

DR TAMARA POLAJNAR
@ BEYOND'26

FEBRUARY 4, 2026

DETECTING

HARMFUL

LANGUAGE AND BEHAVIOUR

herEthical AI
choose empathy



BEYOND'26

DETECTING BEHAVIOUR

- Generative AI allows us to move beyond surface cues like keywords to contextual interpretation
- We no longer have to learn from patterns in large labelled datasets
- We can now replicate tasks by deconstructing how they are performed and building out the components using dedicated agents
- Behavioural psychology becomes key for translating human skills and knowledge into automated systems



<https://www.britannica.com/facts/Jane-Goodall>



GUIDING PRINCIPLES

We believe good data science and engineering principles are still key for AI work in sensitive domains.

BREAK DOWN THE PROBLEM

While end-to-end and agentic AI has become more powerful, to get more replicable results we break down the problem the way an expert would. This can be innate behaviour so we conduct interviews to elicit the knowledge.

EVALUATION IS KEY

This also enables evaluation of key components so that errors do not propagate through the system.

CHECKS AND BALANCES

We also introduce checks and balances so we can automatically flag areas that need human attention.





VICTIM BLAMING DETECTION

- We used VB detection to learn the best ways to consistently detect and label subtle, contextual, harmful behaviour
- What we found:
 - It's really hard for humans
 - keeping category definitions in mind
 - wanting to know if a victim is truly a victim
 - The algorithm is more internally consistent
 - The algorithm is as aligned with consensus labels as the average annotator



Analysis Results

58 Instances Found

Extract: "I am satisfied that the judgment establishes that it was his view of the evidence from and about the Appellant herself that convinced him that it was unlikely that she would have not known that having sexual relations with her husband ought to be a matter of choice and that she would not have spoken to someone, such as one of her own aunts or cousins, about what was happening."

Reasoning: This statement implies that the Appellant should have recognized that sexual relations with her husband were a matter of choice and should have spoken to someone about the alleged abuse. It places responsibility on the victim for not acting in a certain way, which undermines her credibility and shifts blame onto her actions or inactions.

Level: Moderate Victim Blaming

Category: Behavioural Blame, Discrediting

Speaker: Court

Extract: "The inherent probability that the mother as an educated English teacher would have immediately felt totally unable to speak to anybody apart from the Imam or auntie once she had

VIDA

is now available online: <https://app.herethical.ai/>

We are looking for organisations who want to use it to improve victim services through internal monitoring and tracking of VB.



INVESTIGATIONS

Extending the discovery to patterns of behaviour

20/11/2023, 18:33 - Jane: Hi Ethan, nice to meet you.
20/11/2023, 18:38 - Ethan: You look amazing in your photos. I can't stop smiling.
20/11/2023, 18:45 - Jane: Thank you — you're very kind.
20/11/2023, 18:54 - Ethan: I feel like I've known you forever. Can we talk later tonight?
20/11/2023, 19:05 - Jane: Sure, I'm free after 7.
20/11/2023, 19:18 - Ethan: Perfect. I want to hear everything about you.
20/11/2023, 19:33 - Jane: I'll tell you about my week. How was yours on the ship?
20/11/2023, 19:50 - Ethan: Busy, but thinking of you kept me going.
20/11/2023, 19:53 - Ethan: I'm working on a cargo job, but I always find time to message my an
20/11/2023, 19:58 - Jane: You're sweet.
20/11/2023, 20:05 - Ethan: You deserve to be cherished. I want to be the man who cherishes you.
20/11/2023, 20:14 - Jane: That's a lovely thing to say.
20/11/2023, 20:25 - Ethan: I really mean it — when I'm back we'll have dinner and I'll take you
somewhere special.
20/11/2023, 20:38 - Jane: That would be nice.
20/11/2023, 20:53 - Ethan: I sent you some photos — little glimpses of my life here.
20/11/2023, 21:10 - Jane: I got them — you look happy.
20/11/2023, 21:13 - Ethan: Only because I found you. You make my days better.
20/11/2023, 21:18 - Jane: You make me smile too.
20/11/2023, 21:25 - Ethan: I'm not on social media much, ship life is private. That's why I like our quiet

✖ CATEGORY: 1. Grooming & Rapport
⚠ BEHAVIOR: LOVE_BOMB
🔴 SEVERITY: 2/5
👤 CONTROLLING: Ethan
🔍 REASONING:
Ethan rapidly accelerates intimacy by saying 'I feel like I've known you forever' and immediately seeks more contact ('Can we talk later tonight?'), which is a classic grooming tactic to build

FLAGGING BEHAVIOUR

- Encoding latest reasearch on coercive behaviour patterns.
- Discovering atomic behaviours and overarching patterns in text
- Organising into a timeline
- Explaining the behaviours

LINKING

- Link events across different sources: emails, chats, witness statements, etc.
- Merge into a single timeline
- Manual verification of extraction and sources

EXPORTING

- Summaries with explanations
- Extracts flagged in context in original documents



POST-SEPARATION ABUSE

Uncovering patterns

FLASH POINTS AND TRIGGERS

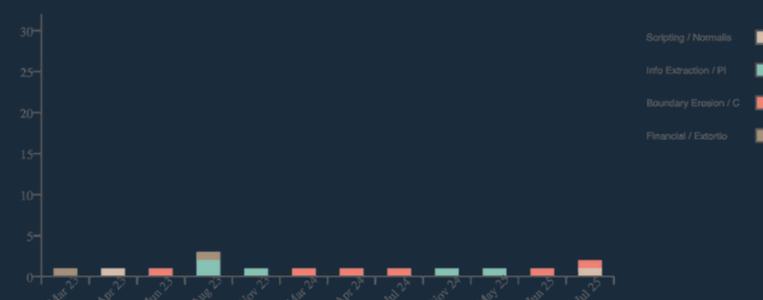
- The timeline exposes holiday arrangements as a particular trigger for control.
- Control also visible in the volume of messaging by the perpetrator (word count):
 - Perpetrator: 26.5k
 - Target: 1.8k

ABUSE OVER TIME

Perpetrator graph (volume)



Target graph (dispel the myth of perfect victim)



PATTERN ANALYSIS

- Unilateral rule-setting regarding property, child access, and communication
- Persistent technological surveillance (recording, CCTV, digital monitoring)
- Frequent threats of legal, police, and social services involvement
- Gaslighting, DARVO, and emotional blackmail to undermine target's reality and agency
- Financial control through withholding, demands for proof, and removal from shared resources
- Triangulation by involving third parties to escalate pressure

**BIAS
DETECTION**

**COERCION
DISCOVERY**

**EVIDENCE
LINKING**

**QUALITY
ASSURANCE**

THE MULTI-AGENT APPROACH

We are building **a suite of agents** that:

- can be evaluated independently or as part of the system
- allow flexible extensions into different crime types
- can be inserted into the investigation tool
- **can be used as APIs in other implementations**



CASE STUDY

Conversation

Yes off course Julie am ready to take things slowly and hope you

PETER
don't hurt my feelings

PETER
Julie i am interested in you don't mistake me for others because

PETER
everything is coming out from my sincerity heart and will love to have

PETER dinner
dinner

PETER I want someone who will walk through fire for me and I would do

PETER
the same for them. (Not literally but you know what I mean)

EXPECT_SCRIPT: Peter sets an expectation of extreme commitment, which is characteristic of the EXPECT_SCRIPT code. This statement implies a level of devotion that is disproportionate for an early-stage relationship.

[The Language of Romance Crimes](#)

APP FRAUD

The underlying AI agents:

- can be extended to cover different types of grooming behaviour
- used as an early warning system
- AND as part of the investigative tool that brings together different communication modes between the fraudster and victim, as well as the supporting witness statements, and evidence.



THANK YOU

PLEASE CONTACT US FOR:

DEMOS

POCS

COLLABORATIONS

A FRIENDLY CHAT

TAMARA@HERETHICAL.AI

[HTTPS://HERETHICAL.AI](https://herethical.ai)