidently bring new organisational processes and structures to Whitehall; and
- The ability to get things done.

When interviewing, have prospective PMs present their project proposal and defend it

7.2 See section on 'Mechanisms to Identify and Start Projects' above, including Heilmeier Catechisms.

Engage with experts, particularly for the hiring process

7.3 In particular, we would recommend you speak with David Spiegelhalter at the Royal Society, Erica Fuchs at Carnegie Mellon University and William B. Bonvillian at MIT.

8. Further Reading for UK ARPA

8.1 Books on DARPA:

- Bonvillian, DARPA Model for Transformative Technologies - PDF version is free
- Danzig, Technology Roulette - the IARPA questions for pitches are on page 22
- The Pentagon's Brain: An Uncensored History of DARPA
- The DARPA Model for Transformative Technologies: Perspectives on the US Defense Advanced Research Projects Agency

8.2 Articles on DARPA:

- What makes DARPA tick?
- What is DARPA? How to Design Successful Technology Disruption
- Defense Advanced Research Projects Agency: Overview and Issues for Congress
- Rethinking the Role of the State in Technology Development: DARPA and the Case for Embedded Network Governance
- Funding Breakthrough Research: Promises and Challenges of the "ARPA Model"
- "Special Forces" Innovation: How DARPA Attacks Problems

8.3 Articles on ARPA clones:

- An Assessment of ARPA-E
- ARPA-E is Here to Stay
- Cloning DARPA Successfully

8.4 UK-specific articles:

- Visions of ARPA: Embracing Risk, Transforming Technology
- DARPA 'lookalikes' must ground their dreams in reality
- Dominic Cummings got his British Darpa. Can he make it work?

References

1) Bonvillian, William. (2019). ‘The DARPA Model for Transformative Technologies: Perspectives on the U.S. Defense Advanced Research Projects Agency’. Open Book Publishers. Available at: https://www.openbookpublishers.com/product/1079
2) Stellman, Jeanne. (2018). ‘The extent and patterns of usage of Agent Orange and other herbicides in Vietnam’. [Article]. Available at: https://assets.aspeninstitute.org/content/uploads/2018/03/Stellman-Nature-2003-Extent-patterns-of-usage-of-AO-in-Vietnam.pdf
3) Stokes, Donald. (1996). ‘Pasteur's Quadrant: Basic Science and Technological Innovation’. Brookings Institution Press

(July 2020)