November 2021
APPG AI Evidence Meeting

BIG INNOVATION CENTRE

Cybersecurity and Regulation:
AI technologies in the prevention of misinformation and cybercrime

PARLIAMENTARY BRIEF

Cybersecurity and Regulation: AI technologies in the prevention of misinformation and cybercrime is *a* Parliamentary Brief based upon the All-Party Parliamentary Group on Artificial Intelligence (APPG AI) Evidence Meeting held online on the 15th November 2021.

This APPG AI is co-Chaired by **Stephen Metcalfe MP** and **Lord Clement-Jones CBE**.

We would like to express our appreciation to the following people for their oral evidence:

- **Nick Pickles,** Senior Director of Global Public Policy Strategy, Development and Partnerships, **Twitter**
- **Professor Martin Innes,** Director of the Crime and Security Research Institute**, Cardiff University**
- **Andy Watkin-Child,** Founding Partner**, The Augusta Group**
- **Dr Elżbieta Drążkiewicz,** Senior Research Fellow at the Institute for Sociology**, Slovak Academy of Sciences**
- **Professor David Rand,** Erwin H. Schell Professor and Professor of Management Science and Brain and Cognitive Sciences**, MIT**

Big Innovation Centre is the appointed Secretariat for APPG AI

- CEO, **Professor Birgitte Andersen**
- Rapporteur: **Dr Désirée Remmert**

The video recording of the Evidence Meeting can be found on our websites.

www.biginnovationcentre.com | Email: info@biginnovationcentre.com | @BigInnovCentre

https://uk.bicpavilion.com/about/appg-artificial-intelligence |
Email: appg@biginnovationcentre.com | @APPG_AI

PARLIAMENTARY BRIEF

Cybersecurity and Regulation: AI technologies in the prevention of misinformation and cybercrime

All Party Parliamentary Group on
Artificial Intelligence

APPG AI Sponsors

The Group supporters – British Standards Institution, Capita, CMS Cameron McKenna Nabarro Olswang, Deloitte, Dufrain, Omni, Osborne Clarke, PwC, and Rialto – enable us to raise the ambition of what we can achieve.

Contents

1. Introduction

The question of how to curb the growth in volume and diversity of deep-fakes, misinformation, and cybercrime on social media platforms will be at the core of our discussion. In recent years, misinformation propagated on various social media platforms appears to have had an important impact on political events, social justice, and public health. Further, the relative anonymity of social media platforms has been abused by cybercriminals to steal private data and commit large-scale fraud.

We would like to hear from our expert speakers how conspiracy theories and misinformation emerge and gain traction, how social media can be protected from such attacks, and which role effective policy and cybersecurity technologies can play in the prevention of the spread of misinformation and cybercrime. We would like to address these questions from different perspectives and invite international speakers to share use cases and discuss relevant policy and regulation that can contribute to the prevention of online misinformation and cybercrime

Panel members:

- **Nick Pickles,** Senior Director of Global Public Policy Strategy, Development and Partnerships, **Twitter**
- **Professor Martin Innes,** Director of the Crime and Security Research Institute**, Cardiff University**
- **Andy Watkin-Child,** Founding Partner**, The Augusta Group**
- **Dr Elżbieta Drążkiewicz,** Senior Research Fellow at the Institute for Sociology**, Slovak Academy of Sciences**
- **Professor David Rand,** Erwin H. Schell Professor and Professor of Management Science and Brain and Cognitive Sciences**, MIT**

We addressed the following questions at the meeting:

- **What are the psychological and social factors that are driving conspiracy theories and misinformation?**
- **How should social media be regulated to prevent the spread of misinformation and cybercrime?**
- **How can AI technologies assist in locating and preventing misinformation?**
- **How can cybersecurity systems protect against online fraud and data theft on social media?**
- **What have we learnt from occurrences of major misinformation campaigns and cyber-attacks in the recent past?**
- **How to educate the public about misinformation and cybercrime on social media platforms?**

This meeting was chaired by **Lord Clement-Jones CBE** and **Stephen Metcalfe MP**.

Parliament has appointed Big Innovation Centre as the **Secretariat of the APPG AI**, led by **Professor Birgitte Andersen (CEO)**. The Project Manager and Rapporteur for the APPG AI is **Dr Désirée Remmert**.

2. Recommendations for policymakers

Policymakers must consider **behavioral patterns as well as the complex sociocultural background in which these phenomenons emerge** when looking for solutions. AI-driven technologies can assist to a certain degree in the detection and deletion of misinformation, however, due to the **ambiguity and subtlety of certain false claims** that are spread on social media, these systems are not yet developed enough to filter our all forms of misinformation.

Due to the incremental growth of data shared on social media platforms, **methods that rely entirely on human intervention would not be scalable** and do thus not offer a long-term solution.

Elite misinformation is increasingly spread through **state-owned media, official institutions, or media outlets** which reach a huge audience; labelling these accounts is only a first step to limiting their damaging impact.

A great share of misinformation is a **result of impulsive decision-making and a lack of critical thinking**. **Nudging** users to engage with the content they are about to share can contribute significantly to a decrease in fake news. Using **collective judgment** to label false information would be a further scalable method to detect false information.

The **financial incentives** that often motivate the spread of manipulative claims must be removed to curb misinformation. As soon as social media posts **cannot be exploited as sales pitches** anymore, the spread of false facts loses their financial appeal.

Policymakers have to ensure that the measures that are applied to curb the spread of misinformation are **regularly audited to check their efficiency**.

Discussing the growing **impact of online misinformation on recent political events** and, in particular, on **attitudes towards vaccines in the context of the Covid-19 pandemic**, our expert speakers agree that there is an **urgent need to take measures against efforts to manipulate public opinion and behaviour** through the spread of false information. The **AI technologies** currently available for detecting and eliminating misinformation online would however **not suffice to reliably track down the complex forms of manipulation and their sources**. For this reason, our experts suggest an **eclectic response to the problem of misinformation** that takes **behavioral patterns** as well as the **complex sociocultural background in which these phenomenons emerge** into account.

The deployment of AI technologies to fight online misinformation: problems and opportunities

Professor Martin Innes, Director of the Crime and Security Research Institute at Cardiff University notes that recenty developed AI systems had the potential to solve tasks such as **monitoring and detecting online communication** as well as **linking different pieces of communicated information** at enhanced speed and scale compared with human analysts. He names functions such as AI "being able to **track and trace versions of particular memes/images/videos,** being able to rapidly connect them to the original source disinforming message – something that currently requires significant human analyst effort." This, he continues, would prove essential to curb the increasing disseminsation of **'shallow fakes' and 'deepfakes'**.

An important issue, however, that would aggravate research into online misinformation, so Innes, would be the **restricted access to data**. For this reason, researchers would often be limited to analyse misinformation on isolated platforms, but would not be able to address the important question of **inter-platform coordination in the dissemination of false information**. As an example he names the recent German elections during which misinforming material was discovered on all major social media platforms. He further stresses that issues surrounding online misinformation like **aggressive misinformation campaigns that target individuals to damage their reputation**, have not been unique to the UK but would be **pertinent across Europe and beyond**. In this context, Innes points to a recent report in which his research groups presents evidence of how **Kremlin sympathisers manipulated reader comment sections of various Western media** targeting stories of geo-political interest to Russia. "Increasingly", he claims, "misinformation is disseminated by a blend of media and social media assets interacting with each other."

Implementing regulation to curb to spread of online misinformation

Regarding the question of regulation, our expert speakers stress that it would be important to **strike the balance between measures that constrain false information while simultaneously protecting free speech rights**. Innes notes that a great share of the misinformation spread online would operate on the basis of "distortion, manipulation, and wilful misinterpration." It would be **effective due to its ambiguity** and, at times, its **"grain of truth**", which made its detection even more difficult.

Pickles argues that the regulatory response in question must recognise that **in a free society there would be people with different perspectives**. However, he criticises that if policymakers set **high level policy objectives** but then **delegate the decision of what makes that threshold to private companies**, keeping misinformation in check would be **extremely challenging**.

Our expert speakers indicate that it would be **impossible to rely entirely on human intervention for content control and removal** on social media platforms. As the volume of data shared on online platforms has increased incrementally during recent years, it would become more and more difficult to detect and remove false information. Pickles explains that

from the perspective of Twitter, “content removal is only ever one option. In the past we have often said if it's bad it must come down. Actually, there's **a range of interventions that we can apply that can improve the health of the ecosystem without having to remove speech**.” In the long term, he asserts, the question of how to verify the truthfulness of a piece of content would be of great significance as **human content control would be difficult to scale**. However, at the moment, available **AI-based systems for content check and removal would not be reliable enough** to be deployed for this task. He explains that currently, the most promising methods for content moderation on social media platforms would be software that **compared different content pieces against each other to identify harmful content based on probabilities** and those that **detect and remove fake accounts**.

David Rand, Professor of Management Science and Brain and Cognitive Sciences at MIT, proposes **two scalable methods to reduce the risk of misinformation**. Social media, he explains, creates an environment in which an **individual’s attention is constantly diverted**. However, his research has shown that if you **nudge people to shift their attention towards the concept of accuracy**, the **likelihood of them sharing false content is significantly reduced**. Further, he suggests, exploiting the phenomenon of **“wisdom of the crowds”** for content check appears to be promising. This, he explains, means that **lay people would identify misinformation** and that platforms could use those ratings to either **label content or downrank it.**

However, policymakers should also be reminded that **incentives that motivate the spread of misinformation are in part financial** and part of larger regulatory issues. Pickles reports that at Twitter they observed that people who are now prominent in the anti-vaccine space on social media would at the same time sell products like supplements. Spreading false content would thus work as a **marketing strategy for their businesses**. ”There's a whole industry of pseudo pharmaceutical products”, Pickles warns, “which are generating revenue from disinformation. **Disinformation is a sales pitch for a product** and if you remove the financial incentive of the product, the disinformation itself becomes unprofitable.”

Political misinformation campaigns by state controlled media and government officials

In recent years, Pickles notes, a change could be observed from using fake accounts and personas to **employing the apparatus of state controlled media, government officials, and embassy accounts**. Rand confirms that “there's a lot of work in political science that shows that **elites and elite messaging can heavily influence the opinion of the public**. Particularly politicians can influence the opinions of members of their party. Elites have built in very large audiences, this means that when political leaders and political talking heads make false statements or misleading statements this can have a **wide reaching negative impact that can be hard to undo**.” As a response to the growing problem of coordinated elite misinformation, Twitter changed several policies to curb efforts to manipulate the public opinion through misinformation spread from official accounts. Firstly, **state media are not allowed anymore to advertise on Twitter** and their **accounts must be labelled** as state media. Innes notes that in this context, there would be a lack of clarity about whether the authoring and amplification of false information or the fact that it is exploited by foreign states

to influence geopolitical events should be addressed first by policymakers.

Knowledge and behavioral changes

Rand suggests applying a **psychological approach** to curb online misinformation. A primary driver of people failing to recognise false information, he explains, was a **lack of careful thinking** before sharing information online. Especially when people are **distracted or feeling emotional**, he continues, they would be more likely to believe falsehoods; this would hold true true regardless of whether the falsehoods align with their politics or do not. One key result from their research, Rand notes, “is that distraction and emotional salience are things that are particularly active on social media relative to other places that people get their news from”. They also found that often, the sharing of inaccurate information was **not motivated by the intention to manipulate others**, but by **a lack of critical reflection on the truthfulness of the message** they were about to share. “If they realize something's not true they don't want to share it”, Rand explains, “but the problem is they never even stop to think in the first place about whether it's true or not. […] If it looks like something got shared by tens of thousands of people you might think that means tens of thousands of people believe it, but probably not. It may just be that tens of thousands of people didn't bother to think about it and hit the share button.”

To **motivate critical reflection** on content that users were about to share, Twitter introduced **nudges** that asked people whether they wanted to read content they were about to retweet unchecked. Due to these prompts, Pickles claims, Twitter could motivate a significant number of people to read the content they were about to retweet.

Dr Elżbieta Drążkiewicz, Senior Research Fellow at the Institute for Sociology at the Slovak Academy of Sciences analyses the issue of online misinformation through a **socio-political lense** that contradicts some of the psychological solutions suggested by Rand and Pickles. **Instead of knowledge deficiency and science denialism** being at the heart of the problem, she claims, her research has shown that the main issues behind the spread of conspiracy theories in partliciular would be **issues of trust**, especially trust in **state institutions and officials**. Detecting this kind of misinformation on social media would be made even harder by the fact that it was rather the **tone and the phrasing of the messages** that would carry the meaning than the vocabulary used. “For misinformation, rumors, or conspiracy theories to gain traction”, she explains, “they have to **resonate with pre-existing experiences of individuals or collective memories and cultural tropes**”. Hence, the issue that had to be addressed was not so much a lack of knowledge as a lack in trust in the state. “To address the problem of conspiracy theories”, Drążkiewicz suggests, “we need **holistic solutions which are tailored to the specific context of each country and society**.” It would not be enough to merely rely on technology to solve these problems, but policymakers must find solutions that will “address **local anxieties and support the rebuilding of social and political trust**”.

Accountability

Innes warns that to **hold social media companies accountable**, a correct **assessment of**

the effect of their measure to curb the spread of misinformation mattered. He reports that recent research his group conducted found that the impact of Facebook's 'de-platforming' of two high profile Covid conspiracy theorists had eventually a **converse effect**. "On the surface it appeared that these measures limited the spread of their ideas to the mainstream public", Innes explains, however, "in both cases **the impact of de-platforming was mitigated by their supporters developing new accounts and groups to continue disseminating their ideas**, and the principals diversifying their social media presence onto other platforms." This would suggest that de-platforming "may constrain the public reach of harmful content, it can also have the **unintended consequence of intensifying the radicalisation of 'devoted' followers** of the target." Policymakers should therefore not leave the decision-making on what are effective measures to curb misinformation to social media companies, but should **implement regulation for the auditing of their decision-making on harmful content and the measures they apply to remove it.** "Platforms should be required to **give regulatory bodies access to their data in an unfiltered form**", Innes stresses, "in order that the regulatory agencies can **conduct their own enquiries** (working with independent experts in analysing these issues) to detect the presence of mis/disinformation being spread on the platforms. This would significantly **enhance transparency and potentially encourage the platforms to pro-actively identify such materials**."

Overview

In sum, in order to prevent the spread of false information on online platforms, policymakers must consider **behavioral patterns as well as the complex sociocultural background in which these phenomenons emerge** when looking for solutions. AI-driven technologies can assist to a certain degree in the detection and deletion of misinformation, however, due to the **ambiguity and subtlety of certain false claims** that are spread on social media, these systems are not yet developed enough to filter our all forms of misinformation. Further, our experts stress that due to the incremental growth of data shared on social media platforms, **methods that rely entirely on human intervention would not be scalable** and do not offer

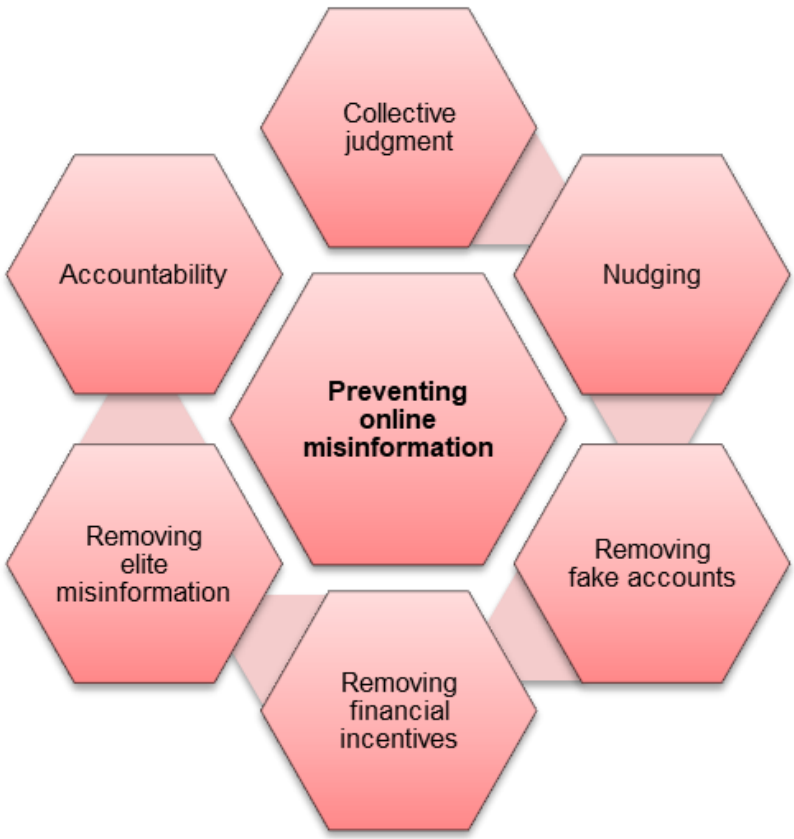
a long-term solution. Moreover, elite misinformation is increasingly spread through **state-owned media, official institutions, or media outlets** which reach a huge audience; labelling these accounts is only a first step to limiting their damaging impact.

Yet, not all misinformation is spread due to ill intent but often, as our experts claim, it is a **result of impulsive decision-making and a lack of critical thinking**. **Nudging** users to engage with the content they are about to share can contribute significantly to a decrease in fake news. Using **collective judgment** to label false information would be a further scalable method to detect false information.

Further, it has been suggested that the **financial incentives** that often motivate the spread of manipulative claims must be removed to curb misinformation. As soon as social media posts **cannot be exploited as sales pitches** anymore, the spread of false facts loses their financial appeal.

However, policymakers should be aware of the sociopolitical context from which these issues emerge and take a **holistic approach in addressing the issue of online misinformation**. Further, they have to ensure that the measures that are applied to curb the spread of misinformation are **regularly audited to check their efficiency**.

Illustration: Preventing online misinformation



3. Evidence statements

Nick Pickles, Senior Director of Global Public Policy Strategy, Development and Partnerships, Twitter

My team at Twitter is responsible for a host of different things. We are the team that works closely with our product and trust and safety teams on designing changes to the product and policies. We run partnerships with some of our critical global partners such as the World Health Organization as well as with NGOs around the world. We are also looking at the the big picture of how we shape regulation.

Last month we published a white paper which sets out some high level principles that should inform and guide policy making. The starting point for that paper is the realization that regulation cannot solve everything. Sometimes when we talk about about the objective to have no misinformation online, there is an early acceptance that it's very unlikely to be achievable, not least given that we all tend to disagree on what constitutes this information. It is often based on our personal views.

From our perspective, the content removal is only ever one option. In the past we have often said if it's bad it must come down. Actually, there's a range of interventions that we can apply that can improve the health of the ecosystem without having to remove speech. This is a really important part of how you strike the balance between a regulatory response, but also one that recognizes that particularly in some of these policy areas in a free society you are going to have people with different perspectives. Particularly when it comes to regulation, one of the challenges we are seeing is is this trend from policy makers to set very high level policy

objectives and then have private companies make the decisions of what meets that threshold.

Certainly, one of the things that we've raised as a concern in the online safety bill is that the very broad definitions inevitably are going to push some of that decision making on to private companies rather than being made by legislators. That's something we've seen around the world and is a trend. On the surface layer you see the content that people are sharing, the underlying infrastructure often contributes to the incentives. Particularly in the space of disinformation, financial incentives are a real part of the challenge. It may seem strange, but if you look at many of the people who are now prominent in the in the anti-vaccine space, they actually sell products like supplements. Those of you who have been following these issues for a long time may recall people selling filters to remove fluoride from water. They were selling the product and actually the products themselves were not regulated. There's a whole industry of pseudo pharmaceutical products which are generating revenue from disinformation. Disinformation is a sales pitch for a product and if you remove the financial incentive of the product, the disinformation itself becomes unprofitable. That removes the incentive to create it.

Secondly I would like to talk about collaboration. Often, we look at the social media space as a silo and we're part of something called the coalition for content providence and authenticity. That's interesting because it's bringing together both people like the New York Times and the BBC, but also Adobe who make the software that you might use to edit videos to use online. In the long term, the question of how you verify a piece of content is something that's going to be incredibly important because the method of having a human watch this content is very difficult to scale. The reason for that, and I had this conversation with with a group of regulators from Ofcom last week, is that the reality of AI versus the perception of AI is quite a challenge right now. AI isn't really here yet, we're basically working on machine learning and what that means is you have two content moderation approaches. You can check if a piece of content is it the same as another piece of content you've already identified, That's the technique used by Microsoft's photo DNA to identify child abuse material now. You can basically get a probability of if this content is like a piece of content you've already identified. Then you have to make a decision about do you remove something based on a probability that's not AI. The real, objective AI is many years in the future but one of the critical things is behavior. There is a whole other form of analysis that doesn't require you to look at content so you can look at how accounts behave and the signals they give you via a phone number and email address or their IPs. You can do a lot of work at that level without needing to read the content. For us it's far quicker to identify a fake account than it is to read all the tweets and decide if they break a misinformation policy.

You don't just have to rely on content, the really hard thing is, you know companies like Twitter have rules if you tell someone “go vote on Wednesday” and the election is on Thursday. We'll take that down because it's telling people the wrong information about how to vote. However, if you state how much money is sent to the EU from the UK - as a company deciding whether that's true or false isn't our role. We've been experimenting with a number of different interventions, like labeling accounts. We've been looking at putting labels on tweets, but also nudges where you're about to retweet something you haven't read and we'll prompt you and

ask if you are sure you want to retweet this as you haven't read it. We actually found in that space 40 of people more read the article because we prompted them. It doesn't require you to make a judgment on the content, you just do it on the behavior. One of the questions that I was asked in the brief was what have we learned. If i can take the example of state disinformation. Five years ago there was a quite famous case, there was an attempt by the Russian internet research agency to start a scare about turkeys having Ebola before Thanksgiving. It didn't really gain much traction and so over the time the tactics evolved. They moved from spreading outright falsehoods to amplifying existing social divisions. It doesn't require you to be false, you would amplify real Americans in this case and try to start arguments.

One of the things that we found was there's been a change in this space from using fake accounts and fake personas into using the real apparatus of the state controlled media and government officials, including some embassy accounts. In that space we changed a number of policies: state media can't advertise anymore, they have a label on them. We don't amplify them or recommend them so if you go to a Russia Today page, you see very clearly this is Russian State media and it's organically available. That balance and a lot of that understanding was made possible because we actually published the archives of the data we removed from those Russian operations. One of the things that we always encourage policymakers to think about is how to expand the legal protections for publishing more data. I will stop there, but as I say this, this user behavior piece is something that is often forgotten about, but I think it is a big opportunity to increase digital literacy without relying on content moderation as the only kind of approach.

Professor Martin Innes, Director of the Crime and Security Research Institute, Cardiff University

Background

Cardiff University's OSCAR (Open Source Communications Analytics Research) programme has developed a specialist expertise in tracking and tracing the causes, construction, communication and consequences of (mis/dis)information operations and campaigns.

Since 2018, the OSCAR team has worked on mis/disinformation episodes in over 20 countries, including: democratic elections; terror attacks; the coronavirus pandemic; and a range of conflicts and disputes. This has encompassed detailed analyses of the accounts, behaviour and content associated with both foreign state actors, including Russia and China, as well as domestic actors.

The OSCAR programme has secured in excess of £6.5M funding including from multiple UK Government departments, the European Union and the Economic and Social Research Council. The research team has published a range of outputs including policy-facing reports and peer reviewed academic journal articles.

1. How can social media be regulated to prevent the spread of misinformation?

1.1 Tackling mis/disinformation is a particularly challenging regulatory problem. Much misinformation operates on the basis of distortion, manipulation and wilful misinterpretation, and is effective precisely because it is ambiguous and often contains a 'grain of truth'. Striking the right balance between measures that constrain harmful communications, whilst protecting free speech rights, is consequently challenging.

1.2 To make progress, there is a need for a common measurement framework, in order that we can agree what mis/disinformation episodes to prioritise and why, and reliably gauge whether counter-measures are doing what we intend. At the current time, we do not have this. It will require investment in time and resources to develop, but is vital for developing a more robust and evidence-informed approach. It should be led by government and scientists, not industry.

1.3 Measurement matters because, without it, social media companies cannot properly be held to account. For example, we have just published a peer reviewed article assessing the impact of Facebook's 'de-platforming' of two high profile Covid conspiracy theorists. On the surface it appeared that these measures limited the spread of their ideas to the mainstream public. However, digging deeper a more complex picture emerges. We found:

In the seven days following his account being de-platformed on 30 April, public mentions of David Icke on Facebook increased by 84%.

Contrastingly, de-platforming Kate Shemirani initially appeared more promising. In the two months following her account removal, post mentions and user engagement on Facebook decreased markedly. However, the suppression effect appeared temporary - her Facebook video shares increased from approximately ten in October and November 2020 to over 60 in the next two months.

In both cases the impact of de-platforming was mitigated by their supporters developing new accounts and groups to continue disseminating their ideas, and the principals diversifying their social media presence onto other platforms.

Evidence suggests that although de-platforming may constrain the public reach of harmful content, it can also have the unintended consequence of intensifying the radicalisation of 'devoted' followers of the target.

1.4 We cannot persist with a situation where social media providers are effectively 'marking their own homework'. They are currently positioned as key decision-makers in defining what the problems are on their platforms, and which warrant taking action. There is little meaningful oversight of what data they choose to release or not, nor mechanisms for auditing their decision-making in terms of their choices to report or not. It is noticeable that over the past year or so, several major platforms (Twitter, Facebook and YouTube) have significantly extended their definitions of problematic content. However, the concern here is that the 'supply' of problematic content meeting these new criteria far outstrips their capacity to act on all of it. As such, how and why are certain issues going to be selected for action, and others not?

1.5 Developing the above point, there is a fundamental tension between a platform's primary concern with reputation management, and the role of democratic governments to minimise public harms. When platforms 'de-platform' users their principal concern is with removing the problematic behaviour from their surface, and so they can be quite content with seeing malign

actors displaced to other platforms.

1.6 So what could be done from a regulatory perspective? Platforms should be required to give regulatory bodies access to their data in an unfiltered form, in order that the regulatory agencies can conduct their own enquiries (working with independent experts in analysing these issues) to detect the presence of mis/disinformation being spread on the platforms. This would significantly enhance transparency and potentially encourage the platforms to pro-actively identify such materials.

2. How can ai assist in locating and preventing misinformation

2.1 We should not over-estimate the capacity of AI to assist, given how, as rehearsed above our models of misinformation construction, communication and consequences are under-developed. In addition to which, the design and delivery of mis/disinformation campaigns are evolving rapidly in terms of their core tactics and techniques, and also cross-platform dimensions.

2.2 Where AI potentially has a role is in terms of enabling enhanced speed and scale in terms of monitoring and detection, in terms of being able to link different pieces of communicated information together. For example, being able to track and trace versions of particular memes / images / videos, being able to rapidly connect them to the original source disinforming message – something that currently requires significant human analyst effort.

2.3 This capability will become increasingly important as the use of ‘shallow fakes’ and ‘deepfakes’ gathers pace. We can also anticipate that appropriately trained AI models could be usefully deployed to detect the often quite subtle data signals indicating that an image may have been manipulated, thus assisting in visual disinformation detection.

3. Learning from previous misinformation campaigns

3.1 OSCAR’s current situation assessment is that most current significant political events and controversies are acting as ‘magnets for misinformation’, attracting overlapping and interacting distortions and deception, authored and amplified by a range of sources.

3.2 There is a ‘unit of analysis’ problem that is neglected and under-appreciated in terms of how it is shaping our understanding of mis/disinformation and what can be done about it. Different researchers are analysing different issues and episodes from a variety of vantage points, and the choices they make alter the insights and findings generated. Most available empirical data that can be confidently attributed to particular misinformation actors, comes from platforms. This gives a good view of activities on that platform, but is arguably less insightful about cross-platform co-ordination. Studies by NGOs are often focused upon particular democratic events such as elections, or themes such as vaccine hesitancy, or climate emergency denial. Academic research has often been steered by the availability of datasets. Consequently, a lot of this work has focused upon Twitter and Facebook, neglecting the situation on other platforms.

3.3 The overarching point though, is that these differences make it hard to synthesise material to build comprehensive understanding of the design and delivery of (dis)information operations and misinformation campaigns.

3.4 Transferred to the policy environment the fact that these issues typically blur together, creates a lack of clarity about whether the primary problem to be tackled is the authoring and amplification of mis/disinformation, or that it is being used by foreign states as part of their influence campaigns. Political and public attention has particularly centred the role of foreign state sponsored operations and campaigns, especially Russia, China and Iran. However, the research evidence base tends to supplement this view by highlighting how hard- and far-right inspired thought communities on social media have been important vectors for constructing mis/disinformation at scale across a number of topics.

3.5 In an academic article synthesising learning from across the OSCAR programme, we trace twin process of 'normalization and domestication' to describe how the societal problem with mis/disinformation is evolving. Normalization is the sense that it is increasingly accepted that mis/disinformation percolates round all events of major political and social significance. Domestication references how tactics and techniques pioneered by foreign state campaigns are increasingly being adopted and adapted by more locally oriented actors.

3.6 We recently completed work on the German election, and the above concepts capture much of what we observed. There were a number of patterns that are likely to recur going forward in many countries, including:

Multiple misinformation authors, operating with different motives including political activists, conspiracy theorists (Querdenker/QAnon), domestic far-right groups and foreign far-right (US). Additionally, there was a lot that we could not attribute to a source because the obfuscation of the author's identity was too sophisticated. Some of this may have been foreign state linked.

Multiple tactics, techniques and procedures were used to construct misinformation, including: apparent targeted hack and leak smear campaigns against individual politicians; campaigns utilising both social media and mainstream media to propagate their message; using narratives from the US 2020 election to attack the integrity of the German processes.

Misinforming material relating to the election was detected across all major platforms including TikTok, Reddit and some new ones such as Gettr.

3.7 One further insight from working on these issues across Europe is seeing how current concerns in the UK are not unique, but reflect trajectories of development that have been developing in other countries for some time (eg. aggressive misinformation campaigns targeting individual (especially female) politicians, to damage their reputations).

4. How to educate the public about misinformation on social media platforms

4.1 The public and political conversation needs to move beyond its current fixation upon

Twitter, Facebook and YouTube. In a report published in September, we presented evidence of how pro-Kremlin trolls have been manipulating the reader comment sections on Western media outlets, relating to stories of geo-political interest to Russia. Based upon detailed analysis we identified:

242 stories that had been targeted in this way in April 2021, spread across 32 high profile media outlets (Daily Mail, The Times, Fox News, Le Figaro, Der Spiegel) in 16 countries.

4.2 The point here is that social media is only a part of how misinforming messages are being constructed and communicated. As individual platforms have improved their defences, our adversaries have innovated and identified new vulnerabilities and to exploit. Indeed, one of the key strategic 'take homes' from our ongoing research programme is that increasingly misinformation is disseminated by a blend of media and social media assets interacting with each other.

Andy Watkin-Child, Founding Partner, The Augusta Group

Key observations to take away

1. With AI, machine learning and digital having a profound impact on the society. How can the UK public and private sector work alongside its partners without adopting similar cybersecurity standards and cyber risk strategy?

2. The direction of US cybersecurity regulation and legislation is very clear. How does the UK meet those requirements and how can the UK benefit from the US legislative and regulatory agenda?

3. US cyber and risk management standards are significantly higher than those in the UK. How can we achieve cybersecurity reciprocity, oversight, assurance, and accreditation with our current standards?

4. There is a global shortage of skilled cybersecurity and cyber-risk management resources. How will the UK up-skill its public and private sector cyber defences?

Introduction

Cyber is a complex risk, it has become the most significant non-financial risk the public and private sector faces. The January 2021 World Economic Forum (WEF) report on 'Global Risks 2021 16th Edition' identify cybersecurity failure as a top 7 global risk1. The WEF March 2021 report on 'Principles of Board Governance of Cyber Risk', identify cybersecurity failure as a top 4 short term (0–2 years) risk behind Infectious disease, livelihood crisis and extreme weather events for the board room2. The number, complexity and severity of cyber-attacks have increased with Ransomware now the predominant attack vector used by nation state

hackers and their proxies. With well documented attacks in 2020/ 2021 such as the SolarWinds, Microsoft Exchange, Colonial Pipeline, JBS meat packing and Kaseya3. Causing supply chain disruption, the loss of government data and services, cost to society and remediation costs. It is reported that Colonial Pipeline paid $5Million to unlock their data, JBS paid $11Million and a demand was made to Kaseya3 to pay $70Million.

Unfortunately, it has been well demonstrated cybersecurity-risk is not being managed. Where cyberattacks were once an extreme loss event they are commonplace. The severity, complexity, frequency and impact of cyber-attacks and their impact to financial statements increasing significantly over the past 10 years. Where Distributed Denial of Service (DDoS), Phishing and SQL injection were a critical attack vector they are now tools to support ransomware. Ransomware is the weapon of choice for hackers, with a proven record of causing significant damage to infrastructure, disruption of supply and generating significant profits for the hackers. Ransomware attacks spiked 170% higher in 2020 than in 2019, the average costs of remediation jumped from $700,000 to $1,850,000 in the same period forcing cyber insurers to increase prices4. Cyber is an unexpected loss at best and an expected loss at worst, requiring intervention through regulation.

The traditional rules of engagement do not apply to cyberattacks

Adversaries are rarely identified, are difficult to locate, their targets can be indiscriminate or specific, the attacker can range from well-funded nation states to school kids buying cyber as a service attack from their bedroom. Cyber-attacks are used as a geopolitical weapon and to inflict cybercrime by Nation States, their proxy's, and private hacking groups.

Where cyber related incidents were once 100-year events, now they are regular occurrences. Cyber has moved from extreme loss to unexpected loss and closer to an expected loss event. Cyber is an issue the public and private sector is coming to terms with. The risk is becoming harder to mitigate, expected Losses are expensive to manage and the costs of remediating unexpected loss is increasing (cyber insurance, capital holding), there is a lack of cyber-skilled resources globally and it is difficult to budget for Cyber-Supply Chain Risk (C-SCRM) attacks which start outside an organisation i.e. NotPetya, SolarWinds, JBS, Colonial Pipeline and Kaseya.

Managing cyber risk requires a defensive cyber strategy

Society relies upon the state to protect it and manage conventional offensive and defensive capabilities, providing physical deterrents and intelligence services. Cyber is a domain of operation and nation states are building up their offensive cybersecurity capabilities. However, society is dependent upon cyberspace and the products and services which are provide by the public and private sector. Critical National Infrastructure (CNI) including energy, oil, gas, water, oil and gas, healthcare, FMCG and defence manufacturers that are overwhelmingly private sector companies. Where governments have little control over their security posture, that is managed through market forces.

Cyber Regulation

Recent successful cyber-attacks have demonstrated that market forces alone are not working. The US government has increasingly adopted cyber legislation in 2021 to address identified failures in the public and private sector to manage cybersecurity. Outlining regulatory enforcement programs for cyber security through the Department of Justice (DoJ) and Treasury (DoT). Executive Order 14017 (Americas Supply Chains, February 20215) initiated reviews of current Supply Chain Risk Management (SCRM) across Federal Agencies cyber capabilities, oversight, assurance, software development, Federal threat sharing, incident response and collaboration. Executive Order 14028 (Improving the Nation's Cybersecurity, May 20216) set out a range of activities for the Federal Government to assess, recommend, and improve the protection of US National Security.

US Cyber regulation, proposed legislation and enforcement

Regulation.

The US adopted the Federal Information Security Modernization Act (FISMA) in 2002, updated in 2014 and legislation has been proposed to update it in 20217.

DFARS 252.204 - 7012 was regulated by the DoD in 2017 requiring cybersecurity to be implemented by their Defence Industry Based (DIB) Globally. Updated in 2020 (DFARS 252.204-7019 and 7020) 8.

Legislation introduced across the House and Senate - 2021

In process

- Cybersecurity Maturity Model Certification (CMMC).
- FISMA (2021).
- H.R 3684 - Infrastructure Investment and Jobs Act9.
- H.R. 5440 - Cyber Incident Reporting for Critical Infrastructure Act9.
- S. 2407, Cyber Incident Notification Act – NB Requires the reporting of cyber incidents to CISA. (S. 2875, Cyber Incident Reporting Act – NB Sets-up Cyber incident review office in CISA)9
- S. 2943, Ransom Disclosure Act9.

Planned

- Security and exchange commission – Cyber Risk Governance (2021 regulatory calendar)10

Enforcement

Department of Justice (DoJ)

- Civil fraud Initiative (October 2021) and the utilization of the False Claims Act (FCA) to pursue companies, that are government contractors who receive federal funds, when they fail to follow required cybersecurity standards11.

Department of Treasury (DoT)

- Cyber Ransomware payments and OFAC (Office of Foreign Asset Control – October 2021) – requiring the reporting of Ransomware attacks to Federal Agencies and OFAC12.

Securities and Exchange Commission (SEC).

- Existing market regulation requires the reporting of material risks to the SEC

The impact of US cyber regulation on the UK and UK companies is to be considered could be significant, however is as yet un-quantified

- The US DoDs DFARS 252.204-7012 required DoD contractors and subcontractors to implement NIST SP 800-171 cybersecurity standards since December 2017. As of November 2020 DoD contractors must submit compliance scores to the DoD. The DoD will use these scores to issue DoD contracts. Failure to provide a score could result in a UK contractors or subcontractors not receiving a contract. US DoD proposals will require all contractors and subcontractors across their supply chains to submit a certificate of CMMC compliance before they can be awarded a contract or subcontract.
- US Enforcement regimes are likely to have a punitive impact of UK contractors.
- US cyber standards are significantly higher than our own. The cost of compliance for UK industry will be high. (cyber essentials vs. NIST SP 800-171, ISO 27001 or CyberSecurity Framework Profiles)
- FISMA (2014) requires Federal agencies and contractors to implement information risk management practices.
- The global lack of appropriately skilled cybersecurity resources will impact the development of the UK economy.

Cybersecurity and Artificial intelligence

- Cyber and AI are inextricably linked.
- AI relies on IT and OT technology platforms and data to function. Platforms that do not operate in isolation and will be connected to networks for the collection and sharing of information.
- As the use of AI develops the data associated with AI platforms becomes more useful to cyber criminals and nation States e.g. Biometric data.
- AI systems take more control of our lives i.e. Driverless cars or airborne taxis. Cyber threats could present a danger.
- Current cybersecurity standards are not mature enough to provide the necessary oversight and assurance of cybersecurity risk management required.

References

1. World Economic Forum: The Global Risks Report 2021(Jan 2021) http://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2021.pdf
2. World Economic Forum: Principles for Board Governance of Cyber Risk (March 2021) http://www3.weforum.org/docs/WEF_Cyber_Risk_Corporate_Governance_2021.pdf
3. Kaseya attack: https://www.abc.net.au/news/2021-07-04/kaseya-cyber-probed-for-russia-links-joe-biden/100266350
4. The cost of a cyber-attack: https://www.insurancebusinessmag.com/us/news/cyber/global-cyber-insurance-pricing-spikes-32--report-259795.aspx
5. Executive Order 14017 : https://www.whitehouse.gov/briefing-room/presidential-actions/2021/02/24/executive-order-on-americas-supply-chains/
6. Executive Order 14028 : https://www.whitehouse.gov/briefing-room/presidential-

actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/
7. FISMA (2021): https://www.nextgov.com/cybersecurity/2021/10/senate-committee-passes-major-fisma-changesincluding-new-definition-major-incident/185914/
8. DFARS Interim Final Ruling: https://www.federalregister.gov/documents/2020/09/29/2020-21123/defense-federal-acquisition-regulation-supplement-assessing-contractor-implementation-of
9. Cyber incident reporting Bills: https://crsreports.congress.gov/product/pdf/R/R46944
10. Securities and Exchange Commission : https://www.sec.gov/news/press-release/2021-99
11. DoJ Cyber fraud Initiative: https://www.natlawreview.com/article/doj-announces-new-cyber-fraud-initiative-and-intent-to-utilize-false-claims-act-to
12. DoT OFAC and ransomware: https://home.treasury.gov/system/files/126/ofac_ransomware_advisory_10012020_1.pdf

Dr Elżbieta Drążkiewicz, Senior Research Fellow at the Institute for Sociology, Slovak Academy of Sciences

In recent years conspiracy theories have been frequently defined as a threat to liberal democracies. As a result various stakeholders are increasingly interested in finding solutions that will help to prevent the spread of conspiracy theories and their impact on social and political spheres. For instance, the European Commission initiated The East StratCom Task Force and established the European Democracy Action Plan, which aims to counter disinformation and address the problem of conspiracy theories. In 2018 Germany introduced the Network Enforcement Act penalising the spread of disinformation. Since 2020 other states have taken similar steps [8-10]. Other countries, such as Sweden, look for solutions in education and awareness-raising campaigns [11-13]. Across Europe, many NGOs (such as the EU's DisinfoLab and International Institute for Strategic Dialogue, Poland's Demagog.org and Centrum Cyfrowe, or Estonia's Propastop) are today focusing on combating conspiracy theories [7, 12, 14-16].

When it comes to medical conspiracy theories (for instance narratives concerning vaccinations, new diseases etc.) many of those interventions begin with the premise that conspiratorial thinking is a problem of misinformation, and knowledge deficiency. But today we know that even highly educated people can express conspiratorial beliefs [22, 111-114], and that less educated people endorse conspiracy theories not because they are less intelligent, but because they feel powerless [84]. We also know that for decades, people were opting-in for vaccination programmes not because they were informed about vaccines, but because their doctors told them to do so, and it was the 'right thing to do'. The Andrew Wakefield controversy can be very informative here. Since false theories regarding the MMR vaccines started to circulate in the late 1990s, countless efforts have been made to increase people's knowledge on vaccinations, to counter the misinformation. However, the

effectiveness of those debunking campaigns is sometimes questioned (Betsch and Sachse 2013). In spite of all informational and 'debunking' campaigns, vaccination uptake is still problematic in certain areas, so much so, that the World Health Organisation announced vaccination hesitancy as one of its main challenges for 2019.

This is because, while conspiracy theories appear to be obsessed with 'facts' and 'information' they are an expression of anxiety caused by specific events (for instance an illness that hits the family, a global pandemic). They can be also a reaction to changing relations (for instance between patient and doctor). Conspriacy theories are also frequently a form of criticism of the imbalances of power – such as those characterising relationships between patients and doctors, patients and healthcare administrators, patients and companies and institutions producing medications or offering medical services {Harambam, 2015 #1417;Højer, 2020 #3237;Fassin, 2011 #1549;Briggs, 2004 #1381;Butt, 2005 #1467;Sobo, 2015 #1019;Sobo, 2016 #3239}. Conspiracy theories are narratives characterized by suspicion, distrust, and a feeling of persecution, describing the world as a space where bad things happen because of the evil plots carefully crafted in secrecy by powerful individuals or groups. Therefore, when designing solutions to the problem of medical conspiracy theories it is important to address the key problem of trust deficiency (not just knowledge deficiency).

Even though on the surface medical conspiracy theories appear to be obsessed with 'facts' and 'truth', many of those narratives point to the fundamental problem of mistrust in medical communities and the deteriorating relationship between patients, healthcare administrators, medical professionals and pharmaceutical corporations. An important factor in shaping those relations is gender, race and socio-economic class. Studies show clearly, that women who experienced oppressive reproductive regimes, who have negative individual or collective experience with healthcare are more frequently opting out of immunisation programmes {Pop, 2016 #2575}{Drążkiewicz, 2021 #3224}. Similar can be said by economically disadvantaged groups who do not have satisfactory access to healthcare {Briggs, 2004 #1381}. The vaccine attitudes of many black people are informed by the memory of unethical medical trails and scandals regarding unwanted sterilisations and eugenic programmes {Momplaisir, 2021 #3448}.

Therefore instead of focusing only on 'what people know', we should also start asking what is it that makes it easier for people to believe or disbelieve in science, to trust or not trust doctors and health administrators? This requires actions that address specific anxieties and support rebuilding of broken relationships. This requires solutions that move beyond focusing exclusively on people who endorse conspiracy theories and instead also target other stakeholders who have have a possibility on impacting peoples believes in medical conspiracy theories: healthcare workers, health administrators, representatives of scientific communities and pharmaceutical companies.

David Rand, Erwin H. Schell Professor and Professor of Management Science and Brain and Cognitive Sciences, MIT

I am a professor at MIT and for the last five years, I've been trying to understand misinformation and fake news on social media - why people fall for it and what to do about it. I'll try to give you an overview of a lot of that work. In addition to doing research, I also have been advising tech companies like Google, Twitter, and facebook on some of their anti-misinformation approaches. So I have some understanding of what's going on in the social media space. We also want to understand why it is that people fail to discern truth from falsehood. This is not just on social media but in general.

We have done a lot of research at this point that suggests that a primary driver of people failing to tell what's true and what's not is simply that they are not thinking hard enough and not thinking carefully enough; just going with their gut responses. We have shown in experiments that when people are distracted, or people are feeling emotional, or they're just people that tend to not think carefully about things they're more likely to believe falsehoods. That's true regardless of whether the falsehoods align with their politics or don't. There are many stories about motivated reasoning, where people's reasoning abilities are hijacked by their identities; getting people to think more doesn't actually help. I'd say don't buy the motivated reasoning hype, like reasoning is good and getting people to engage in more reasoning is in general going to be helpful.

I think one key result and another related result is that distraction and emotional salience are things that are particularly active on social media relative to other places that people get their news from. This may be part of why the social media context is maybe particularly problematic. Another key result is that most people do not want to share accurate news. If they realize something's not true they don't want to share it, but the problem is they never even stop to

think in the first place about whether it's true or not. This is because the social media context is focusing their attention on all this other social stuff, so they never even get to that first step. If it looks like something got shared by tens of thousands of people you might think that means tens of thousands of people believe it, but probably not. It may just be that tens of thousands of people didn't bother to think about it and hit the share button.

I think the third major driver of misinformation and failing for misinformation is the role of cultural and political elites. There's a lot of work in political science that shows that elites and elite messaging can heavily influence the opinion of the public. Particularly politicians can influence the opinions of members of their party. Elites have built in very large audiences, this means that when political leaders and political talking heads make false statements or misleading statements this can have a wide reaching negative impact that can be hard to undo. A large set of people immediately hear it, it gets covered by mainstream news; these are trusted sources. These are the big issues: getting people to think more and to think more about accuracy in particular. Trying to combat coordinated elite misinformation. So if you're asking what can platforms do about misinformation, currently there are two main things that social media platforms are doing and both of these things are good things:

One is using machine learning to identify misinformation and then just downrank it so people are less likely to see it. The other thing is hiring professional fact checkers to fact check articles and then putting labels on them saying “false disputed by fact check”. These are both good, you shouldn't worry about backfire effects. If something is found to be false, fact checkers flag it as false and this will reduce the belief and sharing in it. The problem is that these things aren't scalable, professional fact checking in particular. What we've been doing is trying to find other things that could be added to the anti-misinformation toolkit. In particular, what are scalable ways to reduce the risk of misinformation that doesn't rely on some centralized authority via the government or the platforms deciding what's true or false. We have two interventions to propose. One of them is simply shifting attention back to accuracy. Social media creates an environment where attention is scarce and directed towards all kinds of social things, because you're always getting social feedback who liked it how many people liked it et cetera… But we've shown that it's easy to nudge people to shift their attention back to the concept of accuracy. Once they think for a minute about whether something's accurate or not, they'll reduce the likelihood of sharing bad content. We've done this by showing them random headlines and ask how accurate they think they are. We were giving minimal digital literacy tips that don't teach anyone anything, but just activate the idea of accuracy.

We also conducted a big study where we sent messages to 5 000 Twitter users that were regularly sharing misinformation, asking them to judge the accuracy of a random non-political headline. We showed that this significantly improved the quality of the content that they subsequently shared. These kinds of attention nudges need to be dynamic, you can't always do the same thing. Twitter did this thing where if you went to retweet a link that you hadn't read yet, it asked whether you were sure if you did not wanted to read it first, This was the first time it happened, but then they didn't change it and very quickly I found myself starting to ignore it. I'm sure everybody else was, too. It's like an arms race where you fight a battle for attention. You have to always keep it fresh and novel to keep people's attention. This is something that

platforms are very good at, their whole business model is premised on getting people to pay attention to things they don't want to pay attention to, namely ads. Our idea is that they should be using the muscle that they've developed for getting people to pay attention to ads to get them to pay attention to accuracy. More generally, if you're thinking about any kind of intervention on social media you have to remember the fact that attention is very scarce. Getting people to pay attention and holding people's attention is critical.

Although I wouldn't trust any random social media user's opinion about whether something is true or not, there's at least a hundred years of research showing how the wisdom of crowds can allow the average of a bunch of random people to approximate expert judgments. We've done research showing that with under 20 lay person ratings per headline you can get as much agreement between lay person ratings and fact checkers as you have agreement amongst the fact checkers. This means that having lay people identify misinformation is scalable; platforms can use those lay person ratings to either label content or downrank it. So shifting attention to accuracy and crowdsourcing are two important things to add to the toolkit. Just to conclude, social social media has gotten the vast brunt of the criticism, but it's not the only channel or even necessarily the biggest channel for misinformation. I want to go back to this idea of political elites and mainstream media TV. These are playing a huge role in the misinformation process and are largely getting a free pass. It's really important to hold them accountable as well.

Contact

APPG AI Secretariat

Big Innovation Centre
14-16 Dowgate Hill
London EC4R 2SU
United Kingdom

info@biginnovationcentre.com
www.biginnovationcentre.com

appg@biginnovationcentre.com
https://uk.bicpavilion.com/about/appg-artificial-intelligence