Systems and Methods for Predicting Vulnerability to Institutional Gaslighting: A Research Program Proposal

Joshua M. Garfunkel

Disrupt the Loop

December 2025

Working Paper - Research in Progress

ABSTRACT

Modern institutions increasingly deploy sophisticated psychological tactics that systematically harm vulnerable individuals while maintaining plausible deniability. This working paper outlines a comprehensive research program to address three critical gaps: (1) the absence of validated methods to predict who is vulnerable to institutional exploitation, (2) the lack of objective systems to detect manipulation tactics operating invisibly in real-time, and (3) the absence of metrics to quantify psychological harm that is routinely dismissed as "subjective." We propose an interdisciplinary research framework integrating clinical psychology, computer science, neuroscience, digital ethics, and public health to develop vulnerability prediction systems, manipulation detection algorithms, harm quantification metrics, and therapeutic interventions. This represents paradigm-shifting work with applications spanning clinical treatment, consumer protection, regulatory frameworks, and institutional accountability.

Keywords: institutional manipulation, vulnerability assessment, algorithmic accountability, digital dignity, treatment-resistant conditions, epistemic injustice

---

1. INTRODUCTION

1.1 The Problem

Individuals navigating modern institutional processes—healthcare systems, insurance companies, disability evaluations, government agencies—frequently report experiences of systematic psychological harm. They describe feeling gaslit, manipulated, trapped in circular bureaucratic processes, and

subjected to procedures that erode their sense of dignity and autonomy. Yet these experiences are routinely dismissed as "subjective complaints" or attributed to individual psychopathology rather than institutional design.

This dismissal occurs despite growing evidence that: - Treatment-resistant psychological conditions affect 40-70% of individuals seeking mental health care, suggesting systemic rather than individual failures - Power asymmetries enable institutions to deploy psychological tactics with impunity - Procedural injustice predicts worse health outcomes independent of final decisions - Emerging emotional AI systems may inadvertently amplify rather than reduce these harms

1.2 Critical Research Gaps

Three fundamental gaps prevent progress:

Gap 1: No validated vulnerability prediction. We cannot currently identify who is at risk for institutional exploitation before harm occurs. Existing models rely on fixed demographic categories (age, gender, race, socioeconomic status) that poorly predict actual vulnerability and reinforce harmful stereotypes.

Gap 2: No real-time manipulation detection. Covert manipulation tactics operate invisibly. Victims recognize harm only retrospectively, if at all. No systems exist to detect euphemistic language, responsibility displacement, manufactured urgency, or other documented psychological tactics in real-time.

Gap 3: No objective harm quantification. Psychological harm is dismissed as "how someone feels" rather than quantifiable injury. Without objective metrics linking institutional procedures to measurable outcomes, accountability remains impossible.

1.3 Proposed Solution Framework

This research program proposes novel theoretical frameworks requiring empirical validation:

1. **Vulnerability Assessment Architecture**: A composite index predicting susceptibility to institutional manipulation based on value authenticity, stability, life satisfaction trajectories, and personality flexibility—conceptualizing vulnerability as temporal and architectural rather than demographic and fixed.

2. **Manipulation Detection Systems**: Natural language processing algorithms to identify

covert tactics including moral disengagement mechanisms, emotional market maker strategies, and procedural grinding patterns.

3. **Digital Dignity Index**: A multiplicative harm model integrating subjective narrative violation with objective procedural burden metrics, transforming "subjective complaints" into quantified, legally defensible evidence.

4. **Treatment-Resistant Condition Architecture**: A unified pattern underlying conditions that fail standard interventions, suggesting institutional tactics may deliberately activate this vulnerability.

5. **Therapeutic Innovation**: Multi-component interventions based on biological heterosis principles, addressing the architectural pattern rather than surface symptoms.

---

2. THEORETICAL FOUNDATIONS

2.1 Vulnerability as Temporal Architecture

Current vulnerability models treat risk as fixed demographic traits. We propose vulnerability as a **temporal, state-dependent architecture** reflecting four dynamic factors:

- **Philosophical coherence**: The degree to which an individual's values reflect authentic reflection versus external imposition
- **Value stability**: Longitudinal consistency of values versus context-driven fluctuation
- **Life satisfaction trajectory**: Well-being changes temporally linked to institutional events
- **Personality flexibility**: Adaptive capacity existing on an inverse-U curve (both rigidity and excessive fluidity increase vulnerability)

This conceptualization draws on Frankfurt's theory of authenticity, Schwartz's values framework, the Mischel-Shoda CAPS model, and Heideggerian temporal coherence. It predicts vulnerability more accurately than demographics while avoiding stigmatization—anyone can become vulnerable under specific circumstances.

Research Question 1: Can a composite vulnerability index predict institutional exploitation outcomes better than demographic models?

2.2 Invisible Manipulation Tactics

Bandura's moral disengagement theory describes eight mechanisms through which people engage in harmful behavior while maintaining moral self-regard:

1. Moral justification (reframing harm as serving higher purpose)
2. Euphemistic labeling (sanitizing language)
3. Advantageous comparison (comparing to worse alternatives)
4. Displacement of responsibility (diffusing accountability)
5. Diffusion of responsibility (distributing blame)
6. Distortion of consequences (minimizing harm)
7. Dehumanization (denying human qualities)
8. Attribution of blame (victim-blaming)

These mechanisms operate through language. Yet no systems currently detect them in institutional communications.

Research Question 2: Can natural language processing algorithms identify these mechanisms in emails, medical records, and institutional correspondence with sufficient accuracy for real-time alerts?

2.3 Procedural Burden as Harm Multiplier

Tyler's procedural justice research demonstrates that *how* decisions are made affects well-being independent of outcomes. We propose procedural burden operates as a harm *multiplier*—amplifying the impact of narrative violations.

Measurable procedural burden indicators include: - Circular referrals (Kafka loops) - Redundant documentation demands (document deluge) - Asymmetric deadlines (institutions can delay indefinitely; individuals face strict limits) - Vague demands impossible to fulfill (specificity debt)

Research Question 3: Does a multiplicative model (subjective harm $\times$ procedural burden) predict psychological outcomes better than additive models?

2.4 Treatment-Resistant Conditions as Unified Architecture

Most mental health conditions show 40-70% treatment resistance. Rather than viewing this as diagnostic failure, we propose treatment-resistant cases share a common three-component architec-

ture:

1. **Cognitive flooding**: Excessive pre-linguistic cognitive activations preventing deliberate processing
2. **Physiological destabilization**: Cardiovascular and autonomic dysregulation preventing cognitive stability
3. **Prediction-manufacturing**: Engineering outcomes that confirm catastrophic predictions rather than updating predictions based on evidence

This pattern appears analogous to slot machine manipulation (rapid decision cascades, variable reinforcement, arousal induction), suggesting institutional tactics may deliberately activate this vulnerability.

Research Question 4: Do treatment-resistant cases demonstrate this three-component pattern at higher rates than treatment-responsive cases?

2.5 Heterosis in Psychological Intervention

Biological heterosis (hybrid vigor) occurs when combining distinct genetic lineages produces offspring with enhanced fitness beyond either parent. We propose an analogous principle in psychological intervention: combining three complementary components that mask each other's weaknesses may produce non-linear therapeutic effects:

1. **Linguistic precision training** (reducing cognitive flooding through emotional granularity)
2. **Rhythmic stabilization** (passive cardiovascular entrainment via interval timers)
3. **Narrative interruption** (disrupting identity-consolidating catastrophic predictions)

Research Question 5: Do combined interventions produce effects exceeding the sum of individual components?

---

3. RESEARCH DESIGN OVERVIEW

3.1 Phase 1: Empirical Validation (18 months)

Study 1: Vulnerability Index Validation - Large-scale longitudinal cohort (N 1,000) - Baseline assessment + prospective tracking of institutional exploitation events - Target: Vulnerability index

predicts outcomes with OR 2.0, AUC 0.75

Study 2: Treatment-Resistant Architecture Detection - Multi-method assessment across groups (treatment-resistant vs. responsive vs. healthy) - Experience sampling, biosensor monitoring, neuroimaging substudies - Target: Three-component pattern distinguishes treatment-resistant cases

Study 3: Manipulation Detection Algorithm Development - Expert annotation of institutional communications corpus (N=10,000 documents) - Supervised machine learning classifier development - Target: Precision/recall 0.80 for manipulation tactic detection

Study 4: Digital Dignity Index Validation - Cross-sectional correlation study linking procedural metrics to psychological outcomes - Test multiplicative vs. additive harm models - Target: Multiplicative model explains additional variance (ΔR^2 0.10)

3.2 Phase 2: Algorithm Development (12-18 months)

- Machine learning vulnerability prediction from multimodal data
- Real-time manipulation detection API
- Wearable biosensor integration for treatment-resistant pattern monitoring
- Cryptographic evidence generation system for legal documentation

3.3 Phase 3: Clinical Validation (24 months)

Randomized Controlled Trials: - Tri-component intervention vs. individual components vs. treatment-as-usual - Algorithm-guided intervention selection vs. clinician judgment - Target population: Treatment-resistant anxiety/depression with demonstrated three-component architecture

Primary Outcomes: - Reduction in treatment-resistant pattern markers - Functional improvement - Quality of life enhancement

3.4 Phase 4: Translation and Impact (24+ months)

- Large-scale deployment (N=5,000-10,000 users)
- Institutional audit application (manipulation scoring across organizations)
- Regulatory framework development
- Public transparency initiatives

---

4. INTERDISCIPLINARY REQUIREMENTS

This research requires integration across multiple fields:

Clinical Psychology: Assessment development, clinical trials, treatment-resistant condition expertise

Computer Science: NLP, machine learning, algorithm development, software engineering

Cognitive Neuroscience: Neuroimaging, psychophysiology, biosensor integration

Biostatistics: Longitudinal modeling, psychometric validation, causal inference

Public Health: Health equity, implementation science, population impact

Philosophy/Ethics: Value assessment methodology, autonomy theory, AI ethics

Law: Evidence admissibility, regulatory compliance, privacy protection

Human-Computer Interaction: User experience design, intervention delivery platforms

---

5. EXPECTED IMPACT

5.1 Scientific Contributions

Paradigm shifts: - From demographic to architectural vulnerability models - From diagnosis-specific to architectural treatment frameworks - From single-mechanism to heterosis-based interventions - From subjective to objectively quantified psychological harm - From surveillance AI to autonomy-protective AI

Novel theoretical contributions: - First unified vulnerability framework integrating philosophy, psychology, neuroscience - Computational operationalization of authenticity and autonomy - Architectural model of treatment resistance - Neural mechanisms of semantic diversity effects

5.2 Clinical Applications

- Precision mental health: algorithm-guided intervention selection
- New treatment options for 40-70% who fail standard approaches
- Treatment-resistant condition targeting
- Trauma-informed institutional harm recognition

- FDA-clearable digital therapeutic platforms

5.3 Societal Impact

Institutional accountability: - Objective manipulation scoring - Public institutional rankings - Evidence for regulatory action - Legal documentation tools

Regulatory frameworks: - Iatrogenic AI risk assessment standards - Emotional AI safety guidelines - Consumer protection standards - Institutional practice certifications

Health equity: - Validation of marginalized experiences - Disparity detection algorithms - Accessible interventions - Anti-surveillance commitments

---

6. ETHICAL COMMITMENTS

6.1 Anti-Surveillance Design

This system is designed for **autonomy protection**, not surveillance or institutional control: - User control over all data (deletion rights, export rights) - Local processing where feasible - Transparent, explainable algorithms - No sale of user data - No institutional access without explicit consent

Prohibited uses: Employer surveillance, institutional screening to deny services, government surveillance, predictive policing, any use increasing power asymmetry.

6.2 Health Equity Focus

- Oversample marginalized groups in validation
- Fairness-aware machine learning
- Free/sliding scale access for low-income users
- Culturally adapted assessments
- Community advisory boards
- Participatory research designs

6.3 Open Science

- Open-access publication where possible
- Pre-registration of clinical trials

- De-identified data sharing (with protections)
- Open-source research tools
- Publication of null results

---

7. FUNDING AND TIMELINE

Estimated Program Scope: $12-19M over 6 years

Potential Funding Sources: - NIH (NIMH, NIDA, NIAAA): Treatment development, digital mental health - NSF (CISE, SBE): Human-centered computing, methodology innovation - AHRQ: Patient safety, healthcare quality - PCORI: Patient-centered outcomes research - Foundations: Robert Wood Johnson, Open Society, Mozilla, Wellcome Trust

Expected Outputs: - 30-50 peer-reviewed publications - 10-15 doctoral dissertations - Validated assessment tools - Deployed software systems - Regulatory frameworks - Institutional accountability mechanisms

---

8. PARTNERSHIP OPPORTUNITIES

We seek university partners to transform comprehensive theoretical frameworks into empirically validated research protocols and functional AI systems. This represents an extraordinary opportunity for:

- **Faculty researchers** seeking paradigm-shifting, high-impact work
- **Graduate students** seeking dissertation topics with clinical and social significance
- **Institutions** seeking to establish leadership in digital ethics and human-centered AI
- **Interdisciplinary centers** requiring complex, multi-method research programs

Specific partnership needs: 1. **Research leadership**: PI with expertise in clinical psychology and treatment-resistant conditions 2. **Co-investigators**: Computer science, neuroscience, biostatistics, public health 3. **Infrastructure**: Neuroimaging facilities, biosensor equipment, computing resources 4. **Clinical trial capacity**: Treatment-resistant patient populations, clinical trial infrastructure 5. **Institutional commitment**: Multi-year funding pursuit, intellectual property negotiation

---

9. INTELLECTUAL PROPERTY STATUS

Provisional patent applications have been filed (November 2025) covering theoretical frameworks and system architectures. This working paper intentionally avoids detailed disclosure of specific algorithms, mathematical formulas, or implementation details to preserve patent rights.

University partnership models under consideration: - Research license (university uses IP for research, revenue sharing for commercialization) - Joint ownership (if substantial algorithmic contributions) - University assignment with royalty

All models prioritize social impact alongside commercial viability, with commitments to: - Free/low-cost access for vulnerable populations - Pro bono evidence generation for public interest cases - Open-source research tools - Ethical use restrictions (no surveillance applications)

---

10. CONCLUSION

Institutional manipulation represents a public health crisis hiding in plain sight. Millions experience systematic psychological harm while navigating healthcare, insurance, disability evaluation, and government processes—harm that is dismissed, minimized, and attributed to individual pathology rather than institutional design.

This research program offers a path forward: empirically validated systems to predict vulnerability, detect manipulation, quantify harm, and intervene effectively. It represents not incremental improvement but paradigm-shifting work that could:

- Create new diagnostic categories cutting across traditional boundaries
- Develop AI systems serving human autonomy rather than extraction
- Provide accountability mechanisms for powerful institutions
- Offer new treatments for millions with treatment-resistant conditions
- Establish regulatory frameworks for ethical AI design

The theoretical foundations exist. The intellectual property is secured. The research questions are clear. What remains is empirical validation, algorithm construction, and clinical deployment—requiring university partnership, interdisciplinary collaboration, and sustained funding.

This is work that changes fields, helps millions, and redefines how we understand vulnerability, autonomy, and institutional responsibility in an increasingly digital world.

REFERENCES

Bandura, A. (2016). Moral disengagement: How people do harm and live with themselves. Worth Publishers.

Frankfurt, H. G. (1988). The importance of what we care about. Cambridge University Press.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture, 2(1).

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality. Psychological Review, 102(2), 246-268.

Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. Annual Review of Psychology, 57, 375-400.

Zuboff, S. (2019). The age of surveillance capitalism. Public Affairs.

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

CONTACT

Joshua M. Garfunkel

Disrupt the Loop

Email: [your email]

Website: disruptttheloop.com/papers/VI-DDI-working-paper-2025

LinkedIn: [your LinkedIn URL]

For partnership inquiries: Interested researchers and institutions are invited to contact the author to discuss potential collaboration opportunities, including research leadership, co-investigation, resource sharing, and funding partnerships.

This working paper is intentionally limited in technical detail to preserve intellectual property rights while establishing the conceptual framework and research program scope. Detailed methodologies, algorithms, and system architectures are available under appropriate confidentiality agreements for serious research partners.

Last updated: December 2025

Status: Seeking university research partners

Funding status: Pre-submission (grant applications in preparation)