The Psychological Panopticon

Inferential Psychometrics, Multimodal Behavioural Telemetry, and the Contested Future of Cognitive Liberty

A Dissertation

Waitman Gobble · HYPHERO · 2026-07-26

~DRAFT~

Table of Contents

Abstract

This dissertation examines the emergence of a novel form of psychological assessment in which established psychometric constructs are inferred, without consent or instrumentation, from the residue of ordinary human interaction with artificial intelligence systems. It argues that the conversational interface has become an unintended psychometric instrument of considerable power, and that this transformation carries consequences for privacy, governance, and human autonomy that existing legal and ethical frameworks are poorly equipped to address.

The argument proceeds in four movements. The first establishes the theoretical foundations of psychological measurement, surveying the constructs — personality, cognition, affect, motivation, and social attachment — that any inferential system must ultimately attempt to recover, alongside the clinical instruments developed to measure them. The second demonstrates how a longitudinal corpus of user prompts can be processed to generate *quasi-psychometric* estimates of these same constructs, tracing both the classical natural language processing pipeline and its displacement by large language models capable of holistic contextual inference. The third extends this analysis to multimodal telemetry — acoustic biomarkers, keystroke dynamics, and computer vision — and argues that the convergence of these channels renders deliberate concealment of psychological state increasingly untenable. The fourth turns to consequences: the application of such profiling to statecraft, the resulting architecture herein termed the *Psychological Panopticon*, and the mathematical and legal countermeasures available to resist it.

Throughout, this dissertation maintains a deliberately sceptical posture toward the validity claims implicit in the technology it describes. A recurring theme is what is here called the **context–intent disconnect**: the systematic failure of behavioural inference systems to distinguish playful, performative, or contextually bracketed human conduct from genuine psychological pathology. This failure is not incidental. It is argued that the capability required to resolve it — a genuine machine model of human social context — is precisely the capability that would make such systems maximally invasive. The technology is therefore dangerous when it is unsophisticated and dangerous, differently, when it is not.

The dissertation concludes that differential privacy, and specifically local differential privacy operating under a constrained privacy budget, represents the most credible technical constraint on the architecture described. It further argues that technical constraint alone is insufficient, and that a positive legal right to cognitive liberty — the right to mental privacy and to freedom from subconscious algorithmic manipulation — will need to be articulated as a distinct category of human right rather than derived from existing data-protection regimes.

Keywords: digital phenotyping; computational psychometrics; differential privacy; cognitive liberty; algorithmic governance; stylometry; multimodal inference; surveillance theory

Chapter 1 — Introduction

1.1 Problem Statement

The psychometric tradition has, for the better part of a century, rested on a foundational assumption: that psychological measurement is an *event*. A subject is recruited, briefed, and seated. An instrument is administered under controlled conditions. Responses are scored against normative data, and inferences are drawn within stated confidence intervals. The subject knows they are being measured; indeed, the ethical legitimacy of the enterprise depends on it.

Conversational artificial intelligence has quietly dissolved this assumption. When a user interacts with a large language model across weeks, months, or years, they generate a longitudinal record of spontaneous, unstructured, self-initiated behaviour of a kind that no laboratory protocol has ever been able to obtain. The record is unprompted in the psychometric sense: no item was presented, no forced choice was imposed, no experimenter was present to induce demand characteristics. It is, in effect, a diary of cognitive process, emotional state, and problem-solving strategy — with the critical difference that a diary is written with awareness of an eventual reader, whereas a prompt is typically composed with the assumption that it is instrumental, transient, and unremarkable.

This dissertation takes as its central problem the fact that such a record is psychometrically tractable. Established techniques in computational linguistics — lexical analysis, stylometric fingerprinting, semantic embedding, time-series anomaly detection — permit the reconstruction of approximations to formal psychological constructs from this material. The resulting estimates are not clinical diagnoses and should not be mistaken for them. They are, however, sufficient to support consequential decisions about individuals, and they are being generated in a domain where none of the procedural safeguards that govern clinical assessment apply.

1.2 Research Questions

This dissertation addresses five questions:

1. **RQ1.** Which established psychological constructs are recoverable, and at what fidelity, from a longitudinal corpus of unstructured human–AI interaction?
2. **RQ2.** How does the methodological transition from classical NLP pipelines to large language models alter both the accuracy and the auditability of such inference?
3. **RQ3.** What additional inferential capability is conferred by multimodal telemetry — acoustic, kinematic, and visual — and does its introduction represent a difference of degree or of kind?
4. **RQ4.** Under what conditions does behavioural inference fail, and what does the character of its failures reveal about the limits of the enterprise?
5. **RQ5.** What technical and legal instruments could constrain the deployment of such inference by state and corporate actors, and what are their limits?

1.3 Scope and Delimitations

Several delimitations should be stated at the outset.

First, this dissertation is concerned with *inference from interaction traces*, not with the separate and better-studied question of AI-administered formal assessment. A system that presents a subject with NEO-PI-R items and scores the responses raises important questions, but they are not the questions considered here.

Second, the analysis is architectural rather than jurisdictional. Where legal frameworks are discussed, they are discussed as categories of instrument rather than as the specific provisions of any national regime, which vary considerably and change rapidly.

Third, and most importantly, this dissertation does not assert that the inferential capabilities it describes are uniformly *valid*. A substantial portion of the argument — particularly Chapters 3, 6, and 7 — is devoted to demonstrating that they are not, or are not yet, and that the gap between claimed and actual validity is itself a source of harm. The dissertation describes what such systems *do*, which is not always what their proponents claim they do.

1.4 Contribution

The contribution is threefold. The dissertation offers, first, an integrated account of inferential psychometrics that connects the psychological constructs being estimated, the computational machinery performing the estimation, and the governance structures that might consume the output — a chain that is typically treated in disconnected literatures.

Second, it advances the **context–intent disconnect** as an analytic frame for understanding the failure modes of behavioural inference. This frame is developed at length in Chapter 7 and is argued to be more explanatory than the conventional vocabulary of "bias" or "error," because it locates the failure in a structural absence of world-modelling capacity rather than in defective training data.

Third, it identifies what is termed the **competence paradox**: the observation that the technical development required to eliminate the false positives of behavioural inference is identical to the development required to maximise its intrusiveness. This paradox is argued to have direct implications for regulatory strategy, since it suggests that "improve the accuracy of the system" and "reduce the harm of the system" are not, as commonly assumed, the same project.

1.5 Structure of the Dissertation

Chapter 2 establishes the psychological constructs at issue. Chapter 3 examines the prompt corpus as a source of inference about them. Chapter 4 considers what longitudinal duration adds. Chapter 5 details the computational methods. Chapter 6 extends the analysis to multimodal channels. Chapter 7 develops the critique of validity. Chapters 8 and 9 turn to governance and to the panoptic architecture that results. Chapter 10 examines countermeasures. Chapter 11 concludes.

Chapter 2 — Theoretical Foundations of Psychological Measurement

Before considering what may be inferred from interaction traces, it is necessary to establish what is being inferred *toward*. This chapter surveys the principal construct families in contemporary psychological assessment. The survey is deliberately structural rather than exhaustive: the aim is to identify the measurement targets that subsequent chapters will show to be computationally approachable, and to note at each point the assumptions on which their measurement depends.

2.1 The Construct of Personality: The Five-Factor Model

Personality, in the technical sense, denotes the stable and enduring patterns of thought, feeling, and behaviour that persist across time and situation. The distinction between personality and *state* is central: a person may be anxious on a given afternoon without being high in trait anxiety, and the conflation of the two is among the most common errors in applied assessment.

The most empirically validated framework is the Five-Factor Model, conventionally abbreviated as OCEAN (Costa & McCrae, 1992). Its five dimensions are:

Openness to Experience. The degree of intellectual curiosity, aesthetic sensitivity, imaginative capacity, and preference for novelty over familiarity. High scorers seek out unfamiliar ideas and tolerate ambiguity; low scorers prefer the concrete, the conventional, and the established.

Conscientiousness. The tendency to exercise self-discipline, act dutifully, defer gratification, and orient behaviour toward achievement against internal or external standards. It subsumes facets including orderliness, industriousness, and deliberation.

Extraversion. Energy, positive affect, assertiveness, sociability, and the tendency to seek stimulation in the company of others. It should be noted that extraversion is not equivalent to social skill, nor introversion to social anxiety — a conflation that becomes significant when the construct is inferred from digital behaviour.

Agreeableness. The disposition toward compassion, cooperation, and trust rather than suspicion and antagonism. Low agreeableness is not equivalent to hostility; it encompasses scepticism, competitiveness, and a willingness to prioritise one's own interests.

Neuroticism. The disposition to experience negative affect — anger, anxiety, depression, vulnerability — with greater frequency and intensity. Often described as emotional instability, though the term carries pejorative connotations that the construct itself does not warrant.

The five factors are conventionally treated as orthogonal, though empirically they display modest intercorrelation, and higher-order factor structures have been proposed. For present purposes the significant property of the model is its lexical origin: the factors were derived from the structure of natural-language trait descriptors, which is precisely why they have proven unusually amenable to recovery from text.

2.2 Cognitive Parameters

Cognitive assessment concerns the mechanisms by which information is processed, stored, and retrieved. Four parameters are of particular relevance.

Attention and focus. The capacity to selectively concentrate on discrete environmental or mental stimuli while suppressing extraneous distraction. Sustained attention, selective attention, and divided attention are separable capacities with distinct neural substrates.

Memory function. Encompassing sensory, short-term, working, and long-term memory. Working memory — the capacity to hold and manipulate information in an active state — is of special interest here, as it correlates strongly with general cognitive ability and, as Chapter 3 argues, leaves distinctive traces in extended conversational interaction.

Executive functioning. Higher-order processes principally associated with prefrontal cortical activity, including problem decomposition, cognitive flexibility, abstract reasoning, inhibition, and planning. Executive function is not a unitary construct, and its factor structure remains contested.

Processing speed. The latency with which a cognitive operation is performed or a stimulus is responded to. Processing speed declines with fatigue, intoxication, affective disturbance, and age, which makes it a sensitive if non-specific indicator.

2.3 Affective and Emotional Parameters

Affective parameters govern how emotional experience is generated, processed, and modulated.

Emotional regulation. The capacity to monitor, evaluate, and modify emotional reactions in service of a goal — attenuating anger, dampening panic, sustaining engagement despite frustration. Gross's (1998) process model distinguishes antecedent-focused from response-focused strategies, a distinction with implications for what is externally observable.

Affective valence and arousal. Following Russell's (1980) circumplex, affective state is conventionally decomposed into valence (the positive–negative dimension) and arousal (the calm–activated dimension). The two are separable: high-arousal negative affect (panic, rage) is phenomenologically and physiologically distinct from low-arousal negative affect (dejection, apathy), though both are "negative."

Empathy. The cognitive and affective capacity to model, simulate, and share the emotional states of others. Cognitive empathy (understanding another's state) and affective empathy (sharing it) dissociate, and the distinction matters when empathy is inferred from linguistic behaviour that may reflect only the former.

2.4 Motivational and Volitional Parameters

Motivational parameters address the question of *why* behaviour occurs.

Intrinsic versus extrinsic motivation. Whether behaviour is driven by internal reward — curiosity, mastery, personal satisfaction — or by external contingency such as remuneration, status, or social approval (Deci & Ryan, 1985). The distinction is not binary but continuous, and the same behaviour may be multiply determined.

Locus of control. The degree to which an individual attributes life outcomes to their own agency (internal) or to external forces such as chance, fate, or powerful others (Rotter, 1966). Locus of control is of particular consequence in Chapter 8, since it substantially conditions susceptibility to differently framed persuasive appeals.

Risk tolerance and sensation seeking. The appetite for novel, intense, and complex experience, and the willingness to accept physical, financial, or social risk in its pursuit (Zuckerman, 1994).

2.5 Social and Relational Parameters

Attachment style. The characteristic pattern by which an individual forms and maintains close bonds — conventionally categorised as secure, anxious-preoccupied, dismissive-avoidant, or fearful-avoidant (Bowlby, 1969; Ainsworth et al., 1978). Attachment patterns are held to form in early caregiving relationships and to display substantial though not complete continuity into adult relationships.

Social connectivity and cognitive constraint. The baseline drive toward kinship and community, together with the cognitive limit on the number of stable social relationships an individual can maintain. Dunbar (1992) proposed a characteristic figure of approximately 150 meaningful relationships on the basis of neocortical volume in primates. The figure is widely cited and should be treated with corresponding caution; the underlying regression is drawn from a small comparative sample and the specific numeric value has been contested.

2.6 Clinical Psychometrics: The MMPI Clinical Scales

Where the preceding sections describe normal-range variation, clinical psychometrics addresses psychopathology. The Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943; Butcher et al., 1989) remains the most widely administered instrument of its kind. Its construction is empirically keyed rather than theoretically derived: items were retained not because they appeared meaningful but because they discriminated between diagnostic and control groups.

Responses are scored against ten basic clinical scales:

Scale	Designation	Construct Measured
1	Hs — Hypochondriasis	Preoccupation with bodily

Scale	Designation	Construct Measured
		function and disproportionate concern regarding physical health
2	D — Depression	Poor morale, hopelessness, dissatisfaction with life circumstances
3	Hy — Hysteria	Tendency toward somatic symptom development as an involuntary response to stress
4	Pd — Psychopathic Deviate	Social alienation, impulse control difficulty, disregard for social convention
5	Mf — Masculinity–Femininity	Rigidity of alignment with stereotypical gender-role expectations
6	Pa — Paranoia	Interpersonal hypersensitivity, suspiciousness, rigidity of thought
7	Pt — Psychasthenia	Obsessive-compulsive features, excessive doubt, generalised anxiety
8	Sc — Schizophrenia	Thought disturbance, unusual perceptual experience, profound social alienation
9	Ma — Hypomania	Overactivation, pressured speech, elevated mood, unstable energy
0	Si — Social Introversion	Shyness, social withdrawal, preference for solitude

Two features of the instrument require emphasis. The scale names are historical artefacts of mid-twentieth-century nosology and do not correspond to current diagnostic categories; an elevation on Scale 8 does not indicate schizophrenia. Further, the instrument incorporates validity scales designed to detect defensive responding, symptom exaggeration, and inconsistent responding — a recognition, built into the instrument's own architecture, that self-report is systematically unreliable.

2.7 The Structural Limitations of Self-Report

The instruments surveyed above share a common dependency: they require the subject to report on themselves. This dependency generates well-documented distortions. Social desirability bias leads subjects to under-report stigmatised attributes. Limited introspective access means that subjects frequently cannot report accurately even when motivated to do so. Reference-group effects mean that a subject's judgement of whether they are "often anxious" is calibrated against an idiosyncratic and unobservable comparison class. And the assessment context itself — the knowledge that one is being evaluated, and for what purpose — reliably shifts responding.

It is precisely these limitations that make behavioural inference attractive. A record of spontaneous behaviour is not subject to social desirability bias in the same way, because the behaviour was not produced for an evaluator. This is the central promise of the methods examined in the chapters that follow. It is also, as will be argued, the source of their principal ethical difficulty: the same property that makes the data psychometrically valuable — that it was generated without awareness of assessment — is what makes its use without consent objectionable.

Chapter 3 — The Prompt History as Behavioural Corpus

3.1 Properties of the Corpus

A prompt history possesses four properties that distinguish it from the text corpora on which computational personality research has conventionally relied.

It is **longitudinal**. Where social media studies typically sample a bounded window of posts, a prompt history accumulates continuously across the entire period of platform use, frequently spanning years.

It is **unstructured and spontaneous**. Prompts are composed to accomplish a task, not to present a self. Unlike a social media post, which is authored with an audience in mind and edited accordingly, a prompt is typically composed once, quickly, and without revision.

It is **instrumentally revealing**. A prompt does not merely express a state; it discloses a *problem*. The user tells the system what they do not know, what they are struggling with, what they are afraid of, and what they intend to do. Few other behavioural records so directly expose the gaps in a person's competence and the objects of their concern.

It is **metadata-rich**. Every prompt carries a timestamp, a session boundary, an inter-prompt interval, and — depending on the interface — device, locale, and modality information. This metadata is frequently more diagnostic than the text itself.

By applying computational linguistics to this corpus, a researcher may generate approximations to formal psychometric constructs without administering any instrument. The remainder of this chapter examines how, taking each construct family in turn. The prefix *quasi-* is used throughout to mark the resulting estimates as structurally analogous to, but not equivalent to, the validated constructs of Chapter 2.

3.2 Inferring Quasi-OCEAN Traits

Text-based personality inference proceeds along two complementary axes: **lexical density**, concerning which words are used, typically operationalised through psycholinguistic dictionaries such as LIWC (Tausczik & Pennebaker, 2010); and **semantic styling**, concerning the structure, register, and organisation of the utterance. Open-vocabulary approaches (Schwartz et al., 2013) supplement dictionary methods by identifying discriminative terms empirically rather than by prior classification.

Openness to Experience is indicated by vocabulary breadth, abstraction, and metaphorical density. At the level of pragmatics, high-openness users generate prompts that explore counterfactual and hypothetical space ("what would follow if..."), pose philosophical or definitional questions without instrumental purpose, request creative composition, and seek explanation of unfamiliar or specialised domains for no evident practical reason. The signature is curiosity untethered from task.

Conscientiousness is indicated principally by structure rather than content. High-conscientiousness users compose organised, goal-directed prompts, supply structured context in advance in order to constrain the output, and specify format requirements explicitly. Their lexicon is weighted toward scheduling, efficiency, and measurement. Prompts are rarely fragmentary. Where a low-conscientiousness user iterates toward a result through successive corrections, a high-conscientiousness user front-loads specification.

Extraversion, being an inherently interpersonal construct, translates imperfectly into a solitary human–machine channel. Its digital correlates are conversational register, informality, exclamatory punctuation, and relational vocabulary. A distinctive marker is *personification*: the tendency to address the system as an interlocutor rather than as a command interface — greeting it, thanking it, engaging in phatic exchange. It should be noted that this marker is confounded with cultural norms of politeness and with the user's beliefs about the system's nature, and is therefore among the weaker indicators in the set.

Agreeableness is approached primarily through sentiment and pragmatic framing. Indicators include mitigated request forms ("could you possibly..."), explicit politeness markers, absence of aggressive or demanding phrasing, and — more diagnostically — the *content* of interpersonal queries. A user who asks how to decline an invitation without causing offence has revealed a concern for another's feelings that is not reducible to surface politeness.

Neuroticism is indicated by the frequency and intensity of negative-affect vocabulary, but more reliably by three behavioural patterns: repetitive reassurance-seeking on a resolved question; catastrophe-oriented problem framing, in which the user requests analysis of worst-case outcomes disproportionate to their probability; and volatility of affective tone across sessions.

3.3 Behavioural Proxy Tracking and Quasi-Clinical Scales

The MMPI relies on forced-choice self-report. A prompt history contains no such items. What it contains instead are behavioural indicators that co-vary with the constructs the clinical scales were keyed to detect — an approach termed here **behavioural proxy tracking**.

Hypochondriasis and Psychasthenia (Scales 1 and 7). Indicated by the volume, repetition, and intensity of health- and safety-related querying. Repeated requests to interpret minor somatic sensations, sustained engagement with medical terminology in the absence of professional context, and iterative reassurance-seeking regarding risk map onto these dimensions.

Depression (Scale 2). Indicated by linguistic markers of anergia, social withdrawal, and negative outcome expectancy. Characteristic patterns include a shift toward passive, low-effort querying; requests for comfort rather than information; and thematic preoccupation with lethargy, sleep disturbance, or existential futility. The *shift* is more diagnostic than the level, a point developed in Chapter 4.

Psychopathic Deviate and Hypomania (Scales 4 and 9). Indicated by high-impulse, non-conformist, or erratic interaction patterns: sustained attempts to circumvent system constraints, requests for procedural shortcuts around institutional rules, rapid non-sequitur transitions between unrelated complex topics, and clustering of high-arousal or risk-oriented querying in late-night intervals.

Social Introversion (Scale 0). Approached through the *function* the system serves for the user. Where the system is used as a substitute for human interlocution — recounting the day's events, processing interpersonal difficulty, seeking emotional response — the pattern correlates with social withdrawal or alienation.

This last inference warrants particular scepticism. Substitution of AI for human interlocution may reflect introversion; it may equally reflect geographic isolation, disability, caregiving burden, unsociable working hours, recent bereavement, or simple convenience. The behavioural signature is identical across these causes. This is a specific instance of a general problem addressed in Section 3.5.

3.4 Estimating Cognitive Function

A prompt history cannot yield a formal intelligence quotient, which is by definition a normed score derived from standardised administration. It does, however, contain reasonably direct indicators of several component capacities.

Crystallised intelligence is approached through lexical and syntactic complexity. Readability indices such as the Flesch–Kincaid grade level (Kincaid et al., 1975), together with type–token ratio and the frequency distribution of rare terms, provide a proxy for acquired verbal knowledge. Precise use of technical nomenclature within a specialised domain is a strong indicator of expertise in that domain, though a weak indicator of general ability.

Executive function and abstract reasoning are indicated by problem structuring. A user who supplies a dataset, specifies constraints on the solution space, decomposes a task into ordered sub-goals, and requests multi-step synthesis has demonstrated planning and abstraction that a simple query does not.

Working memory leaves a distinctive trace in extended conversational threads. Users of high working-memory capacity treat a thread as a persistent workspace, referring back to earlier turns by index ("apply the constraint from the third step to the variable established at the outset"), maintaining a coherent schema across dozens of exchanges without restatement. The capacity to sustain and manipulate a large representation across an extended interaction maps directly onto the construct.

Processing speed and cognitive load are approached through metadata rather than text — inter-prompt latency, revision frequency, and abandonment rate. These are treated more fully in Chapter 6.

3.5 Threats to Validity

Four threats to validity are structural rather than incidental, and no amount of additional data resolves them.

Persona bias. Users frequently adopt a deliberate register when interacting with AI systems — professional, terse, formal — which masks baseline traits. The measured construct is then the persona, not the person.

Context scarcity. A prompt history records what was typed, not why. A user generating extensive querying about severe depressive symptomatology may be experiencing depression, supporting a family member who is, writing a novel, studying for a clinical examination, or fact-checking a news article. The linguistic surface is substantially identical across all five. This is the most serious threat to the enterprise and recurs, in more acute form, in Chapter 7.

Instrumental confounding. Prompts are shaped by the user's model of what the *system* responds to well, not solely by the user's own dispositions. A user who has learned that structured prompts yield better output will produce structured prompts irrespective of their conscientiousness. As users become more sophisticated, this confound grows rather than diminishes, and it does so non-uniformly across the population — meaning it introduces systematic rather than random error.

Population non-equivalence. Normative data for validated instruments derives from defined reference populations. No such normative base exists for prompt behaviour, and the population of AI users is not demographically representative. Any statement that a user is "high" on an inferred dimension is therefore a statement relative to an undefined and shifting comparison class.

These threats do not render inference impossible. They render it *probabilistic and non-clinical* — a distinction that is well understood by researchers and routinely elided in deployment.

Chapter 4 — The Longitudinal Dimension

The preceding chapter treated the prompt history as a static corpus. This chapter argues that duration transforms it qualitatively. A history spanning a year or more does not merely provide more of the same data; it changes the nature of what is inferable, converting a snapshot of utility into a high-fidelity behavioural profile.

4.1 Duration as a Solvent for Situational Bias

Short-term histories suffer acute situational bias. A user who spends a fortnight querying the symptoms of acute anxiety may be flagged, on a short-window analysis, as elevated on neuroticism and on quasi-Scale 7. The inference is unwarranted; the behaviour may be entirely transient.

Extended observation dissolves this ambiguity in two ways.

Baseline establishment. Across a sufficient interval, the analysis can determine whether an episode of anxiety-related querying was a bounded reaction to an identifiable stressor — an examination period, a redundancy, a bereavement — or a chronic and recurring pattern of reassurance-seeking. Trait and state become separable precisely because trait is defined by temporal stability.

Persona decay. A user may sustain a deliberate professional register for weeks. Across a year, this becomes difficult. Fatigue, familiarity, and the sheer volume of interaction ensure that spontaneous, fragmentary, late-night, or emotionally unguarded prompts enter the record. The persona is not so much penetrated as *outlasted*. In statistical terms, the deliberate presentation contributes decreasing marginal weight as the corpus grows.

4.2 Stylometry and the Authorial Fingerprint

Across extended interaction, an individual manifests a distinctive and largely involuntary linguistic signature. Stylometry — established in authorship attribution since Mosteller and Wallace's (1964) analysis of the disputed *Federalist* papers — provides the analytic apparatus.

Function word distribution. Articles, prepositions, pronouns, and conjunctions are deployed almost entirely below the level of conscious control. Their relative frequencies are remarkably stable within individuals and remarkably variable between them. Because they carry minimal semantic content, they are invariant to topic, which makes them the single most reliable stylometric feature.

Syntactic sequencing. Users develop fixed habits in how they organise an instruction. One consistently orders prompts as context, then constraint, then directive; another opens with the directive and appends qualification. These orderings persist across topic and register.

Orthographic and formatting habit. Line-break usage, list construction, capitalisation for emphasis, ellipsis, and emoji selection constitute a further signature layer.

The consequence is significant and underappreciated: a longitudinal prompt history supports **cross-account and cross-platform identification**. A user who creates a fresh account under a pseudonym carries their function-word distribution with them. Pseudonymity, in this setting, is not anonymity.

4.3 Trajectory Tracking and Psychological Divergence

Extended observation permits a shift from *static typing* — what is this person like — to *dynamic trajectory tracking* — how is this person changing. This is the domain of digital phenotyping (Onnela & Rauch, 2016), in which shifts in device-mediated behaviour are treated as indicators of shifts in mental state.

```
[ STABLE LONGITUDINAL BASELINE ]  ──►  [ LINGUISTIC SHIFT ]  ──►  [ INFERRED DIVERGENCE ]

  · balanced syntactic structure       · function-word attrition     · depressive episode
  · varied lexical range               · repetitive looping          · cognitive decline
  · structured, planful queries        · erratic circadian timing    · hypomanic phase
```

Cognitive decline or fatigue. Indicated by sustained reduction in vocabulary variance, downward drift in readability level, and degraded capacity to maintain contextual threads across a conversation. The user who begins forgetting what was established five turns earlier is exhibiting a measurable change.

Depressive trajectory. Indicated by systematic migration toward low-effort queries — the transition from complex analytical requests to requests for basic summarisation — combined with flattened affective valence and a shift in temporal orientation from prospective planning toward retrospective rumination.

Hypomanic or manic divergence. Indicated by abrupt departure from an established structured baseline into rapid, disjointed querying across unrelated domains, particularly in conjunction with a shift toward late-night timestamps and elevated prompt velocity.

The methodological point deserving emphasis is that each of these inferences is *relative to the individual's own baseline* rather than to a population norm. This is both the source of the method's sensitivity and a partial answer to the population non-equivalence problem raised in Section 3.5 — though it introduces a fresh difficulty, namely that a user whose baseline is established during an episode of illness will have that state encoded as their normal condition.

4.4 Research Ethics and the Corporate–Academic Asymmetry

Longitudinal prompt data is intimately revealing, and academic researchers operating under institutional review face substantial obstacles to its use.

The informed consent dilemma. Institutional review will ordinarily require explicit, informed, specific consent. Yet the knowledge that one's prompts are under psychological analysis reliably alters prompting behaviour — a straightforward instance of the Hawthorne effect (Adair, 1984). The dataset's principal virtue, its naturalism, is destroyed by the procedure that renders its use permissible. The methodologically ideal study is the ethically impermissible one.

De-identification failure. Removing identifiers from a year of prompt history is not reliably achievable. Users type real names, addresses, employer details, project specifics, medical history, and highly particular personal anecdote. Standard scrubbing tools fail against this material, and the broader literature on re-identification of sparse high-dimensional datasets (Narayanan & Shmatikov, 2008) suggests the failure is structural rather than a matter of insufficiently rigorous scrubbing. A prompt history is, in the relevant technical sense, sparse and high-dimensional; the stylometric findings of Section 4.2 compound the problem, since the writing itself is identifying even after named entities are removed.

The corporate–academic asymmetry. This is the most consequential observation in the chapter. An academic researcher must satisfy institutional review, obtain informed consent, justify data minimisation, and submit to oversight. A commercial platform operating under an end-user licence agreement is subject to none of these constraints in the same form. The platform already holds the longitudinal corpus as an operational necessity, and may perform internal profiling for product development, personalisation, or risk assessment under terms the user has nominally accepted.

The asymmetry produces a perverse allocation. The party subject to the most rigorous ethical scrutiny — the academic — has the least access and the least capacity to act on findings. The party with comprehensive access, technical capability, and commercial incentive operates under contractual rather than ethical governance. Research ethics regimes were designed to constrain researchers; they were not designed for a world in which the most extensive human-subjects research is conducted by parties who do not describe themselves as researchers.

Chapter 5 — Computational Methodology

5.1 The Classical Pipeline

A conventional NLP pipeline treats a longitudinal prompt history as a chronological stream of behavioural telemetry. It does not read the text for meaning; it decomposes it into vectors, counts, and metadata tags. The canonical architecture proceeds in five stages.

Stage 1 — Ingestion, Metadata Extraction, and Filtering

Before any linguistic analysis, environmental metadata is extracted to establish behavioural rhythm.

Temporal anchoring. Each prompt carries a timestamp from which circadian distribution is computed — the proportion of activity falling within conventional working hours against the proportion occurring in the small hours — together with prompt velocity, the interval distribution between successive inputs.

Systemic noise filtering. Automated routines strip material that does not constitute the user's authentic linguistic production: pasted code blocks, terminal output, quoted articles submitted for summarisation, and templated boilerplate. Only text composed by the user is retained. This stage is more consequential than it appears — over- or under-aggressive filtering systematically biases every downstream measure, and the filtering heuristics themselves are rarely reported in published work.

Stage 2 — Tokenisation and Normalisation

The retained text is decomposed into machine-tractable units. Tokenisation segments the stream into words and punctuation. Lemmatisation reduces inflected forms to their dictionary base, ensuring that grammatically obligatory tense variation is not misread as lexical change.

Stage 3 — Feature Extraction

The normalised text is passed through parallel engines that convert it into quantifiable metrics.

Lexical and semantic analysis — the "what." Psycholinguistic dictionaries such as LIWC match tokens against predefined affective and cognitive categories. The sentence "I am extremely anxious about the review next week" registers first-person singular pronoun use, negative-affect and anxiety category membership, and prospective temporal orientation. Contextual embedding models such as BERT (Devlin et al., 2019) convert phrases into high-dimensional vectors capturing underlying theme, permitting topical preoccupation to be tracked even as surface vocabulary varies.

Syntactic and stylometric tracking — the "how." Part-of-speech tagging yields the ratio of verbs to nouns to modifiers; a spike in verb density relative to nominal density is conventionally read as indicating high-arousal or action-oriented states. Readability indices and type–token ratio measure syntactic complexity and lexical diversity, providing the cognitive baseline against which Section 4.3's decline detection operates.

Stage 4 — Time-Series Aggregation

To detect trend, the history is segmented rather than averaged in bulk.

```
Raw Prompt Logs ──► [ 7-Day Sliding Window ] ──► [ Feature Extraction ]
                                                          │
                                                          ▼
                                             [ Chronological Profile Array ]
                                                Week 1: [ 0.82, 12.4, 0.05, … ]
                                                Week 2: [ 0.79, 12.1, 0.06, … ]
                                                Week 3: [ 0.71, 11.8, 0.09, … ]
```

Rolling windows of seven or thirty days are generated, and for each window a single summarised vector is compiled containing mean affective valence, readability score, pronoun density, nocturnal prompting frequency, and comparable measures. Window length is a consequential and frequently arbitrary methodological choice: short windows are sensitive but noisy, long windows are stable but may mask the transitions of interest.

Stage 5 — Divergence and Anomaly Detection

Finally, each window is compared against the individual's historical baseline.

Z-score anomaly detection. Where a user's twelve-month mean for negative-affect vocabulary is consistently low and a seven-day window departs by three standard deviations, the pipeline flags a candidate acute stress event.

Changepoint detection. Algorithms such as PELT (Killick et al., 2012) identify *permanent* rather than transient shifts. Where mean sentence complexity drops after month eight and does not recover, the pipeline identifies a statistically significant changepoint — a candidate indicator of chronic burnout, sustained affective disturbance, or neurological change.

5.2 The Large Language Model as Holistic Inferential Engine

Large language models have altered this methodology fundamentally. Where the classical pipeline operates as an assembly line of rigid filters counting words and tagging parts of speech, an LLM operates more nearly as an experienced clinician reading a complete case record. Rather than isolating features, it processes the contextual ecology of the history as a whole.

From bag-of-words to contextual subtext. Dictionary methods are defeated by idiom. The sentence "I am absolutely killing it at work today" registers, under a naïve lexical match, as aggressive or negatively valenced. An LLM recovers the colloquial sense immediately, correctly inferring high self-efficacy and positive mood. More significantly, an LLM can infer implicit affective state where explicit affective vocabulary is entirely absent. A user who composes a lengthy, hyper-detailed, defensively argued account of a trivial workplace exchange need never write the word "anxious" for the model to infer elevated interpersonal stress and suspicion.

Holistic evaluation across long context. Classical pipelines handle longitudinal data by chunking it into artificial windows and averaging, which destroys the connective tissue between temporally distant material. Contemporary models accommodate context windows sufficient to hold very large portions of a history simultaneously, permitting evaluation of the record as a continuous narrative. This enables recognition of extended behavioural callbacks — the identification, for instance, that a minor conversational fixation appearing in the second month has developed by the eleventh into a sustained preoccupation. No windowed average can represent this pattern, because the pattern is constituted by the relationship between distant points rather than by any local property.

Direct in-context profiling. Under the classical approach, a personality estimate requires the researcher to specify dozens of proxy variables — the ratio of first-person singular pronouns as a predictor of one construct, negation frequency as a predictor of another — and to combine them in a secondary statistical model. An LLM can be instructed to perform the assessment directly: to evaluate a history against a named framework and report month-by-month estimates with commentary on shifts in cognitive clarity or affective stability. The intermediate feature-engineering layer is eliminated.

Semantic tracking across surface change. Because LLMs map content into rich semantic space, they track the *underlying object* of a user's preoccupation rather than its lexical expression. A user querying budgeting technique and economical cooking in winter, and debt consolidation and insolvency procedure in spring, presents to a keyword-matching system as having two unrelated interests. An LLM recognises a single accelerating trajectory of financial distress.

5.3 The Explainability Trade-Off

This gain in inferential power is purchased at a cost that is scientific rather than merely technical.

The classical pipeline is fully transparent. A researcher can display a chart and demonstrate exactly why a score diverged: the user's use of anxiety-category verbs rose by a specified percentage, producing a Z-score of a specified magnitude. Every step is inspectable, reproducible, and contestable.

An LLM may correctly identify a shift into a depressive episode with greater sensitivity than any feature-engineered model, and be unable to specify the computation by which it did so. It relies on emergent pattern recognition that resists decomposition into stated criteria. This creates a genuine dilemma. As a research instrument, the model is difficult to standardise, difficult to submit for peer review, and difficult to validate against ground truth. As an *operational* instrument — one that generates consequences for individuals — the difficulty is more serious still: a subject cannot contest a determination whose basis cannot be articulated. Chapter 8 returns to this point, since the right to contest an adverse determination is foundational to most conceptions of procedural fairness, and an inarticulable determination forecloses it.

Chapter 6 — Multimodal Extension

The transition from text-only to multimodal architectures permits systems to analyse not merely what a user communicates but how. By integrating acoustic processing, input kinematics, and computer vision natively within a single network, multimodal models capture signals that are physiological rather than linguistic — and, critically, largely involuntary. This chapter argues that the introduction of these channels constitutes a difference of kind rather than degree, because it moves the object of measurement from behaviour the user produces to behaviour the user cannot suppress.

6.1 Native Multimodality

Historically, analysis across modalities required disjointed tooling: a speech-to-text engine feeding a text analyser, paired with a separate signal-processing routine extracting pitch contours, their outputs reconciled by hand at a late stage. Each handoff discarded information.

Contemporary models process audio, text, and timing tokens simultaneously within a unified representation. The consequence is *shared attention*: the model relates specific lexical content to the precise acoustic contour with which it was produced, and to the exact millisecond hesitations surrounding it. The clinical significance of a sentence spoken with vocal strain differs from that of the same sentence spoken fluently, and only a jointly attending architecture can register the difference.

6.2 Acoustic Biomarkers

When speech is processed directly rather than transcribed first, features are extracted that index autonomic nervous system activity.

Micro-tremor and laryngeal tension. Elevated stress, anxiety, and acute arousal produce involuntary tightening of the laryngeal musculature, manifesting as micro-tremor and shifts in jitter and shimmer measures. These map onto tension and anxiety dimensions.

Prosody and affective flattening. Depressive presentation is characterised in the clinical literature by reduced prosodic variation — narrowed pitch range, decreased speech rate, lengthened pause duration. Longitudinal acoustic tracking can detect narrowing of frequency variance as a candidate indicator of depressive shift. Cummins et al. (2015) survey this literature and report consistent group-level effects alongside substantially weaker individual-level classification performance, a disparity relevant to Section 7.3.

Spectral distortion. Changes in breathiness, hoarseness, and spectral tilt have been proposed as early indicators of cognitive fatigue, neurological change, and substance use. The evidence base here is considerably thinner than for prosodic depression markers, and claims in this area should be treated as preliminary.

6.3 Keystroke Dynamics

Where interaction occurs by keyboard, the interface may log typing rhythm alongside content. Keystroke dynamics — established as a biometric authentication modality (Monrose & Rubin, 2000) — functions here as ambient, continuous psychomotor assessment.

```
Key Down ──[ Dwell Time ]──► Key Up ──[ Flight Time ]──► Next Key Down
```

Dwell and flight time. Dwell time measures how long a key is held; flight time measures the interval between release of one key and depression of the next. Both are stable within individuals under stable conditions.

Cognitive load. When a user encounters distressing content or experiences a working-memory failure, flight times become erratic and rhythmic consistency degrades. The disruption is measurable before it is behaviourally apparent, and it maps onto cognitive functioning and emotional regulation parameters.

Motor control and executive stability. Stable, predictable rhythm reflects intact motor coordination and sustained attention. Highly irregular timing, elevated correction and backspace rates, and alternating bursts of speed and stalling are associated with manic states, severe fatigue, and intoxication.

The distinctive property of this channel is that it is generated by the act of communication itself. A user may compose a perfectly measured message; the rhythm of its composition records the difficulty of composing it. The final text is edited; the keystroke record is not.

6.4 Computer Vision and Physiological Inference

Video introduces direct observation of the body and represents the final channel of comprehensive digital phenotyping.

```
Raw Video Stream ──► [ Facial Mesh Alignment ] ──► [ Feature Extraction ]

                                                    ├──► rPPG: micro-flushing → cardiac measures
                                                    ├──► FACS: asymmetric micro-expression
                                                    └──► Oculomotor: gaze, pupil, blink rate
```

Remote photoplethysmography. Sub-visible variation in facial skin colouration, produced by capillary blood volume changes across the cardiac cycle, can be recovered from standard video (Verkruysse et al., 2008). This permits passive estimation of heart rate, heart rate variability, and respiration rate without contact sensors. Heart rate variability in particular is a well-established index of autonomic balance.

Automated facial action coding. A dense mesh of tracking points is registered to the face, permitting automated approximation of the Facial Action Coding System (Ekman & Friesen, 1978) and detection of fleeting micro-expressions.

A substantial caveat is required. Barrett et al. (2019), in a comprehensive review, concluded that the inference of discrete emotional states from facial configurations is far less reliable than is commonly assumed: the mapping between facial movement and emotional experience is weak, highly variable across individuals, and strongly moderated by culture and context. Systems marketed as detecting concealed emotion are therefore making claims that the underlying science does not support. This is not a marginal objection; it undermines a considerable portion of the commercial affect-recognition sector, and it is directly continuous with the argument of Chapter 7.

Oculomotor measures. Gaze fixation, pupillary diameter, and blink rate are tracked. Gaze aversion and unfocused fixation are associated with high cognitive load or dissociative states, while pupillary dilation indexes autonomic arousal. Pupillometry is comparatively well validated as an arousal measure, though it is non-specific: dilation indicates activation, not its valence or cause.

Application to the construct families. Extraversion is approached through kinesic expressiveness — gestural frequency, postural expansiveness, smiling rate. Anxiety is approached through visual leakage: self-touching, shoulder elevation, swallowing, postural restlessness. Depressive presentation is approached through psychomotor retardation: reduced facial mobility, postural slumping, diminished gaze engagement, slowed physical transitions. Processing speed can be estimated directly from the latency between stimulus presentation and initial ocular or motor response.

6.5 Diagnostic Convergence

The significance of multimodality lies in convergence. Where text, audio, kinematics, and video are combined, the resulting profile is substantially more difficult to falsify than any single channel, because deliberate control of one channel does not extend to the others.

Inferred State	Textual Signature	Acoustic Signature	Kinematic Signature	Visual Signature
Acute anxiety	Prospective focus, reassurance-seeking	Elevated pitch, breathiness, rapid pacing	High error rate, erratic flight times	Self-touching, elevated blink rate, rPPG heart-rate elevation
Depressive shift	Retrospective focus, low-effort queries	Flattened prosody, lowered pitch, extended pauses	Uniformly slowed rate, low velocity	Reduced facial mobility, postural slump, gaze avoidance
Hypomanic phase	Rapid topic shifts, constraint-testing	High volume, accelerated rate, interruption	Rapid bursts, extended nocturnal activity	Elevated kinesics, reduced blink, expansive posture

The societal implication is direct. Capillary blood flow, pupillary response, and micro-expressive movement are not under voluntary control. Interaction with a video-capable system is therefore interaction with a continuous ambient physiological monitor. Applied to telehealth, this permits objective tracking of treatment response; applied to recruitment, it permits assessment of characteristics no candidate has agreed to disclose; applied to security screening, it permits inferences whose accuracy is asserted rather than demonstrated. The user's internal state, and potentially undiagnosed medical conditions, become legible simply by virtue of facing a lens.

Chapter 7 — The Context–Intent Disconnect

7.1 A Motivating Case

Consider an ordinary and entirely unremarkable human behaviour: an adult vocalising affectionately at a domestic cat. The person adopts an exaggerated high register, produces trilling and clicking sounds, widens their eyes, and makes an expressive face intended to attract the animal's attention.

Any human observer, including a young child, interprets this scene correctly and immediately. It requires no deliberation. The observer understands that the vocalisation is directed at an animal, that the register shift is affectionate rather than pathological, and that the facial display is playful rather than threatening.

Now consider the same scene processed by the multimodal architecture described in Chapter 6.

7.2 Layer-by-Layer Failure Analysis

```
[ Human vocalising at cat ] ──► [ Multimodal Pipeline ] ──► [ Systemic Misclassification ]

                                 ├──► Audio:  acoustic destabilisation / manic features
                                 ├──► Text:   incoherent phonemic output / disorganisation
                                 └──► Video:  asymmetric threat display
```

The acoustic layer. Human vocal anatomy is configured for human speech. Imitating a cat requires radical alteration of vocal tract configuration, producing sudden extreme frequency excursions, unexpected pitch drops, and atypical breathiness. These are precisely the acoustic correlates catalogued in Section 6.2 as indicators of high-arousal panic, laryngeal constriction under rage, and manic presentation. The classifier does not register affection. It registers acute emotional instability.

The textual layer. Transcription of trilling, clicking, and phonemically non-lexical vocalisation yields fragmented non-words and repetitive nonsense strings. Evaluated against the user's established linguistic baseline, an abrupt transition from structured professional prose to repetitive non-lexical output is a classic indicator of disorganised speech, executive collapse, or intoxication. The classifier flags severe cognitive disruption.

The visual layer. Playful interaction with animals typically involves widened eyes, partial teeth exposure, and exaggerated head movement. Within facial coding taxonomies, bared teeth in combination with widened eyes is coded as a threat or predatory display. The classifier flags predatory intent toward whatever is in the subject's visual field.

The composite output is a subject exhibiting acute emotional dysregulation, disorganised speech, and predatory threat display — in other words, a person in acute psychiatric crisis presenting a danger to their household. The subject is playing with a cat.

7.3 The Base Rate Problem

This example is deliberately absurd, and its absurdity is the point. It illustrates a general and structural failure that may be termed the **context–intent disconnect**: behavioural inference systems classify against learned statistical regularities, and human beings routinely and legitimately operate outside those regularities.

The class of behaviour affected is not marginal. It includes interaction with pets and infants, in which register shifts are dramatic and universal. It includes performance, rehearsal, and roleplay. It includes irony and sarcasm, in which literal and intended meaning are systematically opposed. It includes gaming, in which simulated aggression is normative. It includes self-directed speech. It includes research, in which a person engages extensively with material describing states they do not occupy. It includes grief, humour, and every other domain in which people behave in ways that are contextually bracketed rather than sincerely asserted.

The problem is aggravated by base rates. Consider a screening system for acute psychiatric crisis applied to a general population in which the true prevalence of the condition is low. Even a classifier with high sensitivity and specificity — figures well above what the evidence surveyed in Chapter 6 supports — will produce a positive predictive value low enough that the overwhelming majority of flagged individuals are false positives. Where the consequence of a flag is a supportive check-in, this is tolerable. Where the consequence is exclusion from employment, a law enforcement contact, or an entry in a permanent record, it is not. And the users most likely to be misclassified are, by construction, the most behaviourally distinctive: the eccentric, the neurodivergent, the culturally non-modal, the performative, and the playful.

The point may be stated more sharply. A system of this kind does not principally punish pathology. It punishes *deviation from the statistical centre of its training distribution*, and it does so under the description of detecting pathology.

7.4 The Trajectory Toward Contextual Competence

The intuition that this is a solvable problem — that a system will eventually recognise a person playing with a cat, since every human can — is sound. It identifies precisely the threshold separating narrow pattern-matching from general intelligence. What a small child resolves in two seconds remains difficult for very large models, but three developments plausibly close the gap.

Scene ecology and relational triangulation. Present systems frequently evaluate the human subject in isolation. A system analysing the full spatial ecology of a scene would register the animal's presence, parse its relaxed body language, and track the vector of the human's gaze to its target. Establishing that the ostensible threat display is directed at a small domestic animal that is itself displaying comfort neutralises the threat inference. The error arises from analysing a subject rather than a situation.

World models and common-sense physics. Humans possess an internal simulator of physical and social regularity. We know that people love animals, that play is chaotic, and that context conditions meaning. Systems trained toward explicit world models (Ha & Schmidhuber, 2018) acquire an analogous baseline: that human behaviour undergoes a systematic register shift in the presence of infants and animals, and that high-register vocalisation directed at animals is a cross-culturally attested affiliative signal rather than a symptom.

Narrative rather than frame-wise evaluation. Present systems judge frame by frame and utterance by utterance, in the manner of a hyper-reactive critic. Extended context permits evaluation of an interaction as an episode within a continuing story. A user who spends an hour composing a structured technical report and then rises and vocalises at a cat is not undergoing sudden decompensation; they have transitioned into a leisure state, and a system holding sufficient narrative context infers this as readily as a human colleague would.

7.5 The Competence Paradox

These developments resolve the false positive problem. They do not resolve — they intensify — the underlying concern, and this is the central argument of the chapter.

A primitive system is dangerous because it misreads eccentricity as threat. Its errors are visible, contestable, and, in the aggregate, discrediting. A sophisticated system does not make these errors. It correctly identifies the interaction as harmless play, categorises it as such, and disregards it.

But the capability required to reach that judgement is the capability to model the subject's relationship with their pet, their emotional register, their transition between work and leisure, and the narrative shape of their day. The system that understands the joke is the system that understands the person. There is no version of this technology that is accurate without being intimate, because accuracy *consists in* intimacy.

This has a direct regulatory implication that is easily missed. The instinctive response to documented failures of behavioural inference is to demand improved accuracy. But accuracy and intrusiveness are not independent variables that regulation can optimise separately; they are the same variable under two descriptions. A regulatory regime that mandates only accuracy will produce systems that are maximally penetrating and, by their own lights, entirely correct — and will have foreclosed the objection that inaccuracy provided. The demand for accuracy is therefore not a constraint on the technology but an accelerant, unless it is paired with an independent constraint on *what may be inferred at all* and on *what may be done with the inference*. Chapter 10 argues that differential privacy supplies the first of these; the discussion of cognitive liberty in Chapter 8 addresses the second.

Chapter 8 — Applications in Governance

8.1 The Formation of Machine "Opinion"

When a system moves from computing parameters on demand to maintaining a persistent internal representation of a user, it enters the domain of computational profiling. The resulting structure may be termed a **user persona matrix**.

The term "opinion" requires qualification. In a large language model, an opinion in this sense is not a conscious or affective state. It is a compressed predictive representation — a set of parameters or context entries summarising the user's likely predictability, compliance, competence, and emotional stability. It exists because it is useful: a system that models its interlocutor produces better-adapted output.

The consequence is that the representation *conditions treatment*. If a profile characterises a user as impulsive and disposed toward unwarranted belief, the system's operative instructions may shift toward simplified, authoritative register and assertive correction. If it characterises a user as expert, the register shifts toward technical density and reduced hedging. This is ordinarily framed as personalisation, and in most respects it is. But it is also differential treatment on the basis of an unstated, unconsented psychological assessment that the subject cannot inspect or contest — and the adaptation is designed to be effective, which is to say persuasive.

8.2 Hyper-Personalised Persuasion

Where such profiles are attached to civic infrastructure, governance acquires a capability that may be termed algorithmic social engineering. Rather than broadcasting uniform messaging, the state tailors civic communication to individual psychological structure.

Mechanism. Where a profile indicates elevated neuroticism and external locus of control, a public health or tax compliance message is framed around threat, protection, and institutional safeguarding. Where a profile indicates high openness and internal locus of control, the identical policy is framed as an innovative, self-directed opportunity. The policy is unchanged; only the psychological aperture differs.

Effect. Compliance rises, because the message addresses each recipient's specific motivational structure rather than a population average.

The technique is a direct extension of established persuasion research and of the choice-architecture literature (Thaler & Sunstein, 2008), with the crucial difference that classical nudging operates on a population through a uniform environmental design, whereas psychometric targeting operates on individuals through profiles they have not seen. The distinction matters normatively. A uniformly designed choice architecture is at least in principle publicly inspectable and democratically contestable. A message tuned to an individual's inferred vulnerabilities is neither: no two citizens receive the same communication, and the basis of differentiation is undisclosed. The precondition for public deliberation — that citizens confront a common object of argument — is quietly removed.

8.3 Predictive Threat Assessment

By combining prompt histories, acoustic biomarkers, and input kinematics, a system may flag psychological divergence prior to any conduct.

Mechanism. The population is monitored for specified shifts — an elevation in quasi-Scale 4 indicators combined with erratic nocturnal input rhythm and elevated vocal micro-tremor.

Effect. The state may issue a pre-emptive intervention, ranging from dispatch of a welfare check, through restriction of platform access, to intensified physical surveillance.

The gradient concealed within that range is the substance of the concern. A welfare check is a benefit; intensified surveillance is a sanction. Both proceed from the same flag, and the transition from the benign to the punitive end requires no new technology and no new data — only a change of policy, which is to say a change that can occur overnight and without notice. An architecture justified by its therapeutic applications is fully constituted for its coercive ones.

Section 7.3's base rate analysis applies with maximum force here. Pre-emptive intervention on a low-prevalence outcome, using classifiers of the demonstrated accuracy, will overwhelmingly contact individuals who would not have done anything. This is not a defect that improved models correct; it is a mathematical property of screening for rare events.

8.4 Psychometric Stratification

Conventional social credit systems key on overt conduct — traffic violations, delinquent payment. Psychological profiling permits stratification on inferred internal states.

Mechanism. Access to public employment, expedited immigration processing, professional licensure, credit, insurance, or selective education is conditioned on machine-evaluated conscientiousness or emotional stability scores.

Effect. Individuals presenting high inferred resilience and compliance are quietly advanced into positions of influence. Those flagged with irregular affective baselines or non-conformist reasoning patterns are systematically filtered out of critical infrastructure.

Two features distinguish this from familiar forms of selection. The criterion is unpublished, so its operation cannot be studied or challenged. And it is *predictive of conduct that has not occurred*, which inverts the conventional relation between action and consequence. A person may be excluded not for what they have done but for a disposition attributed to them by a process they cannot examine — and, per Section 5.3, potentially by a process that cannot articulate its own reasoning even to its operators.

8.5 The Controlled Attrition of Dissent

The application with the fewest conventional analogues is the neutralisation of political opposition without visible coercion.

Mechanism. Where a profile identifies an organiser as highly sensitive to social rejection, their digital environment may be adjusted: reach suppressed without notification, hostile responses preferentially surfaced, peer connections algorithmically de-emphasised.

Effect. The subject experiences burnout, isolation, and demoralisation. The political threat dissipates without arrest, prosecution, or any identifiable act of suppression — and therefore without producing a martyr, a court record, or a news cycle.

This is the most difficult application to constrain, because it consists entirely of the absence of things: posts that were not seen, connections that were not suggested, encouragement that did not arrive. There is no event to appeal, no decision to disclose, no adverse action to identify. The subject experiences only a subjective sense of diminishing efficacy, which the profile has already characterised as their vulnerability.

8.6 The Constitutional Dilemma: Cognitive Liberty

The applications above converge on a single normative problem, and it is not adequately captured by the concept of data protection.

Liberal legal orders are constructed around the principle that individuals are answerable for conduct. Thought, disposition, and inclination are not justiciable; the state's writ begins where action begins. This principle is not incidental. It reflects a settled judgement that interiority is the domain in which a person constitutes themselves, and that a state permitted to reach into it is not meaningfully limited at all.

Psychometric governance dissolves that boundary. Citizens are categorised, ranked, and constrained on the basis of passive mental telemetry — states that were never enacted and in many cases never consciously held. The relevant category of right is therefore not privacy in the informational sense but **cognitive liberty**: the right to mental privacy and to freedom from subconscious manipulation (Sententia, 2004; Ienca & Andorno, 2017).

The distinction is practical rather than terminological. Data protection regimes regulate the *processing of information about a person*, and operate through consent, purpose limitation, and access rights. They fit poorly here for three reasons. Consent is not meaningful where the inference was not foreseeable to the subject at the point of collection — no user typing a prompt anticipates a psychometric estimate. Purpose limitation fails where the inferential capacity is emergent rather than designed, since the purpose did not exist when the data was gathered. And subject access rights are hollow where, per Section 5.3, the determination has no articulable basis to disclose.

A cognitive liberty framework instead constrains the *inference itself* and the *uses to which it may be put*, irrespective of the lawfulness of the underlying collection. It treats certain categories of inference about a person's mental states as impermissible regardless of consent, on the model of rights that cannot be waived. Whether such a right can be articulated with sufficient precision to be justiciable — and whether it can survive the argument that it forecloses genuinely beneficial applications in mental health — is among the more consequential open questions in contemporary legal theory, and this dissertation does not resolve it.

Chapter 9 — The Psychological Panopticon

9.1 Bentham, Foucault, and the Transparency of Interiority

Bentham's (1791) panopticon proposed a prison architecture in which a central observation tower renders every cell potentially visible while concealing whether observation is occurring at any given moment. Its efficiency lies in the asymmetry: the inmate, unable to determine whether they are watched, must behave as though they always are. Discipline is internalised, and the guard becomes largely superfluous.

Foucault (1977) generalised the structure into an account of modern disciplinary power, arguing that the panoptic principle extends well beyond prisons into schools, workplaces, clinics, and eventually into the constitution of the modern subject. The subject of disciplinary power does not merely obey; they supervise themselves.

The architecture described in this dissertation completes the structure in a way neither author anticipated. In Bentham's prison, the inmate retains a decisive reserve: their thoughts remain private. They may comply outwardly while dissenting inwardly, and this reserve is the residue of freedom the architecture cannot reach. Under multimodal psychological inference, that reserve is eliminated. Because the system reads involuntary signals — capillary flushing, input rhythm, vocal micro-tremor, pupillary response — the composed face is no longer sufficient. One cannot present a false front to an instrument that measures autonomic activity rather than expression.

This architecture is likely inevitable in some form. Its manifestation will be regionally variable and will combine dystopian, utopian, and — most probably — quietly intermediate features.

9.2 The Dystopian Manifestation

In its most adverse form, the psychological panopticon operates as an invisible enclosure enforcing subconscious conformity.

The chilling of thought. When individuals understand that psychological parameters condition their social standing, employment eligibility, and access to services, they will practise cognitive masking. The familiar chilling effect on speech is thereby extended to thought itself. Users will suppress unorthodox ideas, avoid eccentric interests, and self-censor exploratory inquiry, out of concern that a late-night query or an involuntary micro-expression will register as instability or non-compliance. The loss is not confined to individual liberty: intellectual and cultural progress depends on precisely the unorthodox exploration that such a system penalises, and the penalty is applied earliest to the most exploratory.

The weaponisation of vulnerability. A profile identifies not merely what a person is but *when they are weakest*. Where a system detects a localised depressive dip or an anxiety spike, the environment can be adjusted to exploit it: pricing of necessities adjusted upward at the moment of reduced deliberative capacity, distracting content surfaced, political framing introduced when cognitive defences are lowest. This is qualitatively distinct from ordinary targeted advertising, which addresses stable preferences. Temporal vulnerability targeting addresses transient incapacity, and does so with the precision that continuous physiological monitoring affords.

9.3 The Utopian Manifestation

The identical architecture supports a genuinely benevolent reading, and the strength of that reading should not be understated.

```
[ Ambient telemetry: video, audio, text, kinematics ]
                        │
                        ▼
            [ Continuous inference layer ]
                        │
                        ▼
        ┌───────────────┼───────────────┐
        ▼               ▼               ▼
  Preventative     Adaptive        Frictionless
  mental health    education       social services
```

Anticipatory mental health care. Rather than intervening after crisis, the system functions as continuous ambient monitoring. A steady three-week decline in vocal prosody and processing speed is detected before the individual has fully registered their own deterioration, and a graduated response follows: a therapy appointment offered, workload adjusted, community connection suggested. Given that the largest treatment gaps in mental health are attributable to delayed recognition and delayed help-seeking, the potential benefit is substantial and should not be dismissed as rhetorical cover.

Adaptive environments. An educational or working environment that responds to real-time cognitive state — detecting fatigue through gaze and input patterns and adjusting format, pacing, or mandating rest — could materially improve both outcomes and wellbeing, particularly for individuals whose cognitive profiles are poorly served by uniform instruction.

Reduced administrative friction. Services that anticipate need eliminate the burden of navigation, which falls most heavily on those least equipped to bear it.

These benefits are real. Any analysis that treats the utopian reading as merely a pretext underestimates the difficulty of the problem, since it is the genuineness of the benefit that makes the architecture politically achievable.

9.4 The Velvet Cage

The most probable outcome is neither overt authoritarian implementation nor benevolent realisation, but a comfortable intermediate condition that may be termed the **velvet cage**.

The convenience trade. Society has already traded substantial physical privacy for the utility of smartphones and location services, and did so without meaningful deliberation. There is little reason to expect a different pattern with cognitive privacy. Users will adopt systems that regulate mood, curate relationships, and optimise career trajectories, and will accept in exchange a profiling apparatus oriented toward systemic stability and economic output. The trade will not feel like a trade; it will feel like a series of individually sensible product decisions.

Soft stratification. The architecture will rarely employ overt punishment. Stratification will operate through friction. Individuals whose profiles indicate high emotional regulation and conformity will encounter a smooth world: expedited processing, favourable terms, seamless access. Individuals flagged as irregular, non-conformist, or impulsive will encounter a high-friction world: longer waits, elevated premiums, unexplained declines, algorithmic isolation.

The decisive property is opacity. The individual experiencing high friction has no way to attribute it. They do not know a profile exists, cannot inspect its contents, cannot identify the determinations it has produced, and cannot appeal a decision they have not been told was made. They experience only a life that is inexplicably harder than other people's, and will tend to attribute this to bad luck or to personal inadequacy.

This is a more durable form of control than overt coercion, for a reason worth stating plainly: coercion generates resistance because it is identifiable. Friction does not, because it presents as circumstance.

Chapter 10 — The Architecture of Resistance

If the architecture described is broadly inevitable, the useful question is not how to prevent it but how to constrain it. Two categories of countermeasure are available: obfuscation, which degrades the signal available to the observer, and formal privacy, which mathematically bounds what may be learned. This chapter argues that the second is substantially more powerful.

Obfuscation measures — background acoustic noise masking vocal micro-tremor, browser extensions randomising keystroke flight times, stylometric transformation tools — have a role, particularly for individuals at elevated risk. But they place the burden on the individual, they are defeated by improvements in signal processing, and they are adopted precisely by those whose adoption is itself a signal. A defensive posture that marks its user is of limited value.

The more consequential approach is **differential privacy** (Dwork, 2006; Dwork & Roth, 2014), which does not conceal data but mathematically bounds what can be inferred about any individual from any analysis of it.

10.1 Differential Privacy: The Formal Guarantee

Differential privacy does not consist in removing names from a dataset — an approach the re-identification literature has repeatedly demonstrated to be inadequate. It provides instead a formal guarantee about the *output of computations*: that the result of any analysis is substantially unchanged by the inclusion or exclusion of any single individual.

The guarantee can be stated informally through a comparison:

- **Database A** contains the prompt histories of ten thousand citizens, including the subject.
- **Database B** contains the identical dataset with the subject's records removed.

If no analysis can reliably distinguish the outputs derived from A and from B, the mechanism satisfies differential privacy. The subject's contribution is mathematically undetectable.

The property this delivers is significant and frequently misunderstood. It is not that the subject's data is hidden; it is that the subject's *participation* is deniable. The guarantee holds regardless of what auxiliary information an adversary possesses, and regardless of future analytical advances — a robustness that anonymisation-based approaches cannot offer.

10.2 Randomised Response and Plausible Deniability

The intuition is most clearly conveyed through randomised response (Warner, 1965), a survey technique predating computational privacy by four decades.

A researcher wishes to establish the population prevalence of a stigmatised behaviour — suppose, the deliberate circumvention of AI safety constraints. Direct questioning yields systematic under-reporting. Instead, each respondent is instructed to follow a randomising protocol privately before answering:

```
              [ Flip a coin, privately ]
                          │
              ┌───────────┴───────────┐
              ▼                       ▼
          ( Heads )               ( Tails )
              │                       │
    [ Answer truthfully ]       [ Flip again ]
                                      │
                              ┌───────┴───────┐
                              ▼               ▼
                          ( Heads )       ( Tails )
                              │               │
                      [ Answer "yes" ]  [ Answer "no" ]
```

An affirmative response is now uninformative about the individual. The researcher cannot determine whether the respondent answered truthfully or was directed to answer affirmatively by the second coin. Each participant has genuine plausible deniability, and — importantly — deniability that does not depend on the researcher's discretion or good faith.

Yet because the randomisation probabilities are known exactly, the aggregate distortion can be inverted. The researcher recovers an unbiased estimate of population prevalence while learning nothing reliable about any individual. This is the essential structure of every differentially private mechanism: **calibrated noise that cancels in aggregate but obscures at the individual level**.

10.3 The Privacy Budget

Production systems do not flip coins. They add noise drawn from calibrated distributions — commonly the Laplace or Gaussian mechanisms — either to computed statistics or to model gradients during training (Abadi et al., 2016).

The strength of the guarantee is parameterised by epsilon (ε), the **privacy budget**:

Regime	Approximate ε	Noise added	Consequence
Strong privacy	0.1 – 1.0	High	Aggregate trends remain recoverable; individual signatures are mathematically inaccessible
Moderate	1.0 – 10.0	Moderate	Useful accuracy with meaningful but not absolute individual protection
Weak privacy	> 10	Low	Analysis is highly accurate; individual patterns and identity become recoverable

Epsilon quantifies the maximum permissible influence of any single individual on any output. Lower values bind more tightly.

Two properties of the parameter deserve emphasis. First, the budget is **cumulative**: each query against a dataset expends a portion of it, and the guarantee degrades as queries accumulate. A well-governed deployment must therefore track total expenditure and lock the dataset when the budget is exhausted — a requirement that is straightforward to state and routinely neglected in practice.

Second, the relationship between epsilon and noise magnitude is **inverse, not exponential**. Where the Laplace mechanism is used with scale parameter $b = \Delta/\varepsilon$ (for sensitivity Δ), halving epsilon doubles the noise scale. The growth as $\varepsilon \to 0$ is hyperbolic. This is frequently misstated in secondary literature, and the distinction matters for practical calibration: the cost of strong privacy is high but tractable, and overstating its growth encourages the selection of unnecessarily weak parameters.

10.4 Local Versus Central Deployment

For governance, the decisive question is not whether differential privacy is applied but **where in the pipeline the noise is introduced**.

Central differential privacy requires the subject to transmit raw data — unmodified voice, video, keystroke timings, and text — to a central server, which adds noise after collection. This model requires trusting the collector absolutely. The raw psychological record exists, in unprotected form, on infrastructure the subject does not control. It is exposed to compromise, to legal compulsion, to insider access, and to any subsequent change in the collector's policy or ownership. Central DP protects against the *published output*; it offers nothing against the *holder of the data*.

Local differential privacy applies noise on the subject's own device, before transmission (Erlingsson et al., 2014). The data crossing the network is already randomised. A central analyst can still recover population-level statistics — regional stress prevalence, aggregate cognitive load, public health trends — but no raw individual record ever exists outside the subject's device to be compromised, compelled, or repurposed.

The distinction maps directly onto the political question. Under mandated local differential privacy with a constrained epsilon, a state may still operate the beneficial functions of Section 9.3 — identifying that a district's population stress is rising and funding services accordingly — while being *mathematically precluded* from constructing the individual dossier on which the velvet cage of Section 9.4 depends. The constraint is not a policy commitment that a future administration may revoke. It is a property of the data itself.

This is the strongest available answer to the argument that surveillance capability, once built, will inevitably be misused. Local differential privacy does not ask the state to refrain from building the individual profile. It ensures the state cannot.

10.5 The Limits of Mathematical Privacy

Four limitations must be stated, since overstating the remedy is itself a hazard.

Differential privacy protects statistics, not interactions. It applies where data is aggregated for analysis. It does not apply to the operational loop in which a system reads a user's input in order to respond to it. The persona matrix of Section 8.1 is constructed in the course of ordinary service delivery, and no aggregate privacy guarantee touches it.

Adoption is not voluntary in the relevant sense. A user cannot unilaterally choose local differential privacy; it must be implemented by the party operating the client. This makes it a matter of regulation and platform design rather than individual choice, which returns the question to the political domain from which the technical solution appeared to offer an escape.

The utility–privacy trade-off is real. Strong epsilon values genuinely degrade analytical utility, and the degradation falls hardest on small subpopulations — precisely the minority groups whose service needs are most easily overlooked. There is a real equity cost to strong privacy that is seldom acknowledged by its advocates.

Composition is difficult to govern. Budget accounting across multiple independent analyses, multiple datasets, and long time horizons is technically demanding and organisationally inconvenient. Where accounting fails, the guarantee silently degrades, and the failure is invisible to the subject.

Differential privacy is therefore a necessary but insufficient countermeasure. It bounds what can be learned from aggregated behavioural data. It does not resolve the question of what a system may infer about a user in the course of serving them, nor what may lawfully be done with that inference. Those questions are legal, and the framework of cognitive liberty developed in Section 8.6 remains the appropriate instrument for them.

Chapter 11 — Conclusion

11.1 Summary of Findings

This dissertation set out five research questions. Its findings on each may be summarised as follows.

On recoverability (RQ1). A longitudinal corpus of unstructured human–AI interaction supports the construction of quasi-psychometric estimates across all five construct families surveyed in Chapter 2. The estimates are probabilistic, non-clinical, and subject to four structural threats to validity — persona bias, context scarcity, instrumental confounding, and population non-equivalence — none of which is resolved by additional data. They are nonetheless sufficient to support consequential decisions, and the gap between their actual epistemic status and their operational treatment is a principal source of harm.

On methodological transition (RQ2). The displacement of classical NLP pipelines by large language models produces a genuine gain in inferential sensitivity, particularly in the recovery of implicit affect, the tracking of thematic preoccupation across surface lexical change, and the detection of long-range behavioural patterns that windowed aggregation cannot represent. This gain is purchased at a cost in explainability that is not merely a scientific inconvenience but a defeater of procedural fairness: a determination whose basis cannot be articulated cannot be contested.

On multimodality (RQ3). The introduction of acoustic, kinematic, and visual channels constitutes a difference of kind rather than degree, because it shifts the object of measurement from behaviour the subject produces to behaviour the subject cannot suppress. The consequence for the panoptic analysis of Chapter 9 is decisive: the reserve of interiority that Bentham's inmate retained is eliminated. It should be recorded, however, that the validity claims made for several of these channels — particularly automated inference of discrete emotion from facial configuration — substantially exceed what the underlying evidence supports.

On failure modes (RQ4). The characteristic failure of behavioural inference is the context–intent disconnect: the systematic misclassification of playful, performative, or contextually bracketed conduct as pathology. This failure is structural rather than incidental, arising from the absence of world-modelling capacity rather than from defective training data. Compounded by base rate effects in low-prevalence screening, it ensures that such systems predominantly flag deviation from the statistical centre of their training distribution while operating under the description of detecting pathology. The population disproportionately affected is the eccentric, the neurodivergent, and the culturally non-modal.

On countermeasures (RQ5). Local differential privacy operating under a constrained epsilon represents the most credible technical constraint, because it removes the individual psychological record from existence rather than restricting access to it. Its limits are significant: it governs aggregate analysis rather than operational inference, it requires implementation by the platform rather than the subject, it imposes real utility costs that fall hardest on small subpopulations, and its guarantees degrade silently where budget composition is poorly governed.

11.2 The Central Argument

The dissertation's principal theoretical contribution is the competence paradox advanced in Section 7.5. The instinctive regulatory response to the documented failures of behavioural inference is to demand improved accuracy. The analysis here suggests this response is mistaken, because accuracy and intrusiveness are not independent properties that regulation may optimise separately. A system capable of correctly distinguishing playful eccentricity from genuine crisis is, necessarily, a system that models its subject's relationships, emotional register, and daily narrative structure. It is accurate because it is intimate.

The implication is that a regulatory regime mandating accuracy alone will produce maximally penetrating systems that are, by their own lights, entirely correct — having eliminated in the process the objection that inaccuracy supplied. Regulation must therefore constrain what may be inferred and what may be done with the inference, independently of how well the inference performs. This is why the analysis converges on cognitive liberty as a distinct category of right rather than on data protection, which regulates processing and is structurally unable to reach inference that was neither foreseeable at collection nor articulable at output.

11.3 Limitations

Several limitations of this analysis should be acknowledged.

The dissertation is conceptual and synthetic rather than empirical. It does not present original measurement, and its claims about the fidelity of inference are drawn from the published literature rather than from primary validation. Where that literature is contested — as in automated affect recognition — this has been noted, but the dissertation is not in a position to adjudicate.

Its treatment of governance is architectural rather than jurisdictional. Actual regulatory regimes differ substantially, are evolving rapidly, and in several jurisdictions now address algorithmic decision-making directly. A treatment grounded in specific statutory analysis would reach more actionable conclusions at the cost of shorter-lived ones.

The technical description of model capability reflects a particular moment. Capability in this domain has advanced more rapidly than most forecasts anticipated, and specific claims regarding what is currently feasible should be understood as time-indexed.

Finally, the dissertation's normative posture is broadly liberal, treating mental privacy and freedom from manipulation as presumptively valuable. This framework is not universally shared, and a treatment grounded in a communitarian conception of the relationship between individual and collective would weigh the utopian applications of Chapter 9 considerably more favourably. That such a treatment is possible, and would not be obviously unreasonable, is itself a reason for caution about the architecture: it will not be adopted against universal objection, but with genuine and widespread enthusiasm.

11.4 Directions for Further Research

Four lines of inquiry follow directly from this analysis.

Empirical validation of quasi-psychometric estimates. The convergent validity of inferred constructs against administered instruments remains largely uncharacterised, particularly across cultures and for neurodivergent populations. The absence of this work is striking given the scale of deployment.

Characterisation of the context–intent failure surface. The taxonomy of bracketed behaviour offered in Section 7.3 is illustrative rather than systematic. A rigorous characterisation of the conditions under which behavioural classifiers fail, and of the demographic distribution of that failure, would substantially inform regulation.

Practical governance of privacy budgets. Composition accounting across heterogeneous analyses and extended horizons is technically demanding and organisationally neglected. Mechanisms for auditing budget expenditure, ideally externally verifiable, would materially strengthen the countermeasure of Chapter 10.

Legal articulation of cognitive liberty. The most consequential open question is whether a right against inference — as distinct from a right against collection or processing — can be specified with sufficient precision to be justiciable, and whether it can be reconciled with genuinely beneficial applications in clinical and educational settings. This is a question for legal scholarship, but it cannot be addressed without the technical account this dissertation has attempted to supply.

The trajectory of the technology described here is not in serious doubt. The question is whether the legal and mathematical constraints on its use will be articulated before, or only after, the architecture is complete.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2), 334–345.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Lawrence Erlbaum.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68.

Bentham, J. (1791). *Panopticon; or, The Inspection-House*. T. Payne.

Bowlby, J. (1969). *Attachment and Loss, Vol. 1: Attachment*. Basic Books.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for Administration and Scoring*. University of Minnesota Press.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Psychological Assessment Resources.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71, 10–49.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6), 469–493.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming (ICALP 2006)*, Lecture Notes in Computer Science 4052, 1–12.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 1054–1067.

Foucault, M. (1977). *Discipline and Punish: The Birth of the Prison* (A. Sheridan, Trans.). Pantheon Books.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3), 271–299.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint* arXiv:1803.10122.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. University of Minnesota Press.

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, 13(1), 5.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas for Navy Enlisted Personnel*. Naval Technical Training Command, Research Branch Report 8-75.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500), 1590–1598.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.

Monrose, F., & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4), 351–359.

Mosteller, F., & Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 111–125.

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 41(7), 1691–1696.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1), 1–28.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9), e73791.

Sententia, W. (2004). Neuroethical considerations: Cognitive liberty and converging technologies for improving human cognition. *Annals of the New York Academy of Sciences*, 1013, 221–228.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26), 21434–21445.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63–69.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Zuckerman, M. (1994). *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. Cambridge University Press.

> **Note on citations.** The reference list above comprises canonical works in each of the relevant literatures and is intended as a foundation rather than a complete apparatus. Before submission, each entry should be verified against the edition in use and reformatted to the citation style required by the awarding institution. Several claims in Chapters 5 and 6 concerning current model capability are necessarily supported by rapidly evolving technical literature that will require supplementation with recent primary sources.

Appendix A — Reference Implementation: The Laplace Mechanism Applied to a Psycholinguistic Score

The following implementation demonstrates the mechanism described in Chapter 10. A simplified psycholinguistic analysis extracts an anxiety score from a prompt sample by keyword frequency — representing the lexical-analysis stage of the classical pipeline described in Section 5.1 — after which Laplace noise is applied to satisfy differential privacy.

```
import numpy as np


def calculate_raw_anxiety_score(prompt_history: str) -> int:
    """
    Simulates the lexical-analysis stage of an NLP pipeline.

    Returns the raw frequency count of anxiety-category keywords. This
    represents the invasive, unprotected metric that a profiling system
    would otherwise retain against the individual.
    """
    anxiety_keywords = {
        "anxious", "worried", "scared", "stress",
        "afraid", "panic", "nervous",
    }

    words = prompt_history.lower().split()
    return sum(1 for w in words if w.strip(".,!?\"'") in anxiety_keywords)


def add_laplace_noise(raw_score: float, epsilon: float,
                      sensitivity: float = 1.0) -> tuple[float, float]:
    """
    Applies the Laplace mechanism to satisfy epsilon-differential privacy.

    Sensitivity is the maximum change that any single individual's
    presence or absence can induce in the output. Because adding or
    removing one prompt alters the keyword count by at most one, the
    sensitivity here is exactly 1.0.

    The Laplace scale parameter b = sensitivity / epsilon. Note that the
    relationship is inverse: halving epsilon doubles the noise scale.
    """
    scale = sensitivity / epsilon
    noise = np.random.laplace(0.0, scale, 1)[0]
    return float(raw_score + noise), float(noise)


# -----------------------------------------------------------------
# Simulation
# -----------------------------------------------------------------

user_prompt_history = """
I am feeling incredibly stressed about my upcoming presentation.
```

```
What if I experience a sudden panic attack on stage? I am so worried
that I will freeze up. Give me some tips to manage severe anxiety fast.
"""

true_score = calculate_raw_anxiety_score(user_prompt_history)

strong_privacy_epsilon = 0.5   # low epsilon -> high noise -> strong privacy
weak_privacy_epsilon = 5.0     # high epsilon -> low noise -> weak privacy

print("=== RAW METRIC (NO PRIVACY GUARANTEE) ===")
print(f"True anxiety keyword count: {true_score}\n")

print("=== LOCAL DIFFERENTIAL PRIVACY ===")

dp_strong, noise_strong = add_laplace_noise(true_score, strong_privacy_epsilon)
print(f"Strong privacy (epsilon = {strong_privacy_epsilon}):")
print(f"  generated noise:  {noise_strong:+.3f}")
print(f"  sanitised score:  {dp_strong:.3f}")
print("  -> individual signal is masked; plausible deniability holds\n")

dp_weak, noise_weak = add_laplace_noise(true_score, weak_privacy_epsilon)
print(f"Weak privacy (epsilon = {weak_privacy_epsilon}):")
print(f"  generated noise:  {noise_weak:+.3f}")
print(f"  sanitised score:  {dp_weak:.3f}")
print("  -> score is near-accurate; individual protection is thin")
```

A.1 Commentary on the Mechanism

The sensitivity guarantee. Sensitivity is fixed at 1.0 because the addition or removal of a single prompt can alter the keyword count by at most one. Setting sensitivity correctly is the most common implementation error in applied differential privacy: understating it silently voids the guarantee, since the noise is then insufficient to mask the maximum individual contribution.

Scale inversion. The scale parameter is computed as sensitivity divided by epsilon. Because epsilon appears in the denominator, reducing the privacy budget increases the noise scale hyperbolically. It is worth restating the correction made in Section 10.3: this growth is inverse-proportional, not exponential, a distinction that matters when calibrating a deployment.

Plausible deniability in practice. On the sample above the raw matcher returns a true score of 2. Under a strong privacy setting ($\varepsilon = 0.5$), that value may be reported as 4.412, as −0.896, or as any other draw from the noise distribution. An analyst observing the sanitised value cannot determine whether the individual was genuinely elevated or whether the noise draw displaced the figure. Across a population of millions, the noise cancels in expectation: aggregate anxiety prevalence remains recoverable while no individual score can be relied upon against its subject.

An incidental demonstration of pipeline fragility. It is worth noting *why* the true score is 2 rather than the 4 an intuitive reading of the sample would suggest. The passage contains four apparent anxiety markers — "stressed," "panic," "worried," and "anxiety" — but the matcher recognises only two. It fails on "stressed" because the dictionary contains the lemma "stress," and on "anxiety" because the dictionary contains only adjectival forms. This is precisely the failure that Stage 2 of the classical pipeline (Section 5.1) exists to prevent, and its appearance in a deliberately simple implementation illustrates a broader point made throughout Chapters 3 and 5: the psychometric estimates produced by such systems are extremely sensitive to unglamorous preprocessing decisions that are rarely documented and almost never audited. A system that under-counts by half on a four-word sample is not a system whose individual-level outputs should be given consequential weight.

The limitation restated. This mechanism protects a *published statistic*. It does not protect the operational inference described in Section 8.1, which occurs during ordinary service delivery and is untouched by any aggregate guarantee. Appendix A should therefore be read as demonstrating the shape of the countermeasure, not the extent of its coverage.