St George's House
AI – Threats and Opportunities Consultation
(19-20 October 2023)

Final Report

Four issues were examined during this St George's House consultation on AI threats and opportunities, in four working groups:

- Misinformation
- Social disruption
- Dangerous activities
- Lawmaking and regulatory functions

Consciously omitted from the framework for discussion were 'existential risk' and 'bias' (although it was acknowledged that bias was relevant to topics discussed).

Because of the complex definitional issues around AI (see below), this report uses the term 'artificial-intelligence technologies' (AITs) to describe practical applications of AI.

Common Themes

Across the four groups, broad themes repeatedly emerged:

We don't know what we don't know
Seeking understanding of the issue under examination, we find that **we don't know what we don't know**. Observing this allows us more effectively to pose better questions and actions.

Things that we believe we don't know (after looking) include:

- Relevant existing and emerging efforts to address the four areas above
- Unintended outcomes of efforts to use or constrain use of AITs
- The full potential of AITs to avoid measures taken to control use of AITs
- The full spectrum and variety of uses and associated risks for AITs, and also the coincidences or overlaps of different dimensions of risk in using them
- The extent to which AITs now are training on datasets themselves containing outputs of AITs

Definitional issues and glossing over differences are part of the problem
The generalising term 'AI' kneads into a lump multifarious developments, applications, and effects of different data technologies. For productive discussion leading to effective action, we must clarify 'AI' in terms of specific situations, technologies and risks so as to be able to distinguish individual problems susceptible to solution.

This 'human aspirations' undercurrent in discussion of AITs (see below) was directed by largely unspoken definitions for assumed shared ideas underlying such terms as 'right hands', 'destructive', 'well-intentioned', 'ill-will', 'mis-use', etc. Various key terms requiring contextualisation in terms of clearly stated human, social goals or intentions (*cui bono*?) might include:

- Dangerous activities
- Human rights
- Intelligence
- True, false, good, evil, other morals-based vocabulary

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager

Other areas contributing to the multifariousness of AITs-related issues include:

- The lack of inherent intention in technology as a tool – 'dual-use', 'agnostic', 'ambiguous', etc.; the lack of any absolute good or bad in any aspect of AITs
- The scattered, uncoordinated nature of research, use, tracking, and regulation of AITs worldwide
- International differences in definitions, concerns, and attempts to constrain or promote AITs
- Contextualising factors, such as the particular semantic environment within which AITs operate, or the intention/goal for using AITs

Act now, not later

While acknowledging that short-term-risk proponents do not see eye-to-eye with long-term-risk proponents, we do not have to agree on the level of risk to agree that there *is* risk in the here and now.

We readily perceive immediate risk of significant disruption and damage to social, economic and political order resulting from relatively simple applications of currently available AITs. Conspicuously, we foresee abuse of AITs in the run-up to the UK elections in 2024.

We warn strongly against planning only to mitigate long-term risks – and/or arguing that there is no risk at all – from human use of AITs.

As with any call to action for public good (such as those suggested below), necessary resources include planning, organisation, honesty and transparency, prioritisation of people over profit, political will, corporate responsibility, leadership, skills, education, and *funding*; as always, these are patchy in reality, but the patchiness cannot be a reason not to act.

Evaluating 'AI' and AITs: what (by implicit contrast) is 'human'?

Human mis-use of AITs constitutes a greater existential risk to us all than the possible emergence of an out-of-control, independent machine intelligence.

Because AITs can be used to accelerate the gratification and amplify the effect of human urges, it is important to understand the consequences of the (rapid) realisation at scale of such human urges. A crudely pessimistic statement of the problem could be that as we already cannot trust, predict nor understand human beings in their irrationality and bias, so we cannot trust, predict nor understand (a) outputs of artificial intelligences trained on human-centric data or (b) the use of artificial intelligences to realise human aims.

More hopefully, granted that the majority of people behave in a broadly well-intentioned manner towards others (and that dedicated minorities govern social goods such as Wikipedia as well as, for example, social ills such as aggressive subcommunities using bots to provoke widespread reactive prejudices), AITs in the right hands can serve:

(a) to identify and forestall socially destructive acts of disinformation, violence, ill-will etc and
(b) to help protect and advance positive efforts for the broader well-being of the majority.

As human beings we differ from AITs in several respects, which we should examine in terms of AITs' value to us and ours, and find ways to keep, if valuable:

- Human dignity – we know this when we see it; even if it is hard to define, since ancient times it has been connected to perceiving oneself as making a valued contribution through work

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager

- An ability to contextualise in perceiving, interpreting, and evaluating data-sets (e.g. object and concept differentiation)
- An ability to perceive and interpret causes and consequences in terms of experiences such as mercy, forgiveness, error, a sense of responsibility, a sense of justice, desire for approval or attention or admiration, being valued for one's contributions, sense of belonging to a larger whole, pleasure, pain, self-control, self-awareness, emotional response, altruism, etc, that are governed by biologically and socially founded survival instincts

Possibly, AITs could think more like humans by integrating a variety of different learning (and application) approaches and methods in one system.

Working Group Outputs

Misinformation

We distinguish *mis*information as 'information not corresponding with demonstrable fact' from *dis*information as 'information intentionally falsified to forward discreditable purposes'.

Total cynicism is not a solution. Agreed-upon truths are necessary for trust, without which society cannot operate. The watchword should be 'trust, but verify (*before* acting)'. Therefore, we need known, fair and neutral ways to assess truthfulness.

Current use of AITs to detect and prevent antisocial acts such as terrorism, incitement to violence, etc. rely in part on an understanding of the informational ecosystem within which 'artificial' narratives (intended to sway behaviour) appear. It is very hard to fake the whole environment within which a bot network or bad actor is operating, or to fake the whole background of (for example) deepfake video or audio. Organisations such as Bellingcat, which parses context and background to identify the original sources of online information, and which educates journalists in the same techniques, are immensely valuable examples of effective evaluation of truth. Providing as many people as possible with workable 'authenticity-training' will help suppress dis- and misinformation.

We ask for:

- Open source tools for:
    - National AI research
    - Open innovation
    - Standards
    - Access to data
- A sort-out of the data ecosystem

We offer:

- Networking
- Outputs of this consultation
- Ourselves as demonstrators of some of this content
- Cambridge AI centre to suggest cross-party manifesto regarding AI risks and opportunities
- Microsoft election-guard product (extending to the provision of open-source software; the more parties implement and run it, the more trust it can help generate)

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager

We recommend:

- Future events similar to this on neutral ground between interested parties who would not usually encounter one another for in-depth exchange of information and ideas
- Curated resources for workstreams – for example, each participant to share their top five online resources on AI
- A respected voice such as the Prince's Trust to lead conversation about 'good' elections (concerning trust and technology)
- A public 'trustmark' (e.g., through the FCA) for 'goodness in algorithms'

Social disruption

'AI' and 'human dignity' are two elephants in the room. We looked for touchpoints between humanity and AITs, whether these might be involuntary effects, conscious interactions or the creation of new outputs. Such touchpoints emerge in, for example, medicine and social media, and can be used to guide responses to issues around interactions between humans and AITs.

We need agreed, universal definitions for both 'human' and 'AI', since definitions for each can differ widely from context to context. We conclude that human dignity is indispensible, and intimately connected with meaningful work. Thinking further about loss of work, we believe that AITs threaten to disrupt societies by removing or transforming necessary work faster than people can retrain, and/or by exploiting the results of work (e.g., writings, images, shared content) to benefit others than the author(s). We also feel that AITs are unlikely to acquire such human characteristics as the ability to show mercy in judging a legal or medical case, and that there is innate value to human beings in such characteristics.

Disruption as an outcome of AITs is one thing; a resulting loss of human-to-human interaction is another, with effects of its own.

We ask for:

- Principles of human dignity to inform better design, which requires a framework to be developed within which to tackle issues such as mental health impacts, impersonation, etc.
- Raising literacy around AITs along with effective (early) education in interpersonal skills – and a re-examination of whether active retraining will suffice, and occur fast enough, to help people cope with more rapid changes as working adults (such as loss of, or far-reaching change in, certain professions)
- Models which can be openly and readily explained from the point of view of maintaining or increasing human dignity – relating to the need for trusted data and access to it, so that systems in turn can be better trusted
- Transparency for the end-user on *what* AI is included in any product or service: will the major platforms cooperate?
- (In the context of policy changes and more and more rights accruing to gather analytics data,) clear means to opt out of data-harvesting efforts, and clearer EULAs relating to use of data and of AITs; given that the persistence of data is already a problem, should the right to be forgotten extend into models employing AITs?

Dangerous activities

Defining 'dangerous activities' necessarily begs the question 'What do we regard as dangerous, to whom, and why?' which in turn requires agreement on who 'we' are and opens such issues as political attempts to restrict access to technology, likely resulting in parallel technological and (importantly) AIT worlds on the same planet.

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager

Then, a difficulty with risk analysis in the area of AITs (as 'instantaneity machines' for human desires, as we might think of them) is that the stakes for risks as varied as open-source bioweapons, suicide-prompting, induced market crashes, election-rigging, etc, are high or irreversible.

Relying on likely-perverse organic incentives to control powerful, readily available, dual-use technologies would be foolish. Yet, while the faster pace of change increases risks, we still want to keep innovation and competitiveness alive. Therefore, we look to government and regulatory bodies – and, perhaps more urgently and importantly, to international cooperation within industries associated with AITs – to steer AITs firmly away from visible dangers at least.

We ask for and offer to support:

- International prohibition of certain activities (particularly ones which may cause danger of death, climate impact or irreversible human harm), along with internationally recognised definitions of such prohibited activities
- A balanced regulatory framework recognising the dual-use nature of AITs – almost every positive use-case has a negative corollary (e.g., gain of function, gene editing, autonomous vehicles, image recognition and use in social media)
- An international approach to monitoring, deterrence, accountability and sanction, taking an accountability-led approach to software, data and hardware (like that for signing company accounts or aircraft design and operation), with individual and organisational responsibility asserted possibly through a 'Hague-style' international tribunal
- Open reporting of accidents in a secure way to an independent international body, as for cyber and air transport
- Definitions of best practice with benchmarking, such as agreed standards for models, data and sustainability, as are widely used in safety-critical industries, and akin to software B-Corp

We also discussed the importance of near-term action to create AI rules of engagement for elections in democratic nations including UK and USA, to be agreed by the main political parties, and we agreed that this should be the focus of a future conversation.

Finally, in the short term, we seek to widen our network to see more of what is really happening.

Lawmaking and regulatory functions

At the moment, no formal space exists in which to find out regulations relating to AI. Extant information and activities are ad hoc and scattered.

Regulation, too, constrains only those actors able to be constrained (not always an obvious point). AITs, as tools, are inherently conscience-free; we know no way to endow current technology with such conscience that it will never act for evil (and, indeed, humans have trouble reliably doing so with other humans).

There are difficult and unresolved questions about the extent to which to regulate inputs into AITs (primarily training data), outputs of AITs, or both. We would like to see consideration of principles such as copyright opt-ins, and for all content generated using AITs to be labelled as such.

Domain-specific learning models already offer a way to restrict use of or access to information identified as privileged or private (such as medical records). Licensed users are able to visit and consult the dataset under specific conditions without being able to

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager

remove or replicate its data elsewhere, analogous to use of unique manuscripts in a library.

We ask for, and offer, the creation of:

- A website, or wiki, including expert monitoring, defining the regulatory landscape and its bodies, systems, overlaps and disagreements
- A database on AI and data maturity in different projects (we are all connected to some such projects)
- An expertly curated established space for professionals and others to have conversations and provide information and ideas for the above two resources

These would be distinct spaces, but could be linked under one identity/ website to act as one source of information to contribute to practitioners' knowledge of the global regulatory landscape. We can contribute to producing this package and making it public.

We also offer (as a group) to be a resource for each other and others, going forward.

We ask for:

- Compulsory ethics sessions for groups involved with AI (e.g., Cumberland Lodge could offer such a product)
- A trusted assurance scheme to make it harder for 'AI snake oil' to be sold – to be implemented centrally, by specialists, rather than downstream
- More transparent information about where investments into AI are going, to enable future-scanning and scenario planning to help guide proactive steps for good
- A 'regulatory sandbox' supported by government and business, to provide a safe space in which to test AITs' functions against regulatory frameworks – this could operate in specific contexts, such as a 'public sector pilot' for adult social care
- Political parties to commit to not using deepfake content

Ways Forward

There was broad agreement across the four discussions that short-term risks *are* important and deserve immediate focus. Misinformation was broadly perceived as a particular risk as highlighted by responses to fake or misleading 'reportage' surrounding the Gaza conflict. Addressing misinformation in the upcoming UK elections is an obvious point for attention.

Given the variety of points of view and learning opportunities that emerged at the consultation, it will be valuable for St George's House to reconvene members of this group and others to discuss more specific AI risks and solutions.

tldr; summary by ChatGPT

We provided the following prompt to ChatGPT: "Please produce a summary of approximately 200 words of the following report of a consultation meeting, including a bullet on the output of each working group: '[text of this report before "tldr; summary by ChatGPT"]'". This is what it wrote (unedited):

> The St George's House AI Threats and Opportunities Consultation, held on October 19-20, 2023, addressed four key issues: Misinformation, Social Disruption, Dangerous Activities, and Lawmaking/Regulatory Functions. The discussions emphasized the challenges of defining AI and its multifarious applications, urging specificity in addressing issues and risks. Common themes included acknowledging the unknown aspects of AI's impact and the importance of immediate action to address short-term risks.

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager

Working Group Outputs:

1. **Misinformation:**
   - Distinguished between misinformation and disinformation.
   - Emphasized the need for agreed-upon truths and proposed open-source tools, data ecosystem organization, and authenticity training.
   - Recommended a public trustmark for algorithmic "goodness" and a cross-party manifesto on AI risks.

2. **Social Disruption:**
   - Explored touchpoints between human and AI interactions, highlighting the importance of human dignity.
   - Addressed potential disruptions caused by AITs in the workforce and advocated for principles of human dignity in design, literacy around AITs, and transparency in AI product/service content.

3. **Dangerous Activities:**
   - Discussed high-stakes risks associated with AITs, such as open-source bioweapons and induced market crashes.
   - Called for international cooperation to prohibit certain activities, a balanced regulatory framework, and an accountability-led approach to software, data, and hardware.

4. **Lawmaking and Regulatory Functions:**
   - Advocated for a centralized platform to navigate AI regulations, ethics sessions, transparent information on AI investments, and a regulatory sandbox for testing AITs against frameworks.

The report highlighted the urgency of addressing short-term risks, particularly in combating misinformation ahead of the UK elections in 2024, and suggested reconvening to delve deeper into specific AI risks and solutions.

COORDINATING STAFF AT ST GEORGE'S HOUSE
Gary McKeone, Programme Director
Caroline Biggin, Consultation Coordinator, Tel 01753 848851
Catherine Morgan, House Manager