AI Keeping Vulnerable People Safe: Checklist for an AI-enabled world

This checklist enables organisations to think about the possible additional threats to vulnerable members of their community and identifies practical steps that can be taken to help keep them safe. Work through the checklist to identify the actions you will take, amending any as necessary, deleting those you won't and adding any you wish to.

Risk Management

We have identified groups within our community who might be vulnerable and/or not fully competent to make decisions. For example, younger and older people, those with learning disabilities, low educational attainment or in digital poverty.

We have risk assessed the additional and/or increased threat to them and taken appropriate steps to ensure they are safeguarded. Threats might include people:

- Forming emotional dependence on chatbots.
- Receiving harmful or manipulative AI advice from an AI bot.
- Using AI for emotional or crisis support, self-diagnosis of medical issues or legal/financial decisions.
- Being scammed or mislead by AI-generated messages, voices, or images created by others.

Simple Steps to Stay Safe

Below are a series of steps that vulnerable people could take to help keep themselves safe. We will assess these against our community's needs, reword them so that we communicate with them in a way they would want us to and that will also work well for them. We will provide the checklist in large print, easy-read, or other formats for those who may need this.

- I will never trust AI with personal, financial, or any other information I wouldn't want others to know about.
- I won't rely on AI for emotional support, advice in a crisis, or health, financial or legal advice.
- I take care if an AI system seems to be *caring* or *personal* – I know it's not a person.
- If an AI message makes me feel pressured, guilty, or scared, I will stop and talk to a person.
- If something doesn't look or feel right, I won't trust it.
- If it looks or sounds too good to be true, it almost certainly is and I won't trust it.
- If a message feels urgent or asks for money or personal details, I'll treat it as suspicious.
- I won't click on links or open attachments unless I'm sure they're safe.
- I'll double-check sender names and email addresses carefully before replying.
- If I'm in any doubt, I'll stop and check with someone I trust before doing anything.
- I'll report suspicious messages to my email provider, phone network, or organisation.
- I won't feel embarrassed — scammers target everyone, and reporting helps protect others too.

Additional Support

We will identify other steps that will help them to stay safe from risks such as deepfakes and scams. We will provide additional support in these areas, if they might need this, or include these in our safety list above, if not. Here are some examples:

- Check your social media privacy settings — limit what strangers can see.
- Keep personal info private — don't post your address, phone, or bank details online.
- Use strong passwords — don't reuse them, or use a password manager.
- Turn on multi factor authentication (MFA) – extra login protection for your accounts.
- Update your devices — install updates when prompted to fix security gaps.
- Be careful with public Wi-Fi — avoid logging into sensitive accounts, such as your bank.

Our Use of AI Bots

We do not:

- Delegate crisis or emotional support to AI bots and our people know that they should redirect request for support to human services.
- Allow the use of AI bots for crisis or emotional support, or to provide professional advice, such as medical or legal guidance.

Fundraising Bots

In addition, if we use AI bots to help with fundraising, we will make sure they are safe and fair for everyone, with specific safeguards for those who may be vulnerable. We will ensure that the bot:

- Is safe and secure: We will choose a system that is secure, fully data protection compliant, does not store donor information overseas or share/sell information to 3rd parties.
- Is easy to use and understand: We will choose a system that is easy for everyone to use, particularly, those who may struggle with digital/AI systems, including providing information in plain English and/or other accessible formats.
- Spot signs of vulnerability: The AI bot's personality will include the need to safeguard vulnerable individuals, and it will be trained to look for clues that someone may be struggling to understand, is confused, or is just agreeing without really joining in. For example, if someone keeps asking the same questions, repeats what the bot says, or seems unable to explain their choices, the bot will recognise this as a possible sign of vulnerability/lack of competence.
- Respond with care: If the bot notices these signs, it will slow down, use simpler language, and offer to explain things again. It will never pressure anyone to donate and will always give people the chance to ask for help or to speak to a real person.
- Respect preferences and needs: The bot will ask how it can help and what the person needs, rather than making assumptions. It will summarise any decisions or agreements in plain English.

- No undue pressure: We will not write into the bot personality or train it in a way that might encourage it to apply undue pressure, such as the following.
    - Asking for an amount that individuals might not reasonably wish or be able to give.
    - Repeatedly asking someone to donate after they have said no or not respecting a person's request to stop the conversation or to be left alone.
    - Using emotional pressure, such as suggesting that not donating will have negative consequences for others.
    - Making it hard to say no - for example by hiding the "no thanks" option or making it much easier to donate than to decline.
    - Rushing the person or saying they must decide immediately
    - If the bot thinks someone may not fully understand or may not be able to make their own decisions, it will not ask for or accept a donation.
- Easy reporting and robust oversight: There will be a clear and simple way for anyone to report concerns about the bot's behaviour. All bot conversations will be checked regularly by staff to make sure the bot is acting safely and respectfully.
- Staff training and review: Staff and volunteers will be trained to understand how the bot works, how to spot signs of vulnerability, and what to do if something goes wrong. We will keep records of how the bot is used and review these often to spot any issues early.
- Act quickly if there's a problem: If we believe that someone has inadvertently been harmed or exploited, we will act immediately to put things right and report it to our board and any others who may need to be informed. We will also review and improve our systems to prevent it from happening again.

Implementation

Make our people and wider community aware of these additional risks, simple steps they can take to keep themselves safe and how to raise concerns. We will ensure that:

- The safeguarding of vulnerable people is reflected in our AI policies, procedures and training.
- There are clear escalation routes for AI content that may be harmful, manipulative, or distressing.

We use Safety by Design principles in building or choosing AI systems.

Including designing in accessibility, for example, for neuro diverse people.

We work with our IT department or IT support provider to:

- Block access to unregulated AI chatbots on all work laptops and phones.
- Use browser filters or firewall rules to prevent access to unsafe AI websites.
- This checklist will be reviewed at regular intervals to keep it up to date with new AI risks.

Help us to help everyone by sharing this with your network. Charity Excellence Learning: certificated, online AI courses that anyone can undertake.