MODEL OVERVIEW

Behavioural Model for Cyber Security

An interactive model of the cognitive, behavioural, social and organisational sides of cyber risk, with human, AI agent and convergent lenses.

IN THIS DOCUMENT

Why behaviour matters 2

HCRM alignment 3

A multi-layered interactive model mapping **cognitive, behavioural, social,** and **organisational** dimensions of cyber risk and resilience - applied to **human and AI agent** behaviour.

Why Behaviour Matters in Cyber Security

Technology alone cannot secure an organisation. Research consistently shows that **human behaviour is the decisive factor** in the vast majority of security incidents, from clicking phishing links and reusing passwords to ignoring policy and falling for social engineering. Yet most security programmes still focus overwhelmingly on technical controls, treating people as the problem rather than the solution.

The **Behavioural Model for Cyber Security** changes that. Developed by CyBehave and grounded in evidence-based behavioural science, including the COM-B model, the Behaviour Change Wheel, Nudge Theory, and Protection Motivation Theory, this interactive framework provides a structured, multi-layered approach to understanding *why* people behave the way they do in security contexts, and *how* to design interventions that actually change behaviour.

Critically, as autonomous AI agents become embedded in enterprise workflows, the same behavioural questions apply to them. AI agents exhibit functional analogues of cognitive bias, habit formation, authority compliance, and social norm adoption. CyBehave is developing what we are calling **Behavioural Convergence Theory (BCT)**, an emerging body of research that investigates whether and how established human behavioural science frameworks can be meaningfully extended to govern AI agent behaviour. This work is ongoing, but the evidence base is no longer only our own. Since 2025, independent multi-agent systems research has reported conformity thresholds, authority effects, spontaneous conventions, endogenous norm formation and pluralistic ignorance in populations of LLM agents, and has reached for the original human frameworks to explain them. Human behavioural science is proving to be the most robust existing toolkit for understanding agentic AI risk. Read the evidence

What This Model Provides

The interactive visualisation below maps **16 behavioural factors** across four concentric layers, each representing a different dimension of cyber security behaviour. You can overlay threat vectors to see where attacks exploit behavioural weaknesses, intervention functions to explore evidence-based approaches to behaviour change, and measurement dimensions to assess organisational maturity. The three-lens system lets you examine each factor from a human, AI agent, or convergent perspective.

Alignment to Human Cyber Risk Management (HCRM)

Each behavioural factor in this model is assessed for its alignment to established **Human Cyber Risk Management (HCRM)** theories, models, and frameworks, the body of academic and practitioner research that underpins how we understand human behaviour in security contexts. As part of CyBehave's ongoing research into **Behavioural Convergence Theory**, we evaluate how directly these HCRM concepts translate to AI agent behaviour. When viewing factors through the **AI Agent** or **Convergent** lens, each factor displays an applicability badge indicating the strength of this alignment:

Strong Analogy

The HCRM concept has a direct, well-evidenced functional equivalent in AI agent behaviour. The underlying mechanism differs, but the observable outcome and security implications are structurally parallel. For example, *Risk Perception* in humans maps directly to *Threat Scoring* in AI: both produce assessments systematically biased by prior exposure.

Adapted

The HCRM concept requires meaningful adaptation to apply to AI agents. The human framing from established theories does not transfer literally, but a functionally analogous process exists when reinterpreted

through an agentic lens. For example, *Motivation* in humans becomes *Objective Functions* in AI: not the same mechanism, but producing equivalent behavioural outcomes when misaligned.

Limited Analogy

The analogy between the HCRM framework and AI agent behaviour is partial or metaphorical. The concept applies to AI agents only in a loose structural sense, and the human framing must be substantially reinterpreted. These factors represent the frontier of CyBehave's BCT research and require the most careful handling when designing cross-domain interventions.

Cognitive

How individuals perceive, process, and evaluate security risks under cognitive constraints

Behavioural

COM-B framework: the capability, opportunity, and motivation driving secure actions

Social

Group norms, authority dynamics, and cultural influences that shape collective security behaviour

Organisational

Policies, nudges, training architecture, and incident response systems that structure behaviour

HCRM Alignment Summary

Across all 16 behavioural factors in this model, CyBehave's Behavioural Convergence Theory research evaluates how directly established Human Cyber Risk Management concepts translate to AI agent behaviour.

11

STRONG ANALOGY

69% of factors

5

ADAPTED

31% of factors

0

LIMITED ANALOGY

0% of factors

Strong Analogy

LAYER	HUMAN FACTOR	AI EQUIVALENT
Cognitive	Risk Perception	Threat Scoring
Cognitive	Mental Models	Learned Representations
Cognitive	Cognitive Load	Context Constraints
Behavioural	Capability	Agent Capability
Behavioural	Opportunity	Environmental Permissions

LAYER	HUMAN FACTOR	AI EQUIVALENT
Behavioural	Habit Formation	Learned Defaults
Social	Authority Compliance	Instruction Following
Organisational	Policy Architecture	Governance Frameworks
Organisational	Nudge Design	Prompt & Default Engineering
Organisational	Training & Awareness	Alignment & Fine-tuning
Organisational	Incident Response	Kill Switches & Rollback

Adapted

LAYER	HUMAN FACTOR	AI EQUIVALENT
Cognitive	Decision Fatigue	Inference Degradation
Behavioural	Motivation	Objective Functions
Social	Social Norms	Emergent Agent Conventions
Social	Security Champions	Sentinel Agents
Social	Culture & Climate	System-Level Norms

Key Insight

The **Social layer** has the highest concentration of adapted factors (3 of 4), reflecting that social dynamics required the most reinterpretation when the model was first built. It is also now the layer with the strongest independent empirical support: recent multi-agent research has reproduced threshold conformity, normative influence, authority effects, spontaneous conventions and pluralistic ignorance in LLM agent populations. These classifications are under review and several may move to Strong Analogy. The **Organisational layer** is entirely strong analogy, reflecting that governance structures translate most directly to agentic AI systems.

See the evidence behind the Social layer

About this document

This document was built on 25 November 2026 from the page at cybehave.com/behavioural-model, so it carries the guidance as it was published that day. The page is kept up to date; download a fresh copy whenever you need one. This overview covers the reasoning and the alignment summary. The model itself is interactive: on the page you can open each of the sixteen factors, switch between the human, AI agent and convergent lenses, and overlay threats, interventions and measurement.

This page cybehave.com/behavioural-model

Free Champions Programme Health Check cybehave.com/champions-health-check/

The Rise of the Security Champion, free ebook cybehave.com/the-rise-of-the-security-champion

HEROES by CyBehave, free 28-day trial cybehave.com/trial

© 2026 CyBehave Ltd. You are welcome to share this document with colleagues. Please keep it whole and credit CyBehave.