Worried about AI agents? A calm, plain-English guide to staying safe

If you have read a news story about "rogue AI agents" and come away uneasy, you are not alone. The headlines move fast, the words are new, and it can feel like something big is happening just out of sight.

Here is the good news: you do not need to understand the technology to stay safe. The habits that protect you are simple, most of them are free, and many are things careful people already do. This guide explains what AI agents actually are, which everyday risks are worth knowing about, and a handful of steps that bring real peace of mind.

What AI agents are, and what the headlines mean

An AI agent sounds mysterious, but the idea is fairly ordinary. It is a computer program that can carry out a task for a person, step by step, using the apps and accounts that person has connected to it. The scary stories are usually about agents doing something their owner did not intend, which is a real issue for the people and companies who run them. For most of us at home, the practical concern is much more familiar: scammers using AI tools to make old tricks look more convincing.

- **Chatbot:** a program you type or talk to that writes back in plain language. It answers questions and drafts text, but on its own it does not take actions in the world.

- **AI agent:** a chatbot that has been given permission to act, such as sending an email, filling in a form, or booking an appointment. It can only use the tools and accounts a person has connected to it.
- **Agents do not have wants or plans of their own.** They follow instructions. When one goes wrong, it is usually because it misunderstood a request, was given more access than it needed, or was fed bad instructions.
- **Prompt injection:** a trick where someone hides instructions inside an email or web page, hoping an AI that reads it will obey them. This is one reason "rogue agent" stories happen, and it is mainly a problem for the people who build and run agents.
- **What is real:** agents can make mistakes, and the people who run them need good limits, such as asking a person before anything is sent or paid.
- **What is hype:** an agent cannot reach into your phone, email, or bank account on its own. It has no access to your accounts unless someone gives it that access.

The everyday risks worth knowing about

The FBI has publicly warned (www.ic3.gov/PSA/2024/PSA241203) that criminals use AI tools to make fraud more believable and to reach more people with less effort. The scams themselves are not new. They are the same old cons with better spelling, a more familiar voice, or a more convincing picture. Once you know what they look like, they lose much of their power.

- **Polished scam messages. Phishing** is a fake email, text, or message that pretends to be from someone you trust, like your bank or a delivery company, to get you to click a link, sign in, or pay. AI helps scammers write these with clean grammar, so spelling mistakes are no longer a reliable warning sign.
- **Voice-clone calls.** A **voice clone** is a copy of a real person's voice made by AI from short recordings, such as videos posted online. Scammers may use one to sound like a grandchild or relative in trouble who needs money right away. The FCC (www.fcc.gov/consumers/scam-alert/grandparent-scams-get-more-sophisticated) notes that caller ID can be faked too.
- **Deepfakes.** A **deepfake** is a fake photo, video, or recording made with AI that shows a real person saying or doing something they never did. They often show up as a celebrity "endorsing" an investment, or as a video call meant to prove that a stranger online is real.
- **Fake support.** A pop-up, call, or website claims something is wrong with your computer or account and offers to fix it. Some fake sites even have a friendly chat window. Real companies do not ask you to install remote-control software or pay with gift cards.
- **Oversharing with chatbots.** What you type into a chatbot may be stored by the company that runs it. It is a good tool for general questions, but not the place for passwords, account numbers, your Social Security number, or private medical details.

- **The common thread.** Almost every one of these scams leans on the same three things: pressure to act now, a request to keep it secret, and an unusual way to pay, such as gift cards, wire transfers, cryptocurrency, or a payment app.

Simple steps that give you peace of mind

You do not need special software or technical skill to be well protected. A few calm habits cover most of the risks above, and the steps below come from guidance published by the FBI (www.ic3.gov/PSA/2024/PSA241203), the FCC (www.fcc.gov/consumers/scam-alert/grandparent-scams-get-more-sophisticated), and the federal cybersecurity agency CISA (www.cisa.gov/secure-our-world). Pick one or two to start with this week, and add the rest when you have time.

- **Choose a family code word.** Agree on a word or phrase that only your family knows, and share it in person. If someone calls in an "emergency" and cannot give it, that is your answer.
- **Slow down and verify.** If a call, text, or email asks for money or personal details, hang up or close it. Then contact the person or company yourself, using a number you already have saved or the one printed on your card or statement.
- **Keep payments boring.** Do not send gift cards, wire transfers, cryptocurrency, or payment-app money to someone you have only met online or over the phone.
- **Turn on two-factor authentication. Two-factor authentication**, often called 2FA, means signing in takes your password plus a second check, usually a code sent to your phone or shown in an app. Even if someone learns your password, they still cannot get in. Start with your email and your bank.

- **Use a password manager.** A **password manager** is an app that creates and remembers a different strong password for each of your accounts, so you only need to remember one. Many phones and web browsers have one built in.
- **Install updates.** Updates for your phone, computer, and apps often fix security holes. Turning on automatic updates is one of the easiest wins there is.
- **Share a little less in public.** Setting social media accounts to private and limiting public videos of your voice gives scammers less material to work with.
- **If you already clicked something, take a breath.** It happens to careful people every day, and it can usually be fixed. Stop and do not enter anything else. Change that account's password by typing the real website address yourself, and turn on 2FA. If money or card details are involved, call your bank using the number on the back of your card. You can report scams at ReportFraud.ftc.gov, report online fraud to the FBI at ic3.gov (www.ic3.gov), and get a recovery plan for stolen personal information at IdentityTheft.gov.

You are more prepared than you think

The technology in the headlines is new, but the way to stay safe is old and familiar: slow down, check with someone you trust, and never let pressure make the decision for you. Share this page with someone who might find it reassuring, and talk through a family code word at your next dinner or phone call.

At ReadyUpAgents, we build tools that help keep AI agents in bounds, and we share free guides like this one in our articles (readyupagents.com/articles).

Sources: FBI Internet Crime Complaint Center public service announcement on generative AI fraud (www.ic3.gov/PSA/2024/PSA241203), FCC: "Grandparent" scams get more sophisticated (www.fcc.gov/consumers/scam-alert/grandparent-scams-get-more-sophisticated), CISA cybersecurity tips (www.cisa.gov/secure-our-world), ReportFraud.ftc.gov, IdentityTheft.gov.

Free to print, copy and share. ReadyUpAgents is independent and not affiliated with or endorsed by the FBI, FCC, FTC, CISA or any company mentioned here. Product and company names belong to their owners.