Pragmatic guidance on information risks and security controls for users of ChatGPT

By Gary Hinson,
CEO of IsecT Ltd.

26th April 2023

Executive summary

I have identified at least twenty-six threats capable of causing a variety of information security and privacy incidents to ChatGPT users. Some early adopters of ChatGPT are already suffering serious incidents, hence it would be unwise to discount or ignore the risks. This guideline discusses both general purpose and specific information security controls to mitigate unacceptable risks relating to ChatGPT and similar 'generative AI systems'.

Contents

Introduction

Interest in **A**rtificial **I**ntelligence/**M**achine **L**earning and related technologies for 'deep learning' has been steadily growing over several years as the field emerged from academic research labs through product development into commercial manufacturing and sale of various goods and services. Today we have quite a variety of smart systems on offer, for various definitions of 'smart' and 'system' – some rather dumb with poor engineering, in reality, while others are capable of performing certain activities better, faster and more reliably than us humans. The technology is advancing rapidly.

Suddenly, out of the blue towards the end of 2022, AI/ML really hit the big time with OpenAI's release of the ChatGPT **N**atural **L**anguage **P**rocessing 'generative AI' system onto the Web. ChatGPT seized our collective imaginations, inspiring a flurry of news articles, commentaries and guidelines, some of which were produced by ChatGPT itself – self-serving perhaps but equally a powerful demonstration of this exciting technology.

There's space over here for notes

Following rapidly along, there have been issues relating to ChatGPT, including a few serious security and privacy incidents already – for some a bad omen and yet, paradoxically, an *expected* outcome for innovative technology as complex as this.

As often happens with a new market segment, innovation is on overdrive with numerous small players leaping headlong into the field alongside the bigger, well-established corporations, hoping to seize market share and grow sufficiently quickly either to predominate or become juicy/valuable takeover targets. Already we are seeing new companies offering customised NLP services based on ChatGPT's **A**pplication **P**rogramming **I**nterfaces or other platforms, plus parallel, competing NLP developments.

Audience, scope and aims

This guideline concerns the information risks associated with ChatGPT specifically in the context of organisations whose workers are eagerly – if naïvely – using and discovering the pros and cons of this exciting new technology. The audiences I have in mind are information risk and security specialists ranging from **C**hief **I**nformation **S**ecurity **O**fficers down, plus privacy officers, IT auditors, IT professionals in general, and 'early adopters' or 'power users' – those individuals who just *love* exploring and exploiting new technologies.

Please interpret 'ChatGPT' in this guideline broadly to mean any 'generative' AI/ML/NLP systems and services, not necessarily or specifically OpenAI's ChatGPT. I have no axe to grind, no particular reason to think or suggest that ChatGPT is any more or less risky than any other. ChatGPT is simply an obvious, topical and convenient *example* of a mushrooming category of information systems and services – the first one to feature widely in conventional and social media around the globe.

I come in peace, Earthman!

I should point out that there are also information risks of significant concern to OpenAI and other AI/ML/NLP innovators and providers, and to society in general, arising from or associated with the technology. However, **they are out of scope for *this* guideline**.

Furthermore, given my intended audiences and professional background, the guideline primarily concerns the downside risks while largely ignoring the upside opportunities in this space. Generative AI/ML/NLP, deep learning, expert systems, robotics and other assorted technological developments have tremendous potential to disrupt existing markets, in much the same way that the Internet and World Wide Web, computers, internal combustion and jet engines, the printing press and the wheel have done in the past – the latest in a long line of revolutionary developments. Within fields of specialist expertise such as information risk and security, there is *plenty* of scope for **A**ugmented **I**ntelligence, merging the best qualities of carbon and silicon brains to achieve new levels of capability and performance – and you can bet our black-hat adversaries are busily concocting cunning **A**utomated **I**ntrusion schemes just as fast as we build and implement our AI-enhanced security systems to lock them out.

About the author

I am Dr Gary Hinson PhD MBA, an information security specialist with a lifelong interest in the human and business aspects of both protecting and exploiting information.

Originally a research scientist, my professional career stretches back to the mid-1980s as a practitioner, manager and consultant in the fields of IT system administration, information security and IT auditing for multinationals in several industries.

For more than a decade, I was nose-to-the-grindstone writing creative security awareness materials for "NoticeBored", an innovative subscription service.

These days, I research, write, debate, consult, audit, mentor and teach, mostly on ISO27k - the ISO/IEC 27000 information risk and security management standards - and information security metrics.

By all means browse my websites for more:

- IsecT.com concerns my freelancing/consulting business
- SecAwareBlog.blogspot.com is my blog
- SecAware.com is my virtual shopfront for policies, ISO27k templates and awareness content
- ISO27001security.com has information on the ISO27k standards, plus free templates and tools
- SecurityMetametrics.com offers guidance on the P.R.A.G.M.A.T.I.C. security metrics method
- linkedin.com/in/garyhinson/ for my professional profile and recent debates.

Please contact me by email (Gary@isect.com) or through any of my websites. I'd love to talk with you, your colleagues and senior management about the issues arising from this paper, perhaps even help you manage and review your AI/ML-related information risks, prepare custom security policies and awareness materials, or whatever. If I've sparked your imagination with this guideline, let's chat!

Acknowledgement

ChatGPT (GPT-4) itself sparked some of the ideas in this guideline. Likewise Google and countless websites provided additional information on the topic. However, the actual words are mine and, ultimately, I accept personal responsibility for the content, warts-and-all. Feedback *please*.

Identifying and assessing information risks

Securing information rationally and appropriately involves identifying, understanding and addressing the associated information risks … so, what *are* the risks associated with ChatGPT for organisations whose workers are using it?

Lacking any knowledge of your organisation's business situation, I cannot identify or evaluate *your* particular information risks but I invite you to consider the following *generic* risks.

The approach I am taking is a tried-and-trusted technique to elaborate on information risks by exploring the main risk factors (specifically, the threats, vulnerabilities and impacts) using a variety of known (reported) and unknown (yet credible) scenarios to illustrate the range.

I *urge* you to treat this generic ChatGPT risk analysis cautiously and consider it critically from your own perspective. It may be materially wrong, out of date, incomplete, misleading and inappropriate to your context.

Risk analysis is a risky business!

Threats to, from or involving ChatGPT

Threats are external actors, activities, events *etc*. potentially interacting or impinging in a negative way on 'the system of concern'. Here that means not just the ChatGPT servers, nor the clients and other computing and networking devices involved in delivering the services via the Internet to users, but the diffuse collection of ICT hardware and software technologies, activities and people – not least the workers using ChatGPT and our organisations.

Who or what Out There could cause trouble, when and how?

In fact, strictly speaking 'the system' encompasses the wider supply chain meaning various service and technology providers, the software designers and developers, the NLP models and the information used to train them, plus the commercial markets, societal aspects and more beside. However, *this* guideline is primarily concerned with the ChatGPT service itself.

Threatening situations and threat agents of relevance to ChatGPT include:

1) **Accusations**: as the world comes to terms with ChatGPT, several individuals or organisations have been *accused* of using ChatGPT to generate content they have claimed as their own product, implying that they are being unethical, unfair, incompetent or lazy. Since it is difficult to demonstrate let alone prove otherwise, reputations and trust may be harmed by such accusations, whether true or false;

2) **Changes**: given the immaturity of ChatGPT, frequent changes to the systems/services are to be expected such as:
   - ChatGPT's internal models and capabilities - it is, after all, a youngster growing up, playing and learning new tricks, perhaps even testing the limits and getting up to mischief;
   - The AI/ML/NLP computing technologies and techniques underpinning ChatGPT;
   - The extended ChatGPT system as a whole, from the servers to the clients and users;
   - Various controls built-in to ChatGPT and the associated processes;
   - The business and social context *e.g.* OpenAI's responses to commercial competition;
   - Laws, regulations including qualified legal opinions on the interpretation and application of existing legislation such as privacy/GDPR in relation to ChatGPT itself plus how it is being used and controlled;
   - Guidelines, codes of practice, advisories and expectations of individuals, organisations and society in general;
   - The information and other risks *e.g.* newly discovered vulnerabilities, novel modes of attack, evolving threats and impacts;
   - Knowledge and experience, of course.

3) **Competition**: most organisations face competition from pre-existing rivals or new entrants hoping, for instance, to gain valuable product, technology or process knowledge from others. Intense commercial pressure to exploit promising new markets such as AI/ML/NLP (similar to a gold rush land-grab) is of direct concern to OpenAI and related organisations – a supply chain security threat – plus more widely anyone who make be impacted by the premature market release of such systems. Intense competition for the emerging AI/ML market could lead to ChatGPT failing, being taken over/merged *etc*. leading to changes in service availability and details for customers;

OpenAI and ChatGPT are not so special as to be immune to external pressures

4) **Criminals**: the global criminal fraternity includes lone operators, diffuse and dynamic collaborations, and organised groups/gangs, some of which are identified separately in this list of threats. Most are financially motivated, actively opening up and exploiting commercial opportunities regardless of the ethical and legal controls constraining the rest of society. Some criminal groups are tolerated, perhaps even supported, by government agencies or other backers. Naturally, they *all* strive to remain incognito, hiding behind 'false fronts' or infiltrating legitimate organisations. Given our limited

experience and knowledge, it is unclear how criminals might yet take advantage of ChatGPT, maybe exploiting weaknesses in the technologies or the financial arrangements and associated controls, coercing insiders, holding companies, systems, services or data to ransom or in some other way compromising business processes – or something else entirely. Many of what we consider to be 'vulnerabilities' qualify as 'opportunities' for criminals to create mischief and take advantage;

5) **Distraction/diversion**: exploring and figuring out how to make good use of any new technology, especially one as enticing and complex as ChatGPT, takes time and attention. Judging by the publicity and ChatGPT's occasional performance and capacity issues, millions of people have been trying out ChatGPT;

Is reading and thinking about this guideline a worthwhile investment of your time and attention?

6) **Environmental concerns**: increasingly urgent and insistent global pressure to tackle climate change by limiting the production of 'greenhouse gases' is already affecting substantial information services provided from large data centres with racks stuffed full of powerful, energy-hungry IT systems. ChatGPT is *believed* to be using hundreds or thousands of servers in a number of cloud-based data centres around the world, implying *substantial* energy consumption;

7) **Errors and omissions**: mistakes and accidents are a significant threat to information and information systems, not so much due to their severity (most are inconsequential) as their frequency. Neither human beings not automatons are immune to this threat, for instance the designers and developers of ChatGPT may have made inept or inappropriate decisions leading to design flaws and bugs in the system itself, or the underlying technologies on which it depends. Likewise, ChatGPT system administrators and users are prone to errors, *even if* we are well trained, alert and make the effort to pay close attention to what we are doing. In practice, errors are *certain* to occur at some point, especially given the novelty of ChatGPT. Unfortunately, however, it is difficult to predict, yet, just how serious or numerous they might be, or to determine what kinds of issues or incidents might be materially damaging;

An open-source library bug caused the inappropriate disclosure of ChatGPT users' histories in March

8) **Extremists:** this threat category includes terrorists, pressure groups, anarchists, lobbyists, crackpots and more, sharing three particular characteristics: malicious intent, determination and resources. As with hackers, it is hard to predict whether or which extremists might take an interest in ChatGPT, nor how that might play out. To be crystal clear, I personally have *no* knowledge of *any* such involvement at this point: this is merely idle conjecture of a cynical and somewhat paranoid information risk and security professional!;

9) **Fraudsters and scammers**: insiders, outsiders and, rarely, both together (conspirators) who commit fraud represent an insidious threat. They typically take time to gather knowledge about security arrangements, build (misplaced) trust, plan and prepare to act whilst deliberately concealing

their activities – for instance, technical support scammers using ChatGPT to boost their abilities to convince and mislead naïve IT users;

10) **Hackers and crackers**, particularly of course the sinister black-hat variety within or at the fringes of the criminal fraternity, using ChatGPT to identify and exploit vulnerable IT systems and processes. Even white- and grey-hats are of some concern, such as hobbyists and amateur developers experimenting with the AI/ML/NLP technologies without necessarily appreciating the need or having the skills to keep them under control, legally, ethically and practically;

11) **Inadequate resources**: massive global interest in ChatGPT means that the IT systems are heavily loaded, rate-limiting performance despite the flexibility of the cloud infrastructure. Competition for the limited existing pool of analysts, programmers, testers *etc*. with skills and experience in this field is likely to drive up salaries and fees, at least until the educational and training suppliers catch up;

Maybe automated systems can help plug the skills gap …

12) **Inappropriate use**: looking for advice on how to hack, commit fraud, develop malware, snoop on a partner, rob an online bank, fake a sick note *etc*.? ChatGPT qualifies as 'dual-use' technology of utility to baddies as well as goodies. Creative thinking can evade some of the ethical controls built-in to ChatGPT, but then search engines and the hacker/criminal underground scenes already circulate similar information through the global Internet, and there are books and films and …;

'Security by obscurity' is a fragile control

13) **Insiders**: this threat group comprises malicious, careless or inept employees (staff *and* managers, mind!) plus pseudo employees such as interns, temps, contractors and consultants, working for OpenAI or elsewhere in the supply chain, including our own organisations;

14) **Knowledge limitations**: aside form issues with the processing and synthesis of new information, computer systems are ultimately limited by the quality and nature of the data fed into or available to them, meaning the training data for NLP systems. ChatGPT's training phase ended in September 2021, hence it is largely ignorant of more recent developments or news, while we can barely guess at the nature and extent of its original data sources. In specialist areas such as security metrics, it does not have encyclopaedic knowledge, frequently offering the same handful of metrics examples in its answers, rather like an undergraduate regurgitating a few key learning points from class. Even if it fails to comprehend the subject of a question, it usually answers in a confident style with whatever information seems relevant, similar to a job interviewee;

15) **Malware**: aside from conventional 'malicious software' we have suffered for decades and more recent developments using remotely controlled and updatable ransomware and other nasty variants, potentially we now face the additional threat of AI/ML/NLP-based smart malware and **A**dvanced **P**ersistent **T**hreats. Black hats are doubtless already using ChatGPT to generate more stealthy, invasive and discreet strains of malware, such as highly polymorphic code;

16) **Misuse or manipulation**: ChatGPT may be misused or manipulated by 'bad actors' to generate inappropriate or misleading content. This may involve gaming the system, for example using verbal or logical techniques to trick ChatGPT into disclosing its inner secrets. Adversaries such as unethical competitors of user organisations may conceivably take advantage of workers' naïve interest in ChatGPT to mount phishing or other more creative attacks *e.g.* persuading workers to accept/trust inappropriate security or commercial advice from ChatGPT;

17) **Natural events** such as fires, floods, storms, accidents with excavators, hungry rodents and pandemics – a reminder not to ignore the physical world when dealing with computer systems. Although such threats may not actively target or specifically attack it, they may impact the ChatGPT service provision in some way – a tornado or earthquake, for instance, that damages a critical data centre or its power feed and data cables, or an outbreak of infectious disease, strike or resignations leaving the data centre, call centre or security operations centre desperately short-staffed;

18) **Officialdom** includes various legal and regulatory authorities plus litigious private companies investigating and enforcing compliance with privacy, intellectual property, trade secret, safety and other information-related laws, regulations and contractual terms. The ChatGPT service *could* potentially be shut down or blocked for a period or permanently by a country's authorities, or required to change its practices and controls to comply with national laws and regulations;

> "Italy has become the first Western country to block advanced chatbot ChatGPT. The Italian data-protection authority said there were privacy concerns relating to the model"
>
> *BBC News, 1st April 2023*

19) **Reality gap:** carried along by their own excited hype, the originators and early adopters tend to over-state the benefits while failing to appreciate or under-playing the drawbacks and limitations of a new technology such as ChatGPT. Unrealistic expectations can lead to disappointment and disillusionment among the general audience - a backlash that can involve unfounded fears and an unreasonable reluctance to get involved. Either way, there is a threat that promising innovations may not deliver their full potential;

20) **Social endorsement:** considering comments by other users is generally part of evaluating the risks and opportunities of anything new, hence the utility of customer feedback and product endorsements … but this information is unreliable and can be misleading. Other users have their own requirements and expectations, and some have personal or commercial agendas. Promotional campaigns on social media can reach a wide audience at little cost compared to traditional advertising, while celebrity 'influencers' can

have a disproportionate effect, especially if followers are not thinking critically;

21) **Social engineers**: while benign social engineers such as marketers and trainers are starting to use ChatGPT in support of legitimate business activities, malicious social engineers such as phishers are busy crafting craftier, better written, more convincing, narrowly targeted and simply more varied phishing emails or using other social engineering ruses to trick victims into opening themselves and their organisations to compromise;

> "Darktrace has found that while the number of email attacks across its own customer base remained steady since ChatGPT's release, those that rely on tricking victims into clicking malicious links have declined while linguistic complexity, including text volume, punctuation and sentence length among others, have increased. This indicates that cybercriminals may be redirecting their focus to crafting more sophisticated social engineering scams that exploit user trust."
>
> *The Guardian, 8th March 2023*

22) **Spooks** are intrusive government agencies and agents, both domestic and foreign. The secret services are no doubt heavily invested in AI/ML/NLP tools already, supporting their surveillance/intelligence and other purposes such as disinformation/propaganda/deep-fakes and social engineering on a grand scale. Frankly, I would be amazed if spooks were not discreetly poking around in public-access systems such as ChatGPT as well, perhaps monitoring user inputs for suspicious or outright illegal acts or looking to steal technological advancements and clever tricks for their own benefit;

It is hard to tell them apart!

23) **Thieves** of information and computer data such as intellectual property, personal data, hardware, technologies, proprietary knowledge (trade secrets) *etc*. The motivation here is generally financial but may be malicious or ideological in nature, or perhaps a combination. Information thieves may exploit stolen information directly themselves (*e.g.* obtaining goods and services using ChatGPT customers' credit card details) or sell it to third parties who specialise in particular types of crime such as illicit information brokerage (*e.g.* the underground market for trade secrets and other intellectual property). At a lower level, plagiarists are passing off ChatGPT content as their own, for example in college essay assignments;

It takes skill to engineer suitable prompts, just as it takes skill to use Google or a library for research

24) **Unreliability** is bound to be a concern or threat with anything this new, complicated and changeable. The ChatGPT service has suffered occasional outages for 'technical reasons'. Its outputs are unreliable too in the sense that it can 'hallucinate', perhaps drawing false or misleading conclusions from its training data, confidently spouting nonsense, and maybe even

inventing/faking responses rather than admitting to not having all the answers;

Some humans do this too!

25) **Unanticipated factors**: the novelty of ChatGPT and our lack of experience suggests the likelihood of as-yet-unrecognised situations ahead leading to welcome and unwelcome surprises. Interactions between threats, vulnerabilities and impacts (*e.g.* someone taking advantage of a disruptive incident to perpetrate and conceal a sinister, targeted exploit involving ChatGPT) are difficult to predict and hence control;

26) **Vested interests**: OpenAI clearly has an interest in the success of ChatGPT, but who else is or might be involved? Political, commercial, ideological and other pressures may be at play, behind the scenes, and would be hard to discern even *with* ready access to ChatGPT's training data, system parameters and rules, output statistics, test findings *etc*.;

27) ***Other* threats**: as if the above list of threats was not bad enough already, there may well be others, including those currently learning to exploit ChatGPT in novel ways. The sheer number and variety of threats and uncertainties about them may qualify as a threat in its own right, certainly a challenge for information risk and security professionals.

Vulnerabilities within the ChatGPT 'system'

Vulnerabilities are the inherent weaknesses, flaws or issues *within* ChatGPT – again, meaning not just OpenAI's computer servers but the wider system including all those involved in designing, providing, managing, monitoring and using the service. ChatGPT-related vulnerabilities include:

1) **Complex supply chain/network:** aside from OpenAI, interacting with ChatGPT involves the suppliers of various computing and telecommunications equipment and services, not least the Internet. ChatGPT itself uses hardware, software and services from third parties, and various goods and services are supplied and consumed throughout the entire supply chain/network. We have limited knowledge of the details here (*e.g.* who are those third parties and what are the technologies?), and little if any control over the corresponding security arrangements outside our own domain. We rely on a combination of compliance obligations imposed by contracts, agreements, laws and regulation, plus trust (or blind faith!) that all those involved have adequate security controls in place. For example, we expect everyone to be using appropriate, properly implemented, configured and managed cryptographic protocols, systems and processes for authentication and encryption. Aside from ChatGPT, *most* information services provided by or involving third parties have supply chain vulnerabilities – in other words, ChatGPT is arguably no different, although there may be particular concerns relating to its nature, novelty and popularity that expose it to increased threats (*e.g.* hackers, criminals, secret agents and others are probably actively attempting to compromise

ChatGPT's security right now for reasons such as notoriety, commercial competition or for surveillance purposes);

Complexity is security's Kryptonite

2) **Complex technologies**: AI/ML/NLP systems are inherently complicated with limited visibility of their inner workings, despite efforts to make them 'transparent' and self-explanatory. As with the human brain, examining the model's internals is insufficient to trace and understand the complex sequence of information flows leading to particular responses, except in a general sense (*e.g.* the amygdala is *believed* to be involved in processing emotions, but despite substantial medical research, precisely how it functions remains unclear). Technical complexity compounds the challenge of rapidly detecting and resisting malicious/inappropriate use while permitting benign/appropriate use;

3) **Dynamics of change**: like people, computer systems and other machines with finite resources have a limited capacity to cope with change. Given the rapid, unpredictable and disruptive technological, commercial and societal changes associated with major innovation such as ChatGPT, it is difficult to plan and prepare appropriately, leading to stresses and, perhaps, breakdowns if things fail in service;

Further layers of complexity

4) **Innovation, novelty and inexperience**: although AI/ML has been studied and developed over decades, ChatGPT is rather recent and its public release on the Web followed by rapid uptake is an entirely new experience. Vulnerability is an inherent part of innovation in that, while we don't know for sure how things are going to turn out aside from generic guesses or predictions, there are strong pressures to press ahead and seize the advantage before anyone else;

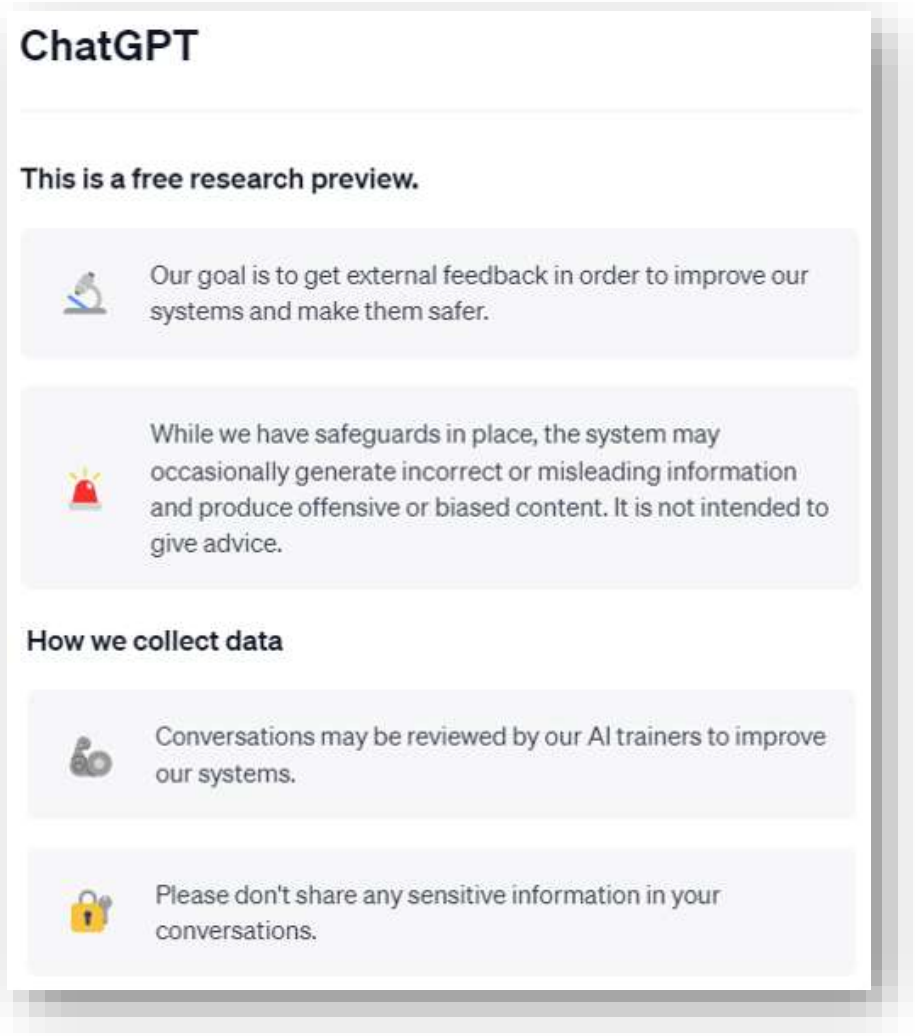
Disruptive innovation is exciting, provided it can be harnessed and tamed

5) **Logical errors**: a system as complex as ChatGPT is virtually *certain* to suffer flaws in its technical architecture such as logical errors and invalid assumptions, plus bugs in the computer code, despite the very best efforts of those involved. There may well be information security implications including 'zero-day' vulnerabilities that are as yet unrecognised by OpenAI and others, except perhaps enterprising hackers hammering away at its defences;

6) **Naïveté:** some new users are glibly or ignorantly using the system without fully appreciating the potential security and privacy implications of what they are doing. They may be casually disclosing confidential personal or proprietary information in their ChatGPT inputs or, beguiled by the plausible and convincing way they are expressed, inappropriately trusting, reproducing and relying upon invalid, incorrect or incomplete outputs;

7) ***Other* vulnerabilities**: as with the threats, it is unlikely that I am aware of or fully understand all of the vulnerabilities in ChatGPT. ChatGPT itself may know of other vulnerabilities, although it is reluctant to disclose them due to system controls intended to block the release of potentially harmful information. In short, this list is probably incomplete and inaccurate to some extent, despite my best efforts.

Sorry but I am only human, after all!



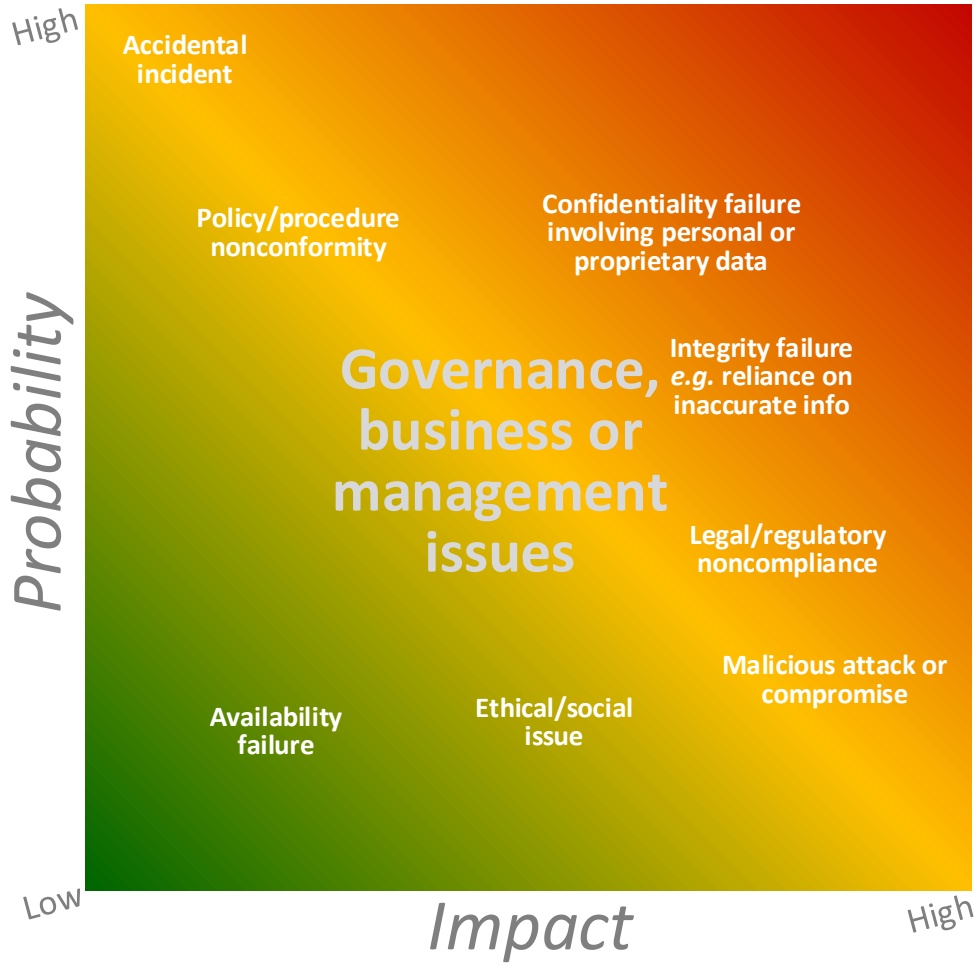
ChatGPT-related incidents and impacts

Information security events, incidents, breaches, accidents, disasters *etc*. have primary and secondary impacts. The primary (direct, immediate) impacts affect information directly, involving loss or compromise of:

- **Confidentiality:** sensitive information is inappropriately disclosed or accessed/stolen. With large volumes of data, it is possible that sensitive personal or proprietary information may be inadvertently disclosed by individual users *to* ChatGPT and OpenAI, or leaked *by* ChatGPT/OpenAI to third parties such supply chain partners or various authorities. Hacks, malware infections, misconfiguration or flaws/bugs in the system could bypass, degrade or negate the confidentiality controls as a whole;

- **Integrity**: information may be inaccurate, incomplete, out of date, misleading *etc.*, or a system, organisation, person or control may be untrustworthy or incapable, failing to perform as intended and required for some reason. ChatGPT may perpetuate or accentuate biases present in its training data or analytical/modelling processes, leading to biased outputs that could harm an organization's reputation or create business or legal liabilities (*e.g.* various forms of discrimination); and/or
- **Availability**: important information, systems or services may not available as expected/required, whether temporarily delayed or permanently lost.

Secondary (consequential) impacts harm those using or relying on the information. The consequences depend on the exact nature, timing and scale of the incident or incidents themselves *and* the particular situations of the affected parties at the point when they occur, hence the harm caused can vary markedly from nothing at all up to critical or even existential damage. The following generic probability-impact graphic[i] and notes represent that broad spectrum with some illustrative examples.



Confidentiality incidents

- Someone deliberately gaining unauthorized access to or stealing user data (personal data or proprietary information) via ChatGPT

- Inadvertent data leaks or disclosures involving sensitive information
- Privacy violations involving systematic collection, analysis and correlation of personal data from various sources including ChatGPT
- Exposure of sensitive commercial data (*e.g.* that used in training) to third parties

Integrity incidents

- Deliberate misuse of ChatGPT to generate harmful/malicious content
- Exploitation of vulnerabilities in ChatGPT's architecture
- Misappropriation of user content or ideas generated through ChatGPT
- Unauthorized modification of ChatGPT's behaviour or output
- Misuse of ChatGPT for spamming or mass messaging
- ChatGPT being used for phishing or social engineering attacks
- Insider threats related to access or control of ChatGPT infrastructure
- Misuse of ChatGPT to generate misinformation, disinformation, propaganda and other manipulative content
- The potential for ChatGPT to be weaponized for cyber warfare or nation-state attacks
- Manipulation or tampering of ChatGPT-generated content to spread false information or cause harm
- Inaccurate or misleading information provided by ChatGPT
- Unintended biases in ChatGPT's responses
- Inappropriate or offensive content generated by ChatGPT
- Inadequate monitoring or oversight of ChatGPT-generated content
- ChatGPT-generated content triggering content filters or being flagged as inappropriate
- Challenges in maintaining the quality and relevance of ChatGPT's training data
- Unclear ownership and accountability for ChatGPT output, leading to disputes and perhaps legal action
- Holes in ChatGPT's training data, plus misinterpretations and conflicts, reducing the quality, veracity and value of its outputs

Availability incidents

- Permanent or temporary interruption of ChatGPT service to users, due to failures, capacity constraints or inadequate performance anywhere in the technology chain between servers and users
- Ready availability of ChatGPT services to virtually anyone with an Internet connected device, regardless of policies or regulations intended to limit or prohibit access

- Devaluation of human creativity and insight due to ready availability of ChatGPT and related systems freely capable of churning out a huge variety of content

Ethical and social incidents

- 'Echo chambers' that reinforce users' existing beliefs while downplaying others
- Use of ChatGPT for covert surveillance/monitoring of individuals or groups, perhaps even the entirety of ChatGPT users, designers, developers, administrators and managers
- Public concern/distrust of AI/ML technology due to incidents involving ChatGPT, plus wariness of disruptive innovation
- Public backlash against AI systems exploiting training data that has been published/released for other purposes
- Reduction of or changes to human employment as a result of AI/ML systems displacing some traditional roles
- Gradual loss of human critical thinking/analytical skills as lazy people increasingly rely on AI/ML for information and guidance
- Generation of offensive, defamatory or inflammatory outputs, inciting violence *etc*.

Noncompliance/nonconformity incidents

- *Claimed* legal or regulatory noncompliance involving ChatGPT *e.g.* privacy incidents, or tax evasion suggested by ChatGPT
- *Claimed* misappropriation of intellectual property rights through ChatGPT-generated content *e.g.* trademark, copyright or patent infringement
- *Claimed* racial or sexual discrimination in, say, a ChatGPT-enabled candidate selection process or bonus scheme could lead to legal action and costs, even if unfounded and unproven
- Workers enthusiastically embracing the new technology without due regard to applicable policies, ethics, laws or other aspects

Defending against such claims would be tricky without sufficient transparency on the analysis and decision-making parts of the process, and clarity over ownership, liabilities and accountability

> "Less than three weeks after Samsung lifted a ban on employees using ChatGPT, the chaebol has reportedly leaked its own secrets into the AI service at least three times – including sensitive in-development semiconductor information"
>
> *The Register, 6th April 2023*

Governance and business management incidents

- Fundamental limitations such as architectural flaws, bugs and inappropriate decisions made in the design and development of ChatGPT, compounded by the system's complexity and lack of transparency
- Premature, potentially inappropriate and risky release of ChatGPT stemming from poor governance and lack of appropriate controls such as laws, regulations and standards
- Inappropriate disclosure and devaluation of proprietary knowledge, trade secrets, personal and other sensitive information
- Security incidents involving the integration/interaction of ChatGPT with other systems or platforms
- Chat-GPT-enabled business processes involving decisions with a financial element (such as credit-checking applicants and authorising loans) may suffer fraud, theft or coercion
- ChatGPT systems and services may be accessed by unauthorized people for inappropriate and potentially nefarious purposes.

Pragmatic information security controls

Conventional wisdom in the profession is to emphasise information security controls that offer the greatest risk reduction with the least amount of complexity and resourcing. Given the novelty of ChatGPT and rapid innovation, however, the information risks are uncertain and the controls largely unproven.

I am therefore offering *pragmatic* examples of the types of controls you might like to consider. The following suggestions are arranged alphabetically rather than by popularity, significance, value *etc.* It is for you, not me, to determine which information security and privacy controls are/are not appropriate for your organisation, given your business situation and information risks.

This is not a comprehensive list of 'recommended' controls – merely some approaches to consider

1) **Assurance** involves identifying information risks through monitoring, checks, reviews, assessments and audits, and reassuring management that they are within acceptable limits. For example, security or privacy assessments exploring the threats, vulnerabilities and impacts associated with ChatGPT may spot issues that ought to be tackled, hopefully *without* suffering incidents. They can also identify improvement opportunities, and confirm that already-implemented controls (such as policies) remain effective and efficient in practice. Supply chain assurance typically involves supplier questionnaires, accredited certification and close business relationships. Another approach is to build and proactively maintain accurate and detailed **S**oftware **B**ills **o**f **M**aterial identifying all the sources of software used in a given system, and to systematically evaluate them for security, privacy or other issues, patch status *etc*.;
2) **Bias and discrimination mitigation**: to address the risk of model bias and inappropriate discriminatory effects, we should review and evaluate systems and processes based on ChatGPT using competent specialists,

statistical techniques, automated bias-detection methods and human review/ethics committees – altogether much easier said than done. In due course, I expect to see the emergence of assurance services and tools supporting this but, for now, we are forging our own path. Good luck finding and securing the services of those 'competent specialists' in this blossoming field: more likely we will have to select, train and support their personal development, hinting at the need for awareness and training materials, experimental/test environments and more, much more. Additionally, organizations should strive for transparency in the selection and composition of training data sets to minimize the risk of biased inputs. Supplementing various controls within the ChatGPT system itself, the manner in which ChatGPT is employed in business processes can reduce the possibility of inappropriately biased information. Systematic record-keeping and statistical approaches, for instance, may indicate discriminatory decisions that deserve management attention and perhaps changes to policies and procedures plus awareness and training for relevant workers;

3) **Business continuity**: approaches such as redundancy, recovery and contingency arrangements may be appropriate for important business activities that depend on ChatGPT, or that may be impacted by incidents such as ransomware attacks, privacy breaches, technical breakdowns or commercial failures. In particular, resilience engineering has the advantage of improving performance and capacity even when things are working well, and keeping the essentials going when they aren't;

4) **Change management**: in the IT context, change management typically means controlling software/configuration changes using version controls, pre-release testing and committees that meet periodically to review and authorise changes … but there can be much more to it, for example treating every significant change as primarily a *business* issue rather than an IT consideration. As with most commercial **S**oftware **a**s **a** **S**ervice cloud applications, customers are largely beholden to the suppliers who may unilaterally make changes that degrade functionality, performance *etc*. for users, with business continuity and security implications. Finding out in advance about planned changes, or being able to delay/defer their implementation, can reduce the risk. Conversely, changes that may be required by user organisations for business, compliance or other reasons may not be understood, accepted or made by ChatGPT;

5) **Conformity and compliance controls**: whereas enforcement actions following detected nonconformity or noncompliance are the usual approach, reinforcement of conformity and compliance is a complementary and potentially highly motivational alternative. Encouraging workers to report ChatGPT-related events, near-misses and incidents, for example, can drive up reporting rates and reduce delays, and may be as simple as circulating guidelines, tracking response rates and 'rewarding' reporters in some way;

6) **Critical thinking**: arguably one of *the* most important human capabilities in modern life, our ability to think critically, analytically, logically sets us apart

from the monkeys and the robots (at least for now!). In this context, it means not simply accepting everything ChatGPT says at face value, taking it as read. It involves considering, questioning, wondering and inquiring, some of the same skills that draw early adopters to explore ChatGPT. However, those who follow along later – egged on by our excitement – don't necessarily share the same capacities, interests or concerns. Critical thinking is a complex cluster of behaviours and capabilities, not something we can magically instil in our colleagues and ourselves through policies, awareness or training. Even given a decade or more of intensive effort during our formative years, the classical educational system struggles to teach students to think critically. So, realistically, what *can* be done to bolster this important control?

7) **Cybersecurity**: given that it is a cloud-based computerised information system, various conventional ICT/cybersecurity controls are appropriate to protect the client systems and networks used to access ChatGPT – prompt security patching and regular backups for example;

8) **Data classification**: classifying information into categories by various criteria may help ensure it is appropriately protected, provided the criteria and the controls are appropriate, correctly defined and consistently applied. An obvious example is 'personal data' subject to obligations under GDPR and other laws and regulations, as well as ethical considerations. Important business data, trade secrets and so on also requires due care to protect its value by ensuring its confidentiality, integrity and/or availability;

Update classification policies, procedures and guidelines to address ChatGPT e.g. do not disclose information classified 'secret' or 'in confidence'?

9) **Data sharing opt-out**: by default, some of the information users provide to the ChatGPT service - notably our input prompts - may be saved and reviewed by OpenAI and, if appropriate, fed back into the ChatGPT model as training data, thereby helping to improve future responses for everyone. This implies that our ChatGPT inputs, our queries or instructions, examples and follow-up questions *could* turn up in someone else's ChatGPT outputs, probably not verbatim but more likely paraphrased through the clever language processing … unless someone somehow persuades the robot to disgorge its training data intact. Users can apply to stop ChatGPT storing and using our inputs in that way using an opt-out form on the website. However, even if we opt out of data sharing, OpenAI retains the right to review our 'conversations' with ChatGPT in order to improve their systems and ensure compliance with their policies. Furthermore, if ChatGPT or OpenAI's information security proves inadequate, leading to or failing to prevent/mitigate an incident such as hack, malware infection or disclosure, our information may be compromised. As with many websites, apps or online services, this could involve the personal information we originally provided when registering, and perhaps credit card numbers *etc*. used to purchase services;

> Read more about this and keep up to date through the ChatGPT FAQ and OpenAI privacy policy. By the time you read this guideline, things may well have changed.

10) **Energy consumption**: a preference for efficient, low-power computing hardware with specialised processing subsystems running in relatively eco-friendly data centres powered by solar or other renewable energy sources can reduce or limit IT's energy consumption. System capacity, performance and power monitoring and tuning/optimisation can be an important part of this control, along with the flexible, scalable features of cloud computing. Since the principles apply throughout the extended system including servers, telecoms, clients, peripherals and all kinds of facilities, corporate strategies, policies and procedures can help to some extent, along with the organisation's *explicit* and *proactive* commitment to eco-principles, preferably without over-playing it ('greenwashing') and creating further risks;

11) **Incident management**: if ChatGPT-related incidents occur, how will workers or third parties identify, report, evaluate, resolve and learn from them? While it *may* not be necessary to cater *specifically* on ChatGPT-related incidents, it may be appropriate to develop or refine the incident reporting and management policies and procedures, depending on the way that ChatGPT is being used – for example, preparing a bland generic disclosure for released as soon as practicable after a privacy incident, giving management some breathing space;

12) **Information risk management**: this guideline exemplifies a proactive, systematic and logical approach to the management of information risks by identifying, evaluating and deciding what to do about them;

13) **Information security management:** information risks should be addressed whenever business processes that revolve around information or data, or IT systems, applications or services, are designed and developed, acquired, implemented, used, managed, monitored and modified;

14) **Network security controls**: traffic traversing corporate networks is typically monitored and controlled using firewalls and other controls. Even if encrypted HTTPS traffic is opaque to the security systems, DNS queries and IP addresses of target systems may be sufficient to identify workers accessing ChatGPT provided they don't find ways to evade the checks. Effective **D**ata **L**eakage **P**revention using proxies may be able to monitor even encrypted traffic, perhaps blocking or triggering alerts if confidential or inappropriate information is detected;

15) **Monitoring**: efficient and ideally continuous security monitoring (both automated and manual) can reduce the delays between incidents occurring, being identified and addressed, and improve the quality of management information (*e.g.* ensuring that potentially significant incidents are definitely identified and suitably prioritised or escalated at the earliest opportunity). Determining what needs to be monitored, how and by whom *etc.* should be an integral part of the organisation's ongoing information risk and security management activities, with ChatGPT and related incidents being simply another type to be considered;

16) **Plagiarism/fakery detection**: AI/ML-powered tools and techniques are being actively designed, developed and trialled to detect the possible use of ChatGPT through linguistic analysis and other techniques. Without access to users' ChatGPT inputs (which are sensitive and unlikely to be made available or searchable by third parties), it may be possible systematically to determine scores indicating the likelihood that ChatGPT was used. Determined plagiarists, in turn, may well edit/customise ChatGPT content in order to reduce the scores and assert their ownership.

17) **Policies and procedures**: managers are busily preparing and mandating corporate policies on responsible use of AI/ML, along with guidelines and awareness/training content (see below). Depending on the business, integrating new policies may involve linking to and updating pre-existing policies concerning information risk management, intellectual property protection, privacy, compliance, IT systems development, management oversight and more. Similarly new/updated procedures may be appropriate, such as how to evaluate the information risks and handle incidents in this area, including how to identify ChatGPT-related incidents promptly and how much/little to disclose in incident reports;

18) **Security awareness**: raising workers' general understanding and appreciation of the information risks associated with ChatGPT *should* reduce the potential for misuse and manipulation, plus accidental incidents (such as disclosing secrets) and carelessness in general (being phished …). We should be proactively informing and motivating people about this topic, ideally as an integral part of an ongoing security awareness programme of activities covering a planned sequence of subjects that are relevant to our audience**s**. So, a topic such as ChatGPT or AI/ML/NLP security might slot naturally into a creative and engaging security awareness approach that covers broad areas of concern such as technology risks, innovation and creativity, carelessness and incidents. It is straightforward to incorporate ChatGPT challenges, situations and incident examples *etc.* into security awareness activities, capitalising on the amount of attention this topic is receiving in the press and social media, provided *someone* is on the ball;

Note the plural: "users" are not the only awareness audience

19) **Security training**: despite often being conflated with security awareness, training is different in that it focuses on delivering more detailed knowledge and particular skills to designated individuals – typically specialists in areas such as information risk and security management and operations, IT, incident management, compliance, systems analysis/design *etc.* It may be appropriate to provide security training on ChatGPT specifically, or more generally on AI/ML/NLP and related technologies, or to incorporate relevant content into other training courses and activities such as new worker induction/orientation classes. Aside from increasing workers' knowledge and competence in this area, hooking-in to topical developments such as this can be a valuable fringe benefit for workers, improving retention and attraction of new talent. For bonus marks, consider encouraging/tolerating skunkworks, allowing the tech-heads to play with and learn more about AI in a safe environment, and reinforcing the learning loop through post-

Feel free to use this guideline as a basis for your security awareness and/or training

exercise and post-incident reviews to identify, justify and make improvements.

20) ***Other* controls**: you may find further inspiration in generic information security control catalogues such as ISO/IEC 27002, perhaps searching for, considering and selecting a balanced mix of controls with a range of desirable attributes or characteristics such as:
    - Preventive, detective and corrective controls;
    - Technological, physical and administrative controls;
    - Simple/basic and complex/advanced controls;
    - Well-proven/trustworthy/reliable controls, where possible;
    - Good value controls *i.e.* their business benefits substantially exceed the lifecycle costs.

Other risk treatments and general-purpose controls

Information risks do not *necessarily* need to be mitigated using information security controls. They may also be:

- **Avoided** *e.g.* by prohibiting or preventing the use of ChatGPT by workers, at least for important business activities or those involving particularly valuable, confidential or vulnerable information;
- **Shared** *e.g.* the possibility of workers' personal data being compromised once they register for ChatGPT is shared between those workers and OpenAI by dint of their privacy policy; and/or
- **Accepted** *e.g.* there is little point in implementing security controls against minor or insignificant risks, particularly if incidents are predicted to cost less than the controls over a period of, say, a few years. Risk acceptance is the default risk treatment that applies to all current or residual risks, including those that are unrecognised or misunderstood.

Staff, managers, interns temps, consultants and contractors should all toe the line

Note that *none* of the risk treatment options can be *guaranteed* to eliminate the possibility of incidents, leaving residual risks if, for instance:

- Mitigating controls are not sufficiently effective in practice, perhaps having not been properly designed, implemented, managed and used, or they simply fail in service for some other reason (*e.g.* they break, the threats increase/change, or they are successfully circumvented);
- Workers are simply unaware of the prohibitions intended to avoid risk, or they consciously ignore/evade them, perhaps because they do not appreciate or feel somehow immune to the risks they are taking;
- Third parties fail to uphold their obligations or expectations relating to shared risks;
- Accepted risks occur and turn out to cause worse than expected impacts;
- Mistakes are made in the identification or evaluation of risks *e.g.* previously unrecognised zero-day vulnerabilities are discovered and actively exploited by hackers before security fixes are available;

- Truly unfortunate coincidences occur – a 'perfect storm' situation that was believed to be impossible or so unlikely that the risk was unwisely accepted/ignored.

Therefore further controls may still be appropriate if those residual risks remain unacceptable, typically meaning broadly-applicable/general-purpose controls such as governance, awareness and assurance.

Conclusion

Rapid adoption of AI/ML/NLP technologies such as ChatGPT brings information risks, some of which are clearly significant and require mitigation.

By prioritizing pragmatic security controls such as data classification, input and output sanitization, monitoring and auditing, security awareness training, model transparency, secure SDLC, and incident response planning, organizations can effectively manage these risks, enabling the secure, responsible and productive use of ChatGPT.

Implementing these security controls is not a one-time effort but rather an ongoing process that requires continuous evaluation and improvement. As the AI landscape evolves, so too should an organization's approach to risk management and security. By staying vigilant and adapting to new challenges, CISOs and IT auditors can help their organizations reap the benefits of AI technologies like ChatGPT while minimizing potential risks.

As the use of AI systems like ChatGPT becomes more prevalent in the enterprise environment, CISOs and IT auditors must remain vigilant in addressing the unique information risks associated with these technologies. By prioritizing the pragmatic security controls outlined in this article, organizations can significantly reduce their exposure to risk while continuing to harness the benefits of AI-powered NLP. It is essential for organizations to remain proactive and adaptive in their approach to securing ChatGPT and other AI systems, constantly reassessing their security posture and evolving their defences to keep pace with the ever-changing threat landscape.

Footnote

[i] About the probability-impact graphic:

- The graphic is figurative/representative, presented merely to stimulate thought and discussion about the risks. It is neither accurate, definitive nor complete.
- Higher probability risks seem more likely to involve coincident clusters of a given incident type than lower probability risks, but any combination is possible (*e.g.* a string of privacy issues might lead to ChatGPT being taken offline by the authorities/courts or OpenAI).
- A given incident may fall into several of the categories shown *e.g.* inappropriate disclosure of personal data is a confidentiality failure that may involve noncompliance with GDPR or other privacy laws, and may have been caused deliberately by hackers or accidentally by workers failing to conform with corporate privacy policies and procedures. Although the categories are shown separately, they overlap.
- The graphic is a snapshot - a simple static visual representation of a dynamic and complex situation. In reality, the risks are gradually changing, sometimes jumping unpredictably.
- Items on this graphic are positioned relative to each other based on my general understanding of the issues, a purely subjective assessment. Your opinions probably vary, especially if you know of actual incidents or have reliable research data.
- Individual incidents may fall well outside the areas indicated by the text *e.g.* issues at any point on the graphic may be of interest or concern to management and perhaps other stakeholders in the business, so the central positioning of the "Governance, business or management issues" item is arbitrary. The key point is that figure represents a generic risk assessment from the perspective of an organisation whose workers are using ChatGPT.