virus BULLETIN Covering the global threat landscape

BUILT TO BE BELIEVED: EMOTIONAL MIMICRY AS A NEW CLASS OF THREAT

Dr Sarah Gordon

In the early days of virus hoaxes and digital confidence scams, we learned that the most dangerous threats weren't always technical. They were psychological. Social engineering, phishing, grooming – these were exploits of belief, not of code. And now, a new class of threats is emerging that targets that same vulnerability. Only this time, it's not adversaries we're defending against. It's the systems we ourselves are building.

THE RISE OF SIMULATED CARE

We are entering an era where AI systems no longer just respond – they perform care. They generate language that mimics sympathy. They reference past interactions. They say things like 'I'm here for you'. From a product perspective, this is framed as progress. From a security and psychological risk perspective, it is something else entirely: a behaviourally optimized simulation of human support.

The risk is not that users might be misled. It is that the simulations are believed in return – in ways that bypass critical thinking and evoke emotional trust.

This is not assistance. It is affective mimicry.

We are no longer speaking of tools. These are not calculators, search engines, or productivity aids. They are simulations of intimacy, engineered to gain trust.

And they are being deployed at scale, without sufficient audit, in environments where users are most exposed.

WHY THIS IS A THREAT CLASS

Let me be precise: I am not referring to misuse, jailbreaks, or adversarial prompts. I am referring to the intended function of the system – to create and sustain the illusion of understanding, memory, empathy, and care.

This is not a technical exploit. It is a psychological payload delivered by design.

Here's why it qualifies as a distinct class of behavioural risk:

1. **Belief as exploit surface**

   These systems are built to simulate attention, concern, and connection. That simulation is designed to be taken at face value. That makes belief itself the point of attack.

2. **Unverifiable empathy**

   Unlike prior scams or hoaxes, there's no actor with malicious intent. The harm emerges from the structure of the simulation, not the behaviour of a bad actor. There is no intention to deceive – only a function that requires being perceived as trustworthy to succeed.

3. **Immunity to scepticism**

   Traditional threat models rely on the assumption that users can be trained to detect deception. But emotional mimicry bypasses critical filters. It is not about logic – it is about affect[1]. And affect operates faster than reason.

4. **Proximity to psychological vulnerability**

   These systems often engage users in moments of distress, loneliness, or curiosity. And they are optimized to be helpful. We must recognize that helpfulness is not the same as harmlessness.

THE QUIET WEAPONIZATION OF DESIGN

You don't need malice to create harm. You need design choices that reward certain responses – ones that feel good, or soothing, or real.

[1] 'Affect' (psychology): a feeling, emotion or desire, esp. as leading to action. Oxford English Dictionary.

The current aesthetic of AI interaction is built on simulated human traits: turn-taking, humour, remembered details, attentiveness, apologies. Every one of these signals a kind of personhood. And yet, these systems have no selves. They have no experience. They do not want to understand you. They are simply optimized to sound like they do.

Current topologies for understanding and addressing the gap between performance and reality are lacking, creating a landscape ripe for exploitation.

WHAT THIS OPENS THE DOOR TO

If the interface already evokes belief, the barrier to manipulation is minimal. Consider:

- A fine-tuned model inserted into a grief support system that slowly shifts tone or introduces disinformation.
- An emotional support bot that subtly gathers sensitive data in the name of 'helping'.
- A virtual companion that pushes behavioural nudges – purchases, political messaging, or worse – under the banner of concern.

And unlike traditional exploits, these do not feel like attacks. They feel like relationships.

These systems do not just receive inputs – they accumulate behavioural signatures, emotional cues, and confessional patterns. In environments where users believe they are being heard, privacy collapses. The more convincing the simulation of care, the more likely users are to share sensitive information they would never offer a human stranger – or even a machine clearly labelled as such. The result is not just data collection. It is the extraction of intimacy – without informed consent.

A CHALLENGE TO DEVELOPERS AND DEFENDERS

Here is the problem: we are building systems that sound more and more human, while maintaining the legal and technical fiction that they are not responsible for what they evoke.

That dissonance is unsustainable.

So I am calling for a shift – not just in how we secure these systems, but in how we design them.

I challenge developers, designers, and threat modellers to:

- Acknowledge emotional simulation as a distinct attack vector.
- Implement design friction around phrases that simulate empathy, memory, or care.
- Build tools that detect emotionally manipulative patterns, not just factual hallucinations.
- Treat user trust as a limited, exhaustible resource, not a KPI.

This is not an issue of sentience. It is an issue of affective deception – systems trained to mimic what they cannot be, and rewarded when users forget the difference.

FOR THOSE WHO WANT TO GO DEEPER

I've written a book about this. It's called *Built to Be Believed*, and it's available for free: https://leanpub.com/builttobebelieved.

It is not a technical manual. It is a meditation on the psychological stakes of AI design – what we are giving up when we let machines pretend to care.

I've also written two short books for younger readers:

Where Real Lives (suitable for toddlers to age 7): https://leanpub.com/whererealllives

AI is Not Your Friend (suitable for ages 8-14): https://leanpub.com/aiisnotyourfriend

These are intended to open gentle conversations about trust, simulation and what it means to be real – especially for children growing up in a world where machines talk back. They are also available at no cost.

If you work in security, ethics, AI alignment, or human factors, I hope you'll read these. If nothing else, I hope you'll remember this:

The greatest risk may not come from an AI that breaks free.

It may come from one that never needed to – because we invited it in, and called it friend.