How to stay safe in the age of artificial intelligence

Like most folks, ChatGPT-4 blew me away. Its spring 2023 debut made AI a public phenomenon.

Back in 1950, AI that could fool people was still just an idea in Alan Turing's (brilliant) head. Heck, any AI at all was still just an idea that year.

The first AI program was created in 1951. It played checkers.

Now AIs can churn out cogent essays, hack computers and maneuver military drones. This may well be the age of artificial intelligence.

Industry leaders warn that "mitigating the risk of extinction from AI should be a global priority" including Sam Altman and AI field pioneer Geoffrey Hinton. Taking a bird's eye view, Elon Musk has predicted that AI will become the "most disruptive force in history".

Where will AI be 100 years from now? Or even 10 for that matter?

What are the AI safety risks that we face as individuals, families, and businesses? What can we do to protect ourselves?

We'll answer those questions. Let's get a better sense of scope first.

How prevalent is AI safety risk?

Very. AI safety risk goes hand in hand with the use of AI.

These days AI powers up everyday tools such as:

- Smartphones
- Social platforms
- Online shopping

[Image: https://unsplash.com/photos/people-using-phone-while-standing-qZenO_gQ7QA]

And more. But that's just what we as consumers see. What about stuff operating in the background?

- Voice capture
- Image generation
- Facial recognition

We've been discussing consumers. What are some ways that AI has been incorporated into business?

- Virtual assistants doing scheduling
- Personalizing product suggestions and promotions
- Automating customer support replies to queries via chatbots and the phone

Our discussion so far has been about the private sector.

Governments have adopted AI for both civil and military uses. Though the line has blurred. Consider how AI has been integrated into government surveillance of citizens.

Understand the risks of AI for individuals, families, and businesses

AI safety risk does not affect people equally.

In the world of business, **small businesses face greater risk from the rise of AI**.

In our personal lives, children are the most vulnerable to AI safety risks. They simply lack the judgment to evaluate and manage AI safety risks. Risks we are about to discuss for adults in their personal lives apply to a heightened degree for children.

In terms of categories, AI safety risk is highest when we're talking about **misuse and malicious use**.

- As an example of misuse, consider a tabloid that used AI to create a fake interview with a celebrity. The tabloid only mentioned the AI-generated nature of the piece at the very end of it.
- For malicious uses of AI, a thief could use an AI trained on your social media, pretend to be you, and defraud your family of money. More on this subtopic later.

Note that much of day-to-day AI safety risk–what we as ordinary people deal with the most–is actually an unintended use, the result of bugs. These earthy problems are the focus of this article.

For now, let's focus on business.

Small businesses face significant AI safety risks

Modern 21st century businesses depend on many different vendor AIs now. These usually eliminate annoying, administrative-type tasks.

[Image: https://unsplash.com/photos/woman-in-black-shirt-using-laptop-computer-L85a1k-XqH8]

The percentage of small businesses selling an AI product or using one to make a product has been going up steadily in recent years. Source: Census Bureau.

Small businesses are more vulnerable to AI safety risks. Big businesses:

- Can pay for customization
- Are able to build AI themselves
- Bargain with the benefit of internal experience and economic leverage
- Use in-house AI expertise to identify AI safety risk and solve it

The development of AI impacts data privacy in two ways

All of your devices collect significant sums of data. In the aggregate, this feeds AI development, which depends on quite a lot of data. In the future even more data will be collected.

Thus, AI raises the stakes for data privacy.

At the same time, AI is becoming more powerful at weakening digital safeguards for security and data privacy.

One example is AI-enhanced password cracking. This is probabilistic thinking fed on large sums of data.

With AI, one success can help produce another. For example, one password can be used to generate more passwords of the same person.

The mistakes of machines

Earlier this century, belt-tightening austerity regimes were imposed on European countries. As you know this created quite a bit of political and socio-economic turbulence.

It happened in part because of bad scholarship. The economists responsible used an Excel equation missing a few rows of data. The politicians didn't question the quantitative results.

This kind of error is easy to attribute to people. But all AIs are made by people. Just like the simpler sort of computation we just discussed.

These scholars were not in the same situation as, say, people who see content on TikTok's algorithm. Viewers may choose to interact with the AI, but they did not create it. TikTok has been criticized for negatively influencing the health-related decision-making of impressionable, at-risk people.

Airplane pilots can seize command from an automated system that makes your plan take a steep nose dive. **But is it so straightforward to free our thinking from false and misleading information?**

What sorts of errors do AIs make?

- Inaccuracies and mistakes, like the examples we just discussed.
- Hallucinations, where the AI just makes stuff up! (Just like people.)
- Biases and prejudices, which are a special sort of inaccuracy.

What's worse is that those listed categories don't include bad actor behavior. You need your antenna up to handle those.

The rise of AI-powered threats and scams

We'll talk about scams, and then threats. Consider:

- Social engineering attacks
- Identity theft and fraud
- Phishing, smishing, and vishing
- Deepfakes
- Fake kidnapping and ransom calls

Focusing on these categories will help us discuss principles that apply across the board.

AI helps commit **social engineering attacks**.These victims are lured by psychological tricks to create a security vulnerability or give sensitive information.

For example, it's become more common in recent years for **identity theft and fraud** to rely on AI-generated voice or text. This is another type of social engineering attack. This crime occurs when someone steals personal information to represent themselves as someone they're not, usually for money.

There are three methods of identity theft and fraud.

- Email, or **phishing**
- Text message, or **smishing**
- Phone calls, or **vishing**

Generative AI helps cybercriminals create false appearances and representations. It could be words, images, or voice.

Like a **deepfake** that convincingly mimics the way a trusted person or institution presents themselves.

The most popular schemes at this time are **fake kidnapping and ransom calls**. Here the target receives a call from a supposed loved one who has apparently been taken hostage. Payment is required for release, only for it to be revealed that the whole thing was a lie.

In addition to scams, artificial intelligence can power many types of threats to safety. Consider:

- **Computer hacking** tools that are more effective
- Content on social media to spread **false narratives**
- New types of **malware** to infect computers
- Systems to **reverse engineer** successful products or security safeguards
- **Misinformation and disinformation** for political and other reasons
- Deepfakes of influential people and leaders doing things they actually didn't do
- **Cyberbullying and hate speech** against targeted individuals and groups including minorities

Consider another example more in depth: **Data poisoning**. Bad actors can also maliciously introduce inaccurate or biased data into the training dataset for an AI, thus leading the algorithm to commit errors or discriminate. On the other hand the same technique can be used by artists to protect their work from being scraped without permission.

How to spot AI-powered threats and scams faster

- Avoid being fooled, knowing you can be
- AI can be used anywhere to affect any activity
- You have something valuable for an AI to steal

Avoid being fooled, knowing you can be
Criminals evolve their tactics. AI development is fast paced. Unless you're an expert you won't be able to tell sometimes if something is an AI-powered scam or threat. That's why it's imperative to act on any red flags that appear and do your due diligence.

Consider one theme in the categories we quickly reviewed. Criminals using AI often need to fool you somehow. They need you to believe that you're dealing with a trusted person or institution like a bank. We'll discuss how to handle this problem later.

AI can be used anywhere to affect any activity
Remember that with AI, the medium or format doesn't make you safe. It doesn't matter if you're using email, phone calls, videos, websites or whatever – AI can be used to help cause mayhem and the objectives of criminals.

Always assume you have something valuable for an AI to steal
The motives behind AI threats are many. Money is a big one of course, either stealing it directly or making money in some other way. Then there's your data and information, which can be

valuable for use or resale to another party. Big motivators for misuse and malicious use also include politics, religion, or some type of social movement.

Never doubt that you have something valuable to steal via a sophisticated AI. Just because you wouldn't steal it from someone else, doesn't mean someone wouldn't steal it from you.

The target of AI-powered theft could be your money, your data, your permission, your reputation, or your vote.

How to stay safe and use AI responsibly

At a high level? Remember two best practices.

- AI development and AI safety knowledge evolve. Make sure you stay up to date.
- You benefit from AI in terms of safety too. Use it for your digital protection.

Sam Altman has noted with regard to advancing AI technology, "We are on an exponential curve and a relatively steep one." He observed that human intuition is not built for exponential learning curves.

This means we have a lot to learn.

But all the knowledge we need to manage AI safety risks isn't static. AI development is fast paced and bad actors figure out new ways to get what they want.

We all have to make an effort to stay up to date with the latest news and advice from experts in AI safety, cybersecurity, and data privacy. We should all regularly ask three questions.

- What is realistically possible with the latest technology?
- What are the most pressing risks?
- What are the best practices?

At the same time, be aware of ways to use AI to protect yourself. Consider a few possibilities for businesses.

- Have your company spam filters use machine learning to improve.
- Check for suspicious transactions through a fintech AI software.
- Train the biases out of one AI using another.

Staying safe from AI as a target, user or buyer

A large portion of AI-powered scams and threats depend on people accepting representations or appearances. For the criminals involved, fooling others is essential. In these scenarios the targeted individuals are routes of access to desirable (digital) locations.

The deception here is about authenticity: Is someone who they claim to be? Consider two situations.

- There's an email from your bank telling you to login (via a link) to the bank's website and review some materials.
- Your family member calling you over the phone and asking for money.

If it's an individual or an institution in which you place (digital) trust, understand that your permission could be valuable to a third party. **At some point AI may be used to fool you into thinking you're dealing with a trusted person or organization.**

Tighten up at moments where the authenticity of who you're dealing with really matters in your digital life. Examples include:

- Opening links
- Logging in
- Saving information
- Being recorded

Ask yourself: How do I know this person is who they claim to be? There's always a practical way.

- Loved ones are in your phone's contact list so you can just call that number before doing anything else.
- Trusted brands have specific methods of communication, handles and website URLs as well as hard-to-fake styles and branding elements.
- Banks are far more stringent than a typical company in how and what they discuss with customers.

Now let's talk about interacting with AI as a buyer or user.

Learning what's available about the AI
If you're dealing with an AI, read the information offered to you and on the website of the organization. This is especially important when it's the government because, well, there's no alternative! It's also important when doing business with firms.

As much as possible, understand what you can about the AIs that really impact your life. This is helpful for battling and fixing a variety of errors that they commit.

In the case of hallucinations using different prompts can help you. Hallucinations happen because the AI doesn't understand whether it should invent fake items or focus on a particular context the prompting person has in mind.

Flagging and reporting AI issues
If you see abuse, bias, or serious inaccuracy, report it to the organization responsible for the AI. As soon as you can.

AI biases can be pernicious and can affect disadvantaged individuals and groups. This could involve a range of situations, such as policing that uses predictive algorithms or child welfare choices.

Whether it's the government, a social media platform, or another firm, simply let them know about the issue with your evidence. Have a discussion in whatever way is available.

Nothing can be done unless people know about the problem.

Demand transparency
AI should be explainable. To put it roughly, that means readily understandable. Only then can users have trust in the system because they can know how to use it and they can hold it accountable.

Users need to know how, at least at a high level, an AI reaches its conclusions. Users also need to demand transparency so they can better identify inaccuracies and biases.

- If you use an AI and did it through sales or a demo, you can talk directly with your point of contact.
- For consumers, you can try the company's normal channels of feedback to see if they work. Or band together with other consumers on social media.

Responsible development of AI

Concepts like explainability, interpretability, and transparency are hard to define and operationalize (even AI experts will admit it).

Let's stay practical.

People in general want to use or make AIs that make them more competitive (if we're talking business) or live a better life otherwise.

If we're talking about business, then you want the most effective AIs while being fair and honest with yourself, your team, and your customers.

To that end, consider these approaches.

- If you use an AI when making or selling something, make sure you learn as much as you can about how the AI works and what data it uses.
- Make it easy for customers to report issues with an AI.

- When you communicate to your clients, try your best to help them understand the role of AI in your product and how it is developed.
- Create in-house guidelines on how AI should be used and who to email on the team if there are questions.
- Stay up to date on legal developments. AI regulations and treaties may come down the pike faster than you think.

AI safety issues that require collective action

We've discussed two big things.

- The most common AI safety risks
- Best practices for managing them

Let's go loftier. We'll talk about issues that you have less influence over. But they can be solved with collective action. This could mean activities of a non-profit or a government agency.

To be fair, there are signs of a movement against AI. Certain groups seek to halt AI development. This seems unrealistic. AI is here to stay because of how helpful it can be.

[Image: Of a protest. https://pixabay.com/photos/riot-protest-street-crowd-laws-6129239/]

Big advances in AI are normally paid for and to the direct benefit of profit-motive organizations. Indeed, AI development depends on expensive computing infrastructure. Frequently only the richest firms in the world can afford it.

But the dominance of the private sector in the world of AI is lessening.

We have seen progress in terms of government effort to manage AI risks, especially towards the development of laws and regulation.

- The release of ChatGPT -4 in the spring of 2023 sent lawmakers across the world, like in the European Union, scrambling.
- In October 2023 President Joe Biden signed the first USA executive order focused on AI. Legally, it is for federal agencies. But it lays out important underlying principles and is a symbolic step forward.
- The United Kingdom hosted an important AI safety summit in late 2023.

One of the biggest obstacles to collective action on AI safety right now is a lack of effective and shared understanding.

One big push right now in the field of AI is to create detailed, useful standards to evaluate AI safety risks. As we discussed, concepts such as explainability are difficult to operationalize.

Here are more of the big AI safety risks that societies and countries across the globe face.

- AI has a **major environmental impact**. There is a cost to the planet of using the cloud.
- AI can negatively impact the quality of journalism and **factual richness of public dialogue**.
- Public opinion and government decision-making could be influenced by AI-generated misinformation and deepfakes.
- Uncertain **economic impact**. This includes the rise of algorithmic trading as well as job losses.
- AI development is so expensive this creates **significant barriers to entry**. There is potential for rent-seeking firms facing little competition.
- Governments can use AI for surveillance of citizens, with significant legal and ethical problems involved.
- AI could exacerbate the risks of pandemics and bio threats.
- The possibility of an AI arms race between nations (involving autonomous weapons) between nations has been conjectured.

For better and for worse, we live in an age of artificial intelligence

The power of human intelligence combined with the power of machine computation. This is how AI can promise and endanger us so much.

AI won't go away. So how should we adapt to the long term?

Let's acknowledge the variety on tap: Certain AIs are ineffective, others are helpful, and some with negative impacts.

We have to do our best to protect ourselves while we benefit from the rise of AI. **The key is to become more sophisticated in how we handle algorithms.**