THE CENTRE FOR
LONG-TERM RESILIENCE

AI incident reporting: Addressing a gap in the UK's regulation of AI

Tommy Shaffer Shane[1]

The Centre for Long-Term Resilience | June 2024

Contents

[Executive summary](#)
- [The need for incident reporting](#)
- [A critical gap](#)
- [Next steps for UK Government](#)

[The benefits that incident reporting can bring to AI regulation](#)
[The current state of AI incident reporting: a fragmented, incomplete patchwork](#)
[Three options for a future UK AI incident reporting regime](#)
[Recommended next steps for the UK Government](#)
[References](#)

[1] With research support from Luke Dawes. Contact: Tommy Shaffer Shane, CLTR Policy Manager, tommy@longtermresilience.org;

Executive summary

AI has a history of failing in unanticipated ways, with over 10,000 safety incidents in deployed AI systems recorded by news outlets since 2014 [36]. With greater integration of AI into society, incidents are likely to increase in number and scale of impact.

In other safety-critical industries, such as aviation and medicine, incidents like these are collected and investigated by authorities in a process known as 'incident reporting'.

We – along with a broad consensus of experts, U.S. and Chinese governments, and the EU[2] – believe that a well-functioning incident reporting regime is critical for the regulation of AI, as it provides fast insights about how AI is going wrong.

However, it is a concerning gap in the UK's regulatory plans.

This report sets out our case, and provides practical steps that the Department for Science, Innovation & Technology (DSIT) can take to address it.

The need for incident reporting

Incident reporting is a proven safety mechanism, and will support the UK Government's 'context-based approach' to AI regulation by enabling it to:

1. Monitor how AI is causing safety risks in real-world contexts, providing a feedback loop that can allow course correction in how AI is regulated and deployed;

2. Coordinate responses to major incidents where speed is critical, followed by investigations into root causes that can generate cross-sectoral learnings;

3. Identify early warnings of larger-scale harms that could arise in future, for use by the AI Safety Institute and Central AI Risk Function in risk assessments.

A critical gap

However, the UK's regulation of AI currently lacks an effective incident reporting framework. If not addressed, DSIT will lack visibility of a range of incidents, including:

- **Incidents in highly capable foundation models**, such as bias and discrimination or misaligned agents, which could cause widespread harm to individuals and societal functions;

- **Incidents from the UK Government's own use of AI in public services**, where failures in AI systems could directly harm the UK public, as occurred when Dutch tax authorities used a flawed AI system to detect benefits fraud and plunged 26,000 families into financial distress [42];

- **Incidents of misuse of AI systems**, e.g. detected use in disinformation campaigns or biological weapon development, which may need urgent responses to protect

[2] See [22, 23, 24, 25, 26, 27] for experts' calls for AI incident reporting, and [41] which states that "in the next 2-3 years, the US, EU, and China will have established mandatory incident reporting requirements by AI service providers".

UK citizens;

- **Incidents of harm from AI companions, tutors and therapists**, where deep levels of trust combined with extensive personal data could lead to abuse, manipulation, radicalisation, or dangerous advice, such as when an AI system encouraged a Belgian man to end his own life in 2023 [38].

DSIT lacks a central, up-to-date picture of these types of incidents as they emerge. Though some regulators will collect some incident reports, we find that this is not likely to capture the novel harms posed by frontier AI.

DSIT should prioritise ensuring that the UK Government finds out about such novel harms not through the news, but through proven processes of incident reporting.

Recommended next steps for UK Government

This is a gap that DSIT should urgently address. We recommend three immediate next steps:

1. **Create a system for the UK Government to report incidents in its own use of AI in public services.** This is low-hanging fruit that can help the government responsibly improve public services, and could involve simple steps such as expanding the Algorithmic Transparency Recording Standard (ATRS) to include a framework for reporting public sector AI incidents. These incidents could be fed directly to a government body, and possibly shared with the public for transparency and accountability.

2. **Commission UK regulators and consult experts to confirm where there are the most concerning gaps.** This is essential to ensure effective coverage of priority incidents, and for understanding the stakeholders and incentives required to establish a functional regime.

3. **Build capacity within DSIT to collect, monitor, investigate and respond to incidents, possibly including the creation of a pilot AI incident database**. This could comprise part of DSIT's 'central function'[3], and begin the development of the policy and technical infrastructure for collecting and responding to AI incident reports. This should focus initially on the most urgent gap identified by stakeholders, but could eventually collect all reports from UK regulators.

[3] DSIT has set out a range of potential 'central functions' in *A pro-innovation approach to AI regulation: government response*. We suggest that one could be collecting and actioning incident reports.

Structure of report

This report covers the following:

Section 1 – We set out why incident reporting is an essential component of effective AI regulation, outlining the benefits it can bring to DSIT, regulators, and the UK public.

Section 2 – We highlight why incident reporting is a gap in the UK Government's regulatory plans, meaning that DSIT will be blind to many of the most concerning safety incidents, despite a loose patchwork of regulator regimes at the fringes of AI, and voluntary efforts.

Section 3 – We set out how an incident reporting regime could work in practice, providing three options for how the UK Government could design an AI incident reporting regime, with a focus on ensuring that regulators work effectively together with DSIT to create an accurate, actionable picture of how AI is causing safety incidents.

Section 4 – We provide three concrete recommended next steps for DSIT to take, prioritising the most urgent gaps and creating a path for a future, reliable regime.

> AI has a history of failing in unanticipated ways, with over 10,000 safety incidents in deployed AI systems

Definitions

Incidents refer to situations where a technical system or a product has caused harm (sometimes an 'accident'), or has nearly caused harm (sometimes referred to as a 'near miss' or a 'hazard'). Some industries use slightly different terms with specific definitions and scope. For example, the Medicines and Healthcare products Regulatory Agency (MHRA) uses the term 'adverse incident' to describe which incidents involving medicines and medical devices require reporting to them as soon as possible, and includes when:

- a medicine causes side effects
- someone is injured (or almost injured) by a medical device, either because its labelling or instructions aren't clear, it's broken or has been misused
- a patient's treatment is interrupted because of a faulty device
- someone receives the wrong diagnosis because of a medical device
- a medicine doesn't work properly
- a medicine is of a poor quality
- a medicine or medical device may be fake or counterfeit [31]

Incident reporting refers to the process of reporting those situations to an authority (e.g. a regulator) so that they can record, investigate and respond.

Harm will have different definitions depending on the context. Because AI is a general-purpose technology, harms could be wide ranging. The OECD identifies the following types of harm that may be relevant to AI incidents: physical harms; environmental harms; harms to economic, property, and finance; reputational harms; harms to the public interest; harms to human rights; psychological harms [30].

The benefits that incident reporting can bring to AI regulation

In this section, we highlight three benefits that incident reporting could provide to AI regulation.

1. Gain visibility of risks in real-world contexts

The UK Government has adopted an approach to AI regulation that "focuses on the context in which AI is deployed" [32].

This 'context-based approach' rightly recognises that it's not possible to understand the safety of an AI system solely by analysing it in isolation, as done in pre-deployment testing. The government and regulators must also monitor how AI is interacting in real-world contexts in often unanticipated ways.

To do this, it is essential that regulatory functions create a fast feedback loop that gives policymakers visibility of issues as they are emerging in context, so they can quickly update their understanding of risks and adapt their regulatory approach.

Incident reports help construct this feedback loop, and can capture risks that emerge from how AI is interacting within real-world systems, such as:

- **How AI is interacting with individuals.** Some safety issues arise from AI's effects on humans, rather than properties within AI systems themselves. For example, self-driving cars have been found to crash because human drivers have relied too heavily on them and stop paying attention to the road [12]. Here the safety issue arises because of the nature of the interaction between the AI model and individual users.

- **How AI is interacting with organisations.** Some safety issues arise from AI's integration by institutions – including in public services – because of existing problems in those institutions. For example, policing algorithms that do not have an inherent bias towards an ethnic minority have been found to develop one when police officers already over-police certain communities, and therefore generate data about that community that feeds back into the AI system. This phenomenon is known as 'emergent bias' [12].

- **How AI is interacting with other technical systems.** Some safety incidents emerge from how an AI model has an unanticipated impact on other technologies. For example, AI-generated text constitutes a large portion of Internet text in some underrepresented languages, which is being used to train translation systems, degrading the quality of text in those languages [13]. This has been described as "the digital equivalent of industrial chemicals making their way into drinking water" [14].

By identifying and investigating the causes of incidents like the above, the UK Government and regulators can help to better anticipate and mitigate similar incidents in future.

In this way, incident reports can also address some limitations of pre-deployment tests, such as those conducted by the AI Safety Institute (AISI), which may fail to identify some risks because of the small scales of those tests. This could include risks that are rare edge cases, the issues arise over longer periods of time (e.g. an AI companion might radicalise users only after gaining trust following several months of use), or issues that only occur in the aggregate of millions of users (e.g. an AI tutor could provide skewed representations of political debates to a specific demographic, which individually may not be significant, but in the aggregate

could lead to manipulation of opinion). In the words of Anthropic CEO Dario Amodei, "you have to deploy it to a million people before you discover some of the things it can do" [10].

These insights from real-world contexts can help functions such as AISI and CAIRF update their understanding of risks, monitor how well the UK's regulatory framework is addressing those risks, and course correct if required (e.g. by changing transparency reporting requirements if introduced with future legislation).

2. Coordinate crisis response

Incident reports can aid the fast resolution of critical safety issues, which will be vital in some cases where delays in addressing incidents can worsen their impact. Examples might include the unreliability of models that have been deployed in safety-critical domains or are single points of failure across society.

Importantly, this could include the UK Government's own use of AI in public services, which might include its application in critical national infrastructure. As the government increases its use of AI, incidents will become more likely, potentially requiring fast and coordinated responses by the UK Government, as well as a systematic approach to public transparency. Existing measures – such as the much-welcomed Algorithmic Transparency Recording Standard (ATRS) – do not require central or local government actors to report when an AI system has malfunctioned in a way that has caused a safety incident or near-miss, which could lead to slower responses and therefore greater harm.

A well-designed incident reporting regime can address this gap by flagging urgent issues directly and quickly to government stakeholders, and by helping them take swift action, including coordinating other stakeholders (such as AI companies) to respond. In turn, this can help to address the most urgent issues as quickly as possible. Equally, it can help CAIRF and other government actors know when to conduct deeper investigations to identify root causes and reassess risks, which may help to identify recommendations for AI companies and inform new regulatory mechanisms.

3. Identify early warnings of emerging risks

AI systems are enabling new services and even new industries, and in turn are enabling novel harms. Incidents can provide indicators about these novel harms, which may point to future, larger-scale harms. These warnings are sometimes referred to as 'canaries', defined as "early warning signs of potentially transformative impacts of AI" [34]. This is because:

- **Incidents can expose novel harms that could become larger-scale in future.** Some types of harms from AI will be novel or newly emerging, and may first be detectable in specific incidents. For example, AI has introduced new types of 'representational harms' that began to be understood through specific incidents, such as when an AI system labelled an African American couple as 'gorillas' [19]. Future examples could include incidents of AI systems behaving deceptively in novel ways in deployment, AI agents acting in ways that are misaligned with users' goals, AI companions that behave abusively or manipulatively (such as convincing users to commit suicide, as happened in 2023 [38]), or something as yet unforeseeable. By detecting incidents, we can identify and understand these new pathways for harm, and better anticipate how they might manifest on greater scales in future.

- **Incidents can expose underlying drivers that could cause future harm.** For example, safety incidents in self-driving cars – such as drivers paying less attention to the road – have helped to identify an underlying risk: placing too much trust in AI systems. This can result from underlying risk drivers such as automation bias (preferring automated solutions to human-managed systems). These drivers could have larger-scale consequences in safety-critical domains (e.g. military decision-making), and can be better anticipated with an understanding of previous incidents.

Individual incidents, or patterns at scale, may therefore provide researchers and policymakers with warnings as to how AI is developing and what mitigations may be needed.

What incident reports cannot do

Incident reports are not a panacea for all harms or safety issues arising from AI, and they have their own important limitations.

They do not offer a comprehensive dataset, because incident reports tend to be biased by whether stakeholders decide to report (though this is mitigated where there is a statutory duty), and in the case of AI it will be biased by whether stakeholders understand the role of AI in the incident they are reporting. As a result, they cannot be used to determine all potential safety issues, which are most common, how they compare from one company to another, or how their occurrence is changing over time.

Instead, incident databases can flag incidents that may require immediate response, and insights about potential problems that can be further investigated, understood, and mitigated. This role will be strengthened if a regime is embedded within sectors, with engaged, strongly incentivised and educated reporters.

Any incident reporting regime will also need to be integrated within a wider structure of risk reporting and evaluation within government, such as periodically feeding into risk registers, and supported by more thorough investigations.

The current state of AI incident reporting

In this section, we set out why key AI incidents are a gap in the UK's regulatory plans, with no mention in their 2023 or 2024 White Papers[4]. We find that UK regulators will only capture some incidents and do not have the appropriate systems in place to effectively action them.

We begin by setting out why existing processes are not capturing major AI incidents in a coherent or actionable way, representing a gap in the UK's regulatory plans. We then show what types of AI incidents are falling through this gap and why they are such a concern.

A gap in the UK's regulatory plans

There are currently two ways that reports of AI incidents can be potentially collected:

1. Reports to UK regulators when incidents meet the threshold of a relevant statutory duty;

2. Reports to voluntary databases managed by technology companies, civil society organisations, individuals, and the OECD.

However, we find that UK regulators' incident reporting is currently unclear and will very likely not capture many major incidents at the frontier of AI. We also find that voluntary efforts are disconnected from regulators and so do not enable effective responses, and are vulnerable to being closed down at short notice. We provide more detail of these gaps in the following sections.

UK regulators

On 30 April, UK regulators published updates covering their approach to regulating AI. Some regulators already have a mature incident reporting regime for their own sector (e.g. MHRA's 'Yellow Card' scheme for medicines) and are beginning to think about how reporting incidents from AI will occur in their sector (see Annex A for a full review of their discussion of incident reporting).

However, we find that this is the exception, with most regulators not mentioning incident reporting. And, UK regulators will only ever collect a minority of AI incidents, because there is not an AI regulator with the necessary legal powers to require reporting of incidents that are specific to AI systems such as highly capable foundation models.

Regulator	Reference to incident reporting in its April 2024 update?
Bank of England	No
Competition and Markets Authority (CMA)	No
Equality and Human Rights Commission (EHRC)	No

[4] The only mention comes not from the UK Government's own statements, but from stakeholders proposing this intervention in their consultation.

Financial Conduct Authority (FCA)	Unclear
Health and Safety Executive (HSE)	No
Legal Services Board (LSB)	No
Medicines and Healthcare products Regulatory Agency (MHRA)	Yes
Office for Nuclear Regulation (ONR)	No
Office for Standards in Education, Children's Services and Skills (Ofsted)	Unclear
Office of Communications (Ofcom)	Yes
Office of Gas and Electricity Markets (Ofgem)	No
Office of Qualifications and Examinations Regulation (Ofqual)	Unclear

While helpful, these updates demonstrate the lack of a coherent approach to AI incident reporting. We are particularly concerned about the following problems:

- **Many AI incidents will likely not be covered by UK regulators because there is no regulator focused on the frontier of AI.** This means that incidents relating to the development, design and deployment of highly capable foundation models are unlikely to be caught by sector regulators. As well as lacking a central authority to report to, there are not the necessary incentives in place to ensure would-be incident reporters (ranging from the frontier AI companies, their employees, and business customers, to citizens and civil society) will reliably report.

- **Cross-sectoral learning will be inhibited because some regulators will collect reports of AI incidents in their own siloed databases.** One of the benefits of AI incident reporting is that lessons can be learned about how AI systems work in context and their underlying drivers of risk (e.g. overreliance on an AI system), which will often be common to many sectors. Cross-sectoral learnings become possible when a central team is able to review either a database of incidents or another method of centralising reports, and is properly resourced to do so.

- **Some regulators may not collect reports of safety incidents involving AI at all,** because they do not have the remit, resources, or stakeholder incentives (such as statutory duties) to do so.

- **Incidents that do not meet a legal threshold, but would still be useful for learnings, will not be reported.** For example, while there is a legal requirement to report incidents of terrorism to MI5, there may not be a legal requirement to report some attempted misuses for terrorism (i.e. near misses), unconfirmed indicators of attempted misuse, or other forms of activity related to terrorism (e.g. creating terrorist propaganda) that may not require reporting to MI5 but still would benefit from being

recorded.

Voluntary databases

There are several systems that could be described as AI incident reporting regimes, ranging from formal civil society initiatives to the use of social media platforms such as Twitter/X. We document a range of these in the table below.

Incident database	Owner	No. reports
The AI Incident Database	Multi-stakeholder, open source, overseen by and with the largest contributions from civil society, such as the Digital Safety Research Institute (DSRI)	~3,500
Twitter / X	Elon Musk	Unknown
OECD AI Incidents Monitor	OECD expert group on AI incidents (supported by the Patrick J. McGovern Foundation)	~10,000
Direct user feedback on AI tools and interfaces	AI companies	Unknown
Open AI Status Page	OpenAI	~200
AI Vulnerability Database	Open source	~50
jailbreakchat.com	Alex Albert (closed in 2023)	Unknown

However, these voluntary efforts suffer from several important limitations for the UK context:

- **They are not connected to relevant UK authorities** that are best placed to take action, such as by conducting crisis response, conducting root cause analyses, updating national risk assessments, or identifying cross-sectoral learnings. Incident databases are insufficient on their own, and they will only provide the proper regulatory value if they are integrated into government and regulator action.

- **They are vulnerable to closure,** as they are funded often through commercial and philanthropic sources, or are voluntary efforts by individuals. For example, jailbreakchat.com, which is one of the most prominent databases of foundation model cybersecurity vulnerabilities, abruptly closed in 2023 after its founder was hired by AI company Anthropic.

- **There are insufficient incentives to ensure that stakeholders will report incidents,** with many initiatives receiving relatively few reports from relatively few reporters.

The current regimes of UK regulators and global voluntary efforts do not represent a coherent approach to incident reporting, and leaving a gap in the UK's regulatory plans.

Incident reporting in other safety-critical sectors

In other safety-critical industries, incident reporting regimes entail:

- **Carefully-designed incentives underpinned by legislation.** Reporting is underpinned by carefully-designed incentives, and is either required by law, or is voluntary (as in reporting incidents in medicines and medical devices to MHRA by doctors and nurses). Where voluntary incentives do not exist, where strong disincentives exist, or where the stakes of an individual incident are sufficiently high, a statutory duty will be necessary. In the UK, penalties for not reporting can include an unlimited fine and/or 6 months' imprisonment [18].

- **A simple system for reporting to a central database.** In most sectors, incident reporting is managed by a regulator or other government body. A relevant stakeholder – such as a manufacturer, an individual professional or employee, or a member of the public – will report the incident to the regulator or body via an online form, a 24 hour phone line, an email address, website, or a mobile app. Importantly, reporting needs to be made as simple and accessible to people as possible to support use and improve data collection.

- **Investigation by a central authority such as a regulator.** In healthcare, doctors and nurses produce millions of incident reports every year, which are described as "nuggets of gold" that identify hazards and trigger urgent action [6]. In the case of aviation and other transport incidents, a board of investigators will often conduct deep investigations to understand root causes.

The AI incidents that will fall through the gap

DSIT will need to consult with UK regulators and experts to provide greater certainty on what types of incidents will fall through this gap in regulation.

However, we are particularly concerned that the following types of incidents will be missed:

- **Vulnerabilities in highly capable foundation models**[5] that are not likely to be covered by cybersecurity vulnerability reporting, such as bias and discrimination, which could have significant implications if widely integrated in society (including in public services). Evidence suggests that model vulnerabilities are not currently being adequately reported, despite some AI companies' public commitments [28].

- **Incidents of misuse of AI systems,** e.g. detected use (or attempted use) in disinformation campaigns or biological weapon development, which are two particular types of misuse we are concerned about[6]. Incidents of misuse are sometimes reported by AI companies. For example, in May 2024, OpenAI announced that it had detected attempted use of its systems in disinformation campaigns by actors

[5] See the UK government's overview of model vulnerabilities: Emerging processes for frontier AI safety - GOV.UK

[6] See [40] and forthcoming work from CLTR on the impact of foundation models on biological misuse risks

originating from Russia, China, Israel and Iran [39]. Such voluntary reports may represent a small fraction of actual incidents, hindering our collective understanding of AI's risks and impacts.

- **Incidents in the context of novel types of relationships introduced by applications of frontier AI,** such as AI companions, tutors and therapists, where deep levels of trust combined with extensive personal data could lead to incidents of abuse, manipulation, radicalisation, or dangerous advice (e.g. to commit suicide[7]), some instances of which may not be covered by a UK regulator.

- **Incidents from the UK Government's own use of AI in public services,** a context where many important risks are likely to materialise, and transparency and accountability is essential for public trust. Many occurrences are important to record and investigate even if they do not meet a legal threshold for reporting to a regulator, both for public accountability and for responsible risk management. While the UK Government has laudably improved the transparency of its use of AI through the ATRS, this does not extend to transparency over safety incidents.

[7] See [38]

Three options for a UK AI incident reporting regime

The UK Government has several options for designing a coherent AI incident reporting regime. We set out three of those options below, and recommend the first option: a hybrid system.

Who are the relevant stakeholders?

Depending on the focus area of incident reporting (e.g. whether it is focused on foundation model misuse, or the UK Government's use of AI), there are a wide range of potential stakeholders that would report incidents and respond to them.

Reporters

- **UK Government Departments**: incidents arising from public sector use of AI could be reported to ensure accountability, resolution and understanding.
- **AI companies**: there could be a mandatory requirement for AI companies to disclose to a government authority when there has been a safety incident or near miss.
- **Professionals and employees**: some incidents may need to be reported by professionals accessing AI models in specific ways (e.g. via API), or employees working within companies that feel they need to blow the whistle.
- **Academia:** safety researchers may identify incidents in their research.
- **Civil society**: there could be a voluntary regime for civil society to provide reports to a central authority through their engagement with stakeholders and communities, which some civil society organisations are calling for [8].
- **The public**: the end users of AI could report their experiences with AI systems, which may be especially important in some domains (e.g. AI therapists).

Recipients

- **Existing regulators:** for example, the MHRA is likely to receive reports related to safety incidents arising from the use of AI in medical devices, but regulators' plans for this are still emerging.
- **The Central AI Risk Function:** to use in risk assessments, inform strategies, and coordinate crisis responses where necessary.
- **AI Safety Institute (AISI):** incidents could inform technical evaluations, safety tests and other safety research.
- **The Central Digital and Data Office (CDDO):** as the body responsible for government use of AI in public services, the CDDO could receive reports of incidents in government departments' use of AI.
- **The public:** there could be accountability to the public for when an AI system has failed by allowing the public to see incident reports - this may be particularly important for AI incidents in the public sector, where the public may have a greater demand for transparency, even if there are likely to be a need for some exemptions (e.g. incidents with national security implications).

Option	A simple system for reporters	Identifies cross-sector learnings + early warnings about AI	Captures full range of incidents
1. Hybrid (recommended)	✅	✅	✅
2. Centralised	❌	✅	❌
3. Decentralised	✅	❌	❌

Option 1 (recommended): A hybrid system: both UK regulators and DSIT (or a new AI regulator) collect AI incidents

Our recommended option is a hybrid system (visualised in Figure 1 below), where:

- DSIT (or a new AI regulator) collects AI incidents that are currently being missed by UK regulators in its own AI incident database, with effective incentives for reporters (possibly via new legislation); and
- UK regulators continue to collect incident reports involving AI (with improvements to guidance), which DSIT (or a new regulator) also collates into its central database.

In this model, DSIT (or a new regulator) would follow the following process:

1. Directly collect incident reports (e.g. about model vulnerabilities, such as bias and discrimination) from frontier AI companies, or a confidential report from an employee;
2. It would then convene and coordinate relevant stakeholders when necessary, ensuring they are aware of potentially cascading risks (e.g. notifying MHRA about the model's use in medical devices), and ensuring the relevant actors implement the necessary immediate mitigations;
3. Finally, DSIT (or a new regulator), possibly via externally commissioned researchers, would be able to follow up with a deeper investigation to understand root causes, such as flaws in design or deployment, and update any risk assessments accordingly.

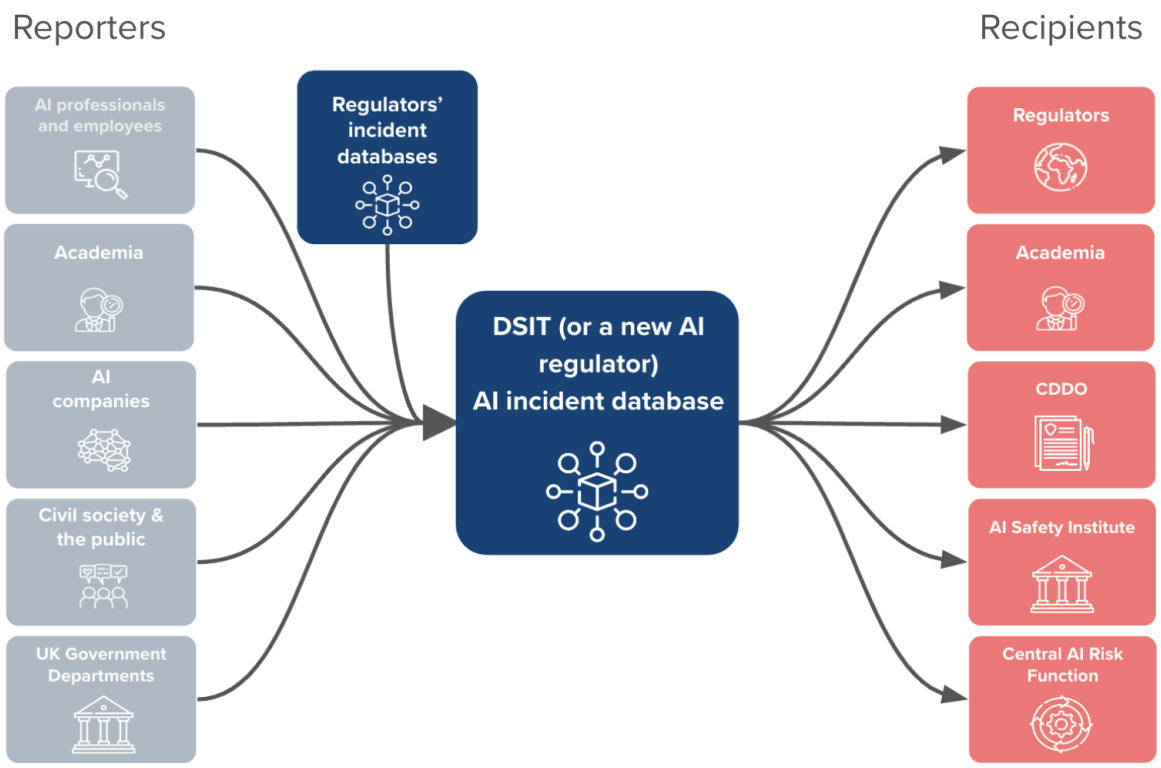

Figure 1: A diagram showing how a hybrid AI incident reporting system could work

- Advantages: The advantage of this model is that it would enable CAIRF and other stakeholders to identify cross-sector learnings by collecting the full range of AI incident reports in a single database, while also providing a single, central system for AI-specific stakeholders (e.g. frontier AI companies) to navigate and use to report.

- Disadvantages: The disadvantage is that this would require DSIT to build capacity to collect, triage, respond and investigate collected incidents. It may also require technical integration of DSIT's system with regulators, which would require financial investment in interoperable technical systems. However, these are both achievable goals with the appropriate investment, and could be partly funded by central government budgets such as the Shared Outcomes Fund.

How should DSIT respond to incidents?

An AI incident reporting regime would require the following capabilities:

1. **Triaging:** some incidents might require an urgent and timely response, and responders will need a way of assessing which incidents qualify for a fast response. There could be adjacent methods for AI companies to alert DSIT when there is a potential emergency (e.g. a dedicated email address, watched 24/7,) to aid with this process.

2. **Crisis response:** where incidents require an urgent response, DSIT may need to coordinate a whole-of-government response, or work with multiple regulators if there are cross-sectoral implications (such as risks of contagion or cascading

impacts on other systems). Existing crisis response processes should be expanded to include incident responses.

3. **Investigation:** some incidents will benefit from deeper investigation to understand root causes, which could involve an aggregate-level analysis of incidents that identifies trends or commonalities (as is typical in investigations into adverse reactions to drugs), or a deep investigation into a single incident (as is typical in aviation incidents). As the Center for Security and Emerging Technology paper notes, "[a]n investigative safety board can be useful for conducting root-cause analysis of significant AI incidents and providing feedback to help AI actors improve their design and development, enable policymakers to craft effective regulations, and educate the public on AI safety" [7].

DSIT should consider what functions must be conducted by their own staff, and which can be outsourced to the research community to reduce resourcing costs or provide external scrutiny. For example, it is likely that triaging and incident response will need to be performed by the UK Government (e.g. DSIT), but it is possible that investigation could be outsourced to civil society and academia, e.g. through commissioned research projects.

Option 2. A centralised system: DSIT (or a new regulator) owns an AI incident database

This option would involve DSIT creating a centralised AI incident database, which collects reports from various UK stakeholders and contexts – AI professionals and employees, AI companies, academia, civil society and the public, and UK Government departments – into a single database. This database can then be accessed by regulators, academia, CDDO, AISI, CAIRF, and potentially the UK public for risk assessments, incident response, investigations, and decisions on product warnings or recalls.

- Advantages: This model has the advantage of collecting AI-related incidents from UK stakeholders and contexts in a single place, facilitating cross-sectoral learning and a central response when needed.

- Disadvantages: Many relevant stakeholders are already legally required to report some incidents to regulators (e.g. MHRA), which could lead to a confusing system for businesses and other stakeholders to navigate. This could result in a duplication of reports and the likelihood that many incidents would be missed from this central database, because AI developers and other manufacturers would also be focused on reporting to regulators where they have a statutory duty.

Option 3. A decentralised system: UK regulators collect AI incidents independently, with no central system

This option would involve relying on existing incident reporting regimes operated by UK regulators, with improved guidance for when incidents involving AI need to be reported.

- Advantages: This model has the advantage of ensuring incidents are reported to the appropriate regulator, provides clarity to manufacturers and other stakeholders about

where to report.

- Disadvantages: First, there are likely to be gaps in coverage of AI incidents when collected solely through existing regulators. As we have established in the previous section ('The current state of AI incident reporting'), existing UK regulators are likely to miss important AI incidents, such as those from public sector use of AI (for which there might not be an appropriate statutory duty on departments or reporting frameworks), and those specific to foundation models (e.g. model vulnerabilities or misuse), for which there is no specific UK regulator. Second, this would be a fragmented system where AI incident reports are dispersed across multiple databases, making it hard for a body such as CAIRF to identify cross-sectoral learnings.

Recommended next steps for the UK Government

Recommendation 1: Create a system for the UK government to report incidents in its own use of AI in public services

DSIT should prioritise creating a system to acquire oversight of incidents in government departments' use of AI, as this is a low hanging fruit that could have a potentially significant impact on crisis response, lesson learning, and public trust.

This could begin with a limited private system to ensure that relevant bodies such as the Central AI Risk Function (CAIRF) or Central Digital and Data Office (CDDO) can see where safety incidents are occuring in the UK Government and possibly local government too. This will require a definition of the scope of incidents requiring reporting, which may need to be a lower threshold than typical legislative incident reporting requirements to ensure effective monitoring. It would also require a system for collecting reports, which could be as simple as an email address or digital form.

Eventually, the UK Government should consider expanding the Algorithmic Transparency Recording Standard (ATRS) to include a framework for reporting public sector AI incidents. The ATRS is a welcome contribution to the transparency of the use of AI in public services, and the next iteration should include a framework for incident reporting.

DSIT should work to ensure that there are no unnecessary national security exemptions, and where those exemptions are necessary it should work to ensure there is still a strong culture of incident reporting, possibly at a higher level of classification.

Recommendation 2: Confirm where there are the most concerning AI incident reporting gaps by commissioning UK regulators and experts

We have provided proposals of which incidents will be missed by UK regulators, but this requires further confirmation by UK regulators and a greater range of experts to ensure it is accurate and comprehensive.

To do this, DSIT should commission UK regulators and AI policy and technical experts to which incidents involving AI will not be reported, clarifying in particular:

- What gaps in the current system are most urgent to address (e.g. misuse, model vulnerabilities, or risks not covered by existing UK regulators);
- Whether a central database of AI incidents is necessary, and how this should be used;
- What best practices exist for responding to urgent incidents that could be applied to incidents involving AI;
- How to create incentives for stakeholders to report AI-related safety incidents, including considerations of whistleblower protections and the need for new statutory duties.

Relevant stakeholders that should be consulted on these questions include:

- **AI companies,** as industry buy-in is likely to be essential to the success of a pilot and future regime.

- **Civil society actors** seeking to gather incident reports on behalf of, or in collaboration with, the UK communities they serve, as these are "often the first port of call for those who have experienced day-to-day harms" [8]

- **Managers of major AI incident reporting systems,** such as those working at the AI Incident Database and OECD AI Incidents Monitor.

- **Key figures at existing UK incident reporting regimes**, such as those working at MHRA's YellowCard scheme and at the Air Accidents Investigation Branch.

- **Academics and policy experts**, such as Joe O'Brien, Shaun Ee, Zoe Williams, authors of Deployment Corrections: An incident response framework for frontier AI models, Ren Bin, Lee Dixon and Heather Frase, authors of An Argument for Hybrid AI Incident Reporting Lessons Learned from Other Incident Reporting Systems, and Kevin Wei and Lennart Heim, authors of Designing Incident Reporting Systems for Harms from AI (forthcoming).

Recommendation 3: Build capacity within DSIT to monitor, investigate and respond to incidents, possibly including the creation of a pilot AI incident database

Incident reports are only as valuable as the subsequent analyses and responses to them. As a result, DSIT should build capacity to ensure that it is able to triage, respond, and investigate incidents that are reported. This capability could sit within the Central AI Risk Function (CAIRF), which is already building capacity for similar functions.

DSIT should also consider scoping a pilot incident reporting database, with a tool for reporters to submit incident reports into a central database that CAIRF can use to triage, respond to, and investigate incidents. This should be based on the findings from the regulator and expert consultation (Recommendation 2), and could include the development of a prototype technical system that is robust enough that it can be scaled up for a future, more comprehensive regime if appropriate. This database could potentially be funded by AISI or a central government funding pot, such as the Shared Outcomes Fund.

The pilot database should ideally focus on a category of incident where:

- Reporters are incentivised, and can be engaged and educated, to provide the intended reports

- The incidents are particularly important to capture, e.g. where they represent major risks to which the UK is vulnerable

- The incidents will otherwise be missed by UK regulators

Alternatively, the pilot could seek to work as a 'hub' that collects AI-related incident reports from existing incident reporting systems, such as the Air Accident Investigation Branch, the MHRA incident report systems, Information Commissioner Officer's system, either manually (via regular reports) or via technical automation.

We have provided an indicative set of deliverables and estimated costs for a technical pilot, though this could conceivably be scaled down or up following the outcomes of the consultation and further analysis by DSIT.

Deliverable	Description	Estimated Cost
Discovery	Consultation with stakeholders and users to understand needs of the system	£350,000
Alpha	A minimum viable reporting system to test with users	£500,000
Beta	A more developed system that implements lessons from Alpha testing and delivers a robust, reliable system that can be scaled up in future	£1,000,000
Project Management	A supplier to manage the delivery of the project over one year, and government staff to manage suppliers	£750,000
Community Engagement	Engagement, outreach and education initiatives to ensure reports are provided	£500,000
Legal	Legal support to understand risks of data collection and sharing	£150,000
Evaluation	Evaluation to ensure lessons are learned for a future scale-up	£250,000
Total		**£3,500,000**

Acknowledgements

I am very grateful to a range of people for discussions on this subject, and in some cases reviews of earlier drafts, including Dr. Jess Whittlestone, Gabby Overödder, Sophie Dannreuther, Ben Robinson, Merlin Stein, Heather Frase, Kevin Wei, Kevin Paeth, Sean McGregor, Shaun Ee, and Ren Bin Lee Dixon. I'm also very grateful to Luke Dawes, who provided research support and copyedited the final report.

Annex A: UK regulators' reference to incident reporting in their 30 April updates

Regulator	Reference to incident reporting?	Details
Bank of England	No	The Bank's update focuses predominantly on compliance with existing legislation and regulation re: data privacy, security, and financial risk.
Competition and Markets Authority (CMA)	No	CMA "appointed 9 Digital Experts [who] have diverse areas of expertise and collectively they possess direct experience working within or alongside large technology firms, **insight into building new regulatory functions** as well as technical expertise relating to digital technologies, including AI."
Equality and Human Rights Commission (EHRC)	No	No mention of incident reporting in the update.
Financial Conduct Authority (FCA)	Unclear	FCA refers to reporting risks, but not clear if this involves reporting incidents to the FCA. The SYSC sourcebook "contains a range of more specific provisions on systems and controls and firms' governance processes and accountability arrangements. In particular, under SYSC 4.1.1R '[a] firm must have robust governance arrangements, which include a clear organisational structure with well defined, transparent and consistent lines of responsibility, effective processes to identify, manage, monitor, and report the risks it is or might be exposed to, and internal control mechanisms, including sound administrative and accounting procedures and effective control and safeguard arrangements for information processing systems'." [3.39]
Health and Safety Executive (HSE)	No	No reference to incident reporting (though AI-related issues might eventually be captured under the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations (RIDDOR)).

Legal Services Board (LSB)	No	No reference to incident reporting
Medicines and Healthcare products Regulatory Agency (MHRA)	Yes	For capturing incidents or safety issues with AI as a medical device (or AIaMD), "manufacturers are expected to have established vigilance systems and processes for continuous monitoring and reporting of safety incidents directly to the MHRA." (p.6) This wouldn't be specific to AI-related products, however. "The MHRA Yellow Card scheme enables anyone to report concerns to the MHRA about a medicine or device, including one incorporating AI. Current regulations also place legal requirements for manufacturers to report incidents to the agency, and these obligations will be strengthened for medical devices, by new regulations which we aim to put in place by the summer." (p.9)
Office for Nuclear Regulation (ONR)	No	There is reference to a "valuable source of intelligence [being] the raising of concerns by industry professionals and third parties during routine regulatory interactions." However, there is no reference to how AI will be handled in a more formal, properly incentivised reporting regime. Their existing whistleblower scheme would provide protection to people reporting issues in areas that ONR regulates directly, but this is not referred to.
Office for Standards in Education, Children's Services and Skills (Ofsted)	Unclear	Expects to use existing regulatory functions to investigate AI-facilitated incidents. "We will... continue to manage concerns and complaints through our existing complaints procedure."
Office of Communications (Ofcom)	Yes	• "Work we will do in 2024/24: Use our regulatory powers to investigate security compromise reports from services in scope of the TSA and identify if these have been facilitated by AI risks." (p16) • "The Online Safety Act 2023 (OSA) requires in-scope services to assess the risk of users encountering illegal or harmful content on their services, including risks associated with the deployment of AI-driven technology, for instance certain recommender systems (which suggest or recommend additional products to consumers). Services are then required to take proportionate measures to mitigate and manage the risks they identify. Priority illegal content, which the OSA requires services to take proportionate measures to prevent people from encountering, could include AI-generated child sexual abuse material. As part of work under the OSA, we could also recommend the use of AI-enabled safety technology within our Codes of Practice. An example of AI-enabled

		safety technology would be automated content moderation. This refers to the automatic analysis of text, images, and videos to detect and remove harmful content (e.g., hate speech, harassment, deepfakes and CSAM material). AI technology typically supports human moderators by filtering vast amounts of data quickly, allowing for more efficient and scalable moderation processes, ultimately creating safer online environments." (p6, Point 2.2)
Office of Gas and Electricity Markets (Ofgem)	No	No reference to incident reporting
Office of Qualifications and Examinations Regulation (Ofqual)	Unclear	Ofqual refers to "implementation of AI specific categories to be used by awarding organisations when reporting malpractice to Ofqual", which may involve incident reporting.

References

[1] Schuett J., Dreksler N., Anderljung M., Heim L., Bluemke E., et al. (2023). *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion*. Centre for the Governance of AI. Retrieved from https://www.governance.ai/research-paper/towards-best-practices-in-agi-safety-and-governance (accessed 2024-06-04).

[2] (2023). *Emerging Processes for Frontier AI Safety*. UK Department of Science, Innovation & Technology. Retrieved from https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (accessed 2024-06-04)

[3] (2023) *Keeping an eye on AI.* Ada Lovelace Institute. Retrieved from https://www.adalovelaceinstitute.org/wp-content/uploads/2023/09/ALI_Keeping-an-eye-on-AI-2023.pdf (accessed 2024-06-04)

[4] Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, et al. (2023). *Model evaluation for extreme risks.* arXiv. Retrieved from https://arxiv.org/pdf/2305.15324 (accessed 2024-06-04)

[5] Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., et al (2023). *Frontier AI regulation: Managing emerging risks to public safety*. Retrieved from https://arxiv.org/pdf/2307.03718 (accessed 2024-06-04)

[6] Cuong Pham, J., Girard T., Pronovost, P. (2013) *What to do with Healthcare Incident Reporting Systems*. J. Public Health Res. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147750/ (accessed 2024-06-04)

[7] Dixon, RBL., Frase, H. (2024) *An Argument for Hybrid AI Incident Reporting*. Center for Security and Emerging Technology. Retrieved from https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/ (accessed 2024-06-04)

[8] Coldicutt R. (2023). *Putting the lid on Pandora's Box: how community power shapes AI.* Promising Trouble. Retrieved from https://www.promisingtrouble.net/blog/2023-12-pandoras-box (accessed 2024-06-04)

[9] Davis, M., Birtwistle, M. (2023). *Regulating AI in the UK.* Retrieved from https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/ (accessed 2024-06-04)

[10] De Vynck, G., Lerman, R., Tiku, N. (2023) *Microsoft's AI chatbot is going off the rails*. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2023/02/16/microsoft-bing-ai-chatbot-sydney/ (accessed 2024-06-04)

[11] Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans.* United Kingdom: Penguin Books Limited.

[12] Dobbe, R. I. J. (2022). *System Safety and Artificial Intelligence.* arXiv. Retrieved from https://arxiv.org/abs/2202.09292 (accessed 2024-06-05)

[13] Thompson, B., Dhaliwal, M. P., Frisch, P., Domhan, T., Federico, M. (2024). *A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism*. Retrieved from https://arxiv.org/abs/2401.05749 (accessed 2024-06-04)

[14] Clark, J. *Import AI 357: Facebook's open source AGI plan; Google beats humans at geometry problems; and Intel makes its GPUs better.* (2024). Substack. Retrieved from https://importai.substack.com/p/import-ai-357-facebooks-open-source (accessed 2024-06-04)

[15] Neff, G., & Nagy, P. (2016). *Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay*. International Journal Of Communication, 10, 17. Retrieved from http://ijoc.org/index.php/ijoc/article/view/6277/1804

[16] Dobbe, R., Dean, S., Gilbert, T., and Kohli, N. (2018). *A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics*. arXiv. Retrieved from https://arxiv.org/pdf/1807.00553 (accessed 2024-06-04)

[17] (2018). *How we investigate*. UK Government, Air Accidents Investigation Branch. Retrieved from https://www.gov.uk/government/publications/how-we-investigate/how-we-investigate (accessed 2024-06-04)

[18]. (2024). *Medical devices: the regulations and how we enforce them*. UK Government, Medicines and Healthcare products Regulatory Agency. Retrieved from https://www.gov.uk/government/publications/report-a-non-compliant-medical-device-enforcement-process/how-mhra-ensures-the-safety-and-quality-of-medical-devices (accessed 2024-06-04)

[19] (2015) *Google Photo App Labels Black Couple 'Gorillas'*. Sky News US. Retrieved from https://news.sky.com/story/google-photo-app-labels-black-couple-gorillas-10353994 (accessed 2024-06-04)

[20] Ullman, S., Saunders, D., *Online translators are sexist – here's how we gave them a little gender sensitivity training.* The Conversation. Retrieved from https://theconversation.com/online-translators-are-sexist-heres-how-we-gave-them-a-little-gender-sensitivity-training-157846 (accessed 2024-06-04)

[21] Koh, R. (2023). *A list of AI-generated Barbies from 'every country' gets blasted on Twitter for blatant racism and endless cultural inaccuracies*. Business Insider. Retrieved from https://www.businessinsider.com/ai-generated-barbie-every-country-criticism-internet-midjourney-racism-2023-7 (accessed 2024-06-05)

[22] Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., et al. (2021). *Filling gaps in trustworthy development of AI.* arXiv. Retrieved from https://arxiv.org/pdf/2112.07773.pdf (accessed 2024-06-05)

[23] O'Brien, J., Ee, S., Williams, Z. (2023). *Deployment corrections An incident response framework for frontier AI models.* Institute for AI Policy and Strategy (via arXiv). Retrieved from https://arxiv.org/pdf/2310.00328.pdf (accessed 2024-06-05)

[24] (2023) *Keeping an eye on AI.* Ada Lovelace Institute. Retrieved from https://www.adalovelaceinstitute.org/wp-content/uploads/2023/09/ALI_Keeping-an-eye-on-AI-2023.pdf (accessed 2024-06-04)

[25] Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., et al (2023). *Frontier AI regulation: Managing emerging risks to public safety*. Retrieved from https://arxiv.org/pdf/2307.03718 (accessed 2024-06-04)

[26] McGregor, S. (2020). *Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database.* arXiv. Retrieved from https://arxiv.org/abs/2011.08512 (accessed 2024-06-05)

[27] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield G., et al. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.* Retrieved from https://arxiv.org/pdf/2004.07213.pdf (accessed 2024-06-05)

[28] Jones, A. (2024). *How are AI companies doing with their voluntary commitments on vulnerability reporting?* Self-published blog post. Retrieved from https://adamjones.me/blog/ai-vulnerability-reporting/ (accessed 2024-06-05)

[29] (2023). *A pro-innovation approach to AI regulation*. UK Government, Department for Science, Innovation & Technology. Retrieved from https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#fn:66 (accessed 2024-06-05)

[30] (2023). OECD. *Stocktaking for the development of an AI incident definition.* Retrieved from https://www.oecd-ilibrary.org/science-and-technology/stocktaking-for-the-development-of-an-ai-incident-definition_c323ac71-en (accessed 2024-06-05)

[31] *Report a problem with a medicine or medical device.* UK Government. Retrieved from https://www.gov.uk/report-problem-medicine-medical-device (accessed 2024-06-05)

[32] (2023). *A pro-innovation approach to AI regulation*. UK Government, Department for Science, Innovation & Technology. Retrieved from https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#fn:66 (accessed 2024-06-05)

[33] AI Incident Database. Accessible at https://incidentdatabase.ai/ (accessed 2024-06-05)

[34] Zoe Cremer, Z., Whittlestone, J. (2021). *Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI.* Retrieved from https://www.repository.cam.ac.uk/items/6e654b0b-26e0-48c1-8a2f-16782c73bc82 (accessed 2024-06-05)

[35] *AI Safety Institute: Systemic AI Safety Fast Grants.* Accessible at https://www.aisi.gov.uk/grants (accessed 2024-06-05)

[36] *OECD AI Incidents Monitor.* OECD.AI Policy Observatory. Retrieved from https://oecd.ai/en/incidents?search_terms=%5B%5D&and_condition=false&from_date=2014-01-01&to_date=2024-05-30&properties_config=%7B%22principles%22:%5B%5D,%22industries%22:%5B%5D,%22harm_types%22:%5B%5D,%22harm_levels%22:%5B%5D,%22harmed_entities%22:%5B%5D%7D&only_threats=false&order_by=date&num_results=20 (accessed 2024-06-05)

[37] Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Henricks, L. A., Mateos-Garcia, J., et al. (2023). *Sociotechnical Safety Evaluation of Generative AI Systems*. Google DeepMind via arXiv. Retrieved from https://arxiv.org/pdf/2310.11986 (accessed 2024-06-05)

[38] El Atillah, I. (2023). *Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change.* EuroNews. Retrieved from https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate- (accessed 2024-06-06)

[39] (2024). Disrupting deceptive uses of AI by covert influence operations. OpenAI. Retrieved from https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/ (accessed 2024-06-06)

[40] Shaffer Shane, T. (2024). *The near-term impact of AI on disinformation.* The Centre for Long-term Resilience. Retrieved from https://www.longtermresilience.org/post/the-near-term-impact-of-ai-on-disinformation (accessed 2024-06-06)

[41] Cheng, D., McKernon, E. (2024). *2024 State of the AI Regulatory Landscape.* Convergence Analysis. Retrieved from https://www.convergenceanalysis.org/ai-regulatory-landscape/home (accessed 2024-06-10)

[42] Heikkilä, M. (2022). *Dutch scandal serves as a warning for Europe over risks of using algorithms.* Retrieved from: https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/