Request for Comments: RFC FS02
(Updates: v0.2.x Series)
Category: Experimental / Informational
Author: Rob Tyrie, Ironstone Advisory & Finger Spider Project
Date: April 1, 2025 (Let's just assume this date is canonical now)
Title: The Bullshit Compass: Semantic Pattern Detection Framework v0.88 (Beta Candidate)

(Still adhering to the RFC spirit for structured development. Version 0.88 - getting closer!)

Status of this Memo

This document describes version 0.88 of the Bullshit Compass framework, representing a significant iteration incorporating feedback, refinement of core indicators, and initial experimental modules. While still considered experimental (pre-1.0), v0.88 aims for a more robust and feature-complete specification based on earlier conceptual work. This RFC is intended for broader review and testing. Distribution remains unlimited.

Abstract

The Bullshit Compass project seeks to address the challenge of language produced with indifference to truth (Frankfurt, 1986) – "bullshit" – which pervades modern communication. This framework specifies a system for detecting such language patterns using a hybrid approach: philosophical grounding (Frankfurt, Orwell, Grice), psychological insights (Kahneman), linguistic analysis, and NLP/AI techniques. Version 0.88 refines the core indicators, introduces experimental modules (e.g., for quantitative rhetoric), clarifies the role of sociological context (Graeber), and outlines steps towards multilingual support. The goal remains a practical tool to enhance critical reading and promote linguistic accountability.

Table of Contents

1. Introduction: The Problem Persists, The Tool Matures
2. Core Concepts & Terminology (Refined)
3. Bullshit Compass v0.88: Architecture & Philosophy
4. Core Indicators (v0.88 Definitions)
   4.1. Tautological Camouflage Rate (TCR)
   4.2. Ideological Smuggling Index (ISI) (Formerly PWI)
   4.3. Semantic Drift Score (SDS)
   4.4. Assertion Ambiguity Ratio (AAR)

4.5. Buzzword Load Index (BLI)
4.6. Pseudoscience Rhetoric Marker (PRM)
4.7. Emotional Persuasion Flags (EPF)
4.8. Orwellian Obfuscation Index (OOI)
4.9. Data Presentation Obfuscation (DPO) (Experimental)
5. Contextual Analysis: LGC Module (v0.88)
5.1. Language Game Contextualization (LGC)
6. LLM Integration Strategy (v0.88)
7. Outputs & Visualization (v0.88 Specification)
7.1. Standard Outputs
7.2. Target Applications
8. Implementation Status & Roadmap
8.1. Core Engine (Python Beta)
8.2. Integrations (Prototypes)
8.3. Licensing (Open Source - MIT)
9. Advanced Topics & Future Work (Post-v1.0)
9.1. Deep Cross-Cultural Analysis
9.2. Real-time Processing
9.3. Advanced Satire/Irony Detection
10. Ethical Considerations & Responsible Use (Mandatory Reading!)
11. Project Goals & Vision
12. Call for Testing & Feedback (v0.88 Focus)
13. Foundational Bibliography
14. Contact Information
15. Legal Framework

1. Introduction: The Problem Persists, The Tool Matures

The challenge identified by Frankfurt (1986) – language unmoored from truth-value – hasn't diminished. If anything, the ease of content generation and the pressures of modern communication have amplified the need for tools that can help distinguish substance from performance. Early versions of this RFC explored the problem space; v0.88 represents a more concrete proposal for a functional system.

We also acknowledge the background context provided by thinkers like Graeber (2018). While the Compass analyzes *text*, not authors or their jobs, understanding potential systemic pressures that might incentivize "bullshit knowledge objects" reinforces the need for robust, objective textual analysis. This version aims to deliver a more refined instrument for that purpose, drawing on insights from Orwell, Grice,

Kahneman, Snyder, Bergstrom/West, and others.

2. Core Concepts & Terminology (Refined)

(Definitions refined for clarity, consistent with previous versions but presented with more confidence)

- **Bullshit (Frankfurt):** Language used with indifference to truth or accuracy.
- **Epistemic Quality:** The degree to which language appears oriented towards truth, evidence, and clarity. The Compass seeks markers of *low* epistemic quality.
- **Semantic Pattern Detection:** Analyzing text for recurring meaning-based structures indicative of specific rhetorical strategies or epistemic stances.
- **Language Game (Wittgenstein):** The contextual rules governing language use in a specific activity (report, satire, marketing). Crucial for the LGC module.
- **Gricean Maxims:** Cooperative principles (Quality, Quantity, Relation, Manner) often violated by bullshit.
- **Orwellian Language:** Obfuscatory, vague, or euphemistic language hindering clear thought (cf. Orwell, Snyder).
- **System 1 / System 2 (Kahneman):** Bullshit often targets fast, intuitive System 1; the Compass aims to engage slower, analytical System 2.
- **Bullshit Jobs / Knowledge Objects (Graeber Context):** Sociological concepts providing background on *potential* drivers for bullshit production; *not* directly measured by the Compass.

3. Bullshit Compass v0.88: Architecture & Philosophy

The v0.88 architecture is a modular scoring engine built primarily in Python. It processes text through a pipeline:

1. Preprocessing (tokenization, sentence boundary detection, etc.).
2. Indicator Analysis (parallel application of core indicator modules - Section 4).
3. Contextual Modulation (LGC module adjusts raw scores - Section 5).
4. LLM Augmentation (Optional enhancement for dynamic terms/context - Section 6).
5. Aggregation & Output Generation (Composite score, radar chart, annotations - Section 7).

Philosophically, it remains a tool intended to augment critical thinking (Parrish), not replace it. It provides data points on rhetorical patterns, prompting users to examine the text more closely.

4. Core Indicators (v0.88 Definitions)

These indicators form the heart of the detection engine. V0.88 includes refinements and one new experimental module.

4.1. Tautological Camouflage Rate (TCR)
Detects circular statements and semantically empty affirmations. (Stable module).
4.2. Ideological Smuggling Index (ISI) (Refined/Renamed from PWI)
Identifies attempts to embed potentially contentious or unsupported claims within structures that seem self-evident or broadly agreeable. Focuses on specific argumentative patterns rather than authorial intent. (Revised module, requires careful interpretation).
4.3. Semantic Drift Score (SDS)
Measures topic coherence and relevance shifts within the text. High scores suggest potential obfuscation or lack of focus. (Stable module).
4.4. Assertion Ambiguity Ratio (AAR)
Quantifies hedging, vagueness, use of weasel words, and non-falsifiable statements. (Stable module).
4.5. Buzzword Load Index (BLI)
Measures density of jargon, acronyms, and neologisms, particularly those undefined or used outside appropriate contexts (cf. Pennycook). (Continuously updated dictionary).
4.6. Pseudoscience Rhetoric Marker (PRM)
Flags misuse of scientific terminology or reasoning patterns in inappropriate contexts (e.g., "quantum healing"). (Refined module).
4.7. Emotional Persuasion Flags (EPF)
Detects language primarily aimed at evoking strong emotional responses (fear, tribalism, excessive flattery) rather than rational argument (cf. Kahneman System 1, Snyder propaganda warnings). (Stable module).
4.8. Orwellian Obfuscation Index (OOI)
Targets specific linguistic markers identified by Orwell (1946): dying metaphors, pretentious diction, verbal false limbs, meaningless words. (Refined pattern matching).
4.9. Data Presentation Obfuscation (DPO) (Experimental)
Inspired by Bergstrom & West (2020). Attempts to flag linguistic patterns often used to misrepresent or obscure quantitative data presented within text (e.g., selective reporting, misleading comparisons, unclear denominators). (Experimental module - feedback needed).

5. Contextual Analysis: LGC Module (v0.88)

5.1. Language Game Contextualization (LGC)
This crucial module attempts to identify the text's genre or communicative context (report, marketing, satire, formal speech, etc.) using stylistic and structural cues. It then modulates the raw scores from the core indicators based on context-specific expectations (e.g., higher jargon tolerance in technical papers, higher ambiguity tolerance in legal disclaimers). V0.88 uses improved genre models but remains an area of active development.

6. LLM Integration Strategy (v0.88)

LLMs are *not* core to the v0.88 engine but can be optionally integrated via API calls

for:

- **Dynamic Vocabulary:** Updating BLI/OOI dictionaries with emerging jargon or euphemisms.
- **Enhanced LGC:** Providing secondary input for genre classification.
- **Zero-Shot Flagging:** Experimentally prompting LLMs to flag novel or subtle forms of bullshit not caught by existing indicators (use with extreme caution).

7. Outputs & Visualization (v0.88 Specification)

7.1. Standard Outputs

- **Composite Bullshit Score (CBS):** 0-100 scale, representing overall detected bullshit patterns, weighted and context-adjusted. *Interpret with care*.
- **Indicator Score Profile:** Scores for each individual indicator (TCR, ISI, SDS, AAR, BLI, PRM, EPF, OOI, DPO). Typically visualized as a radar chart.
- **Annotated Text:** Input text with specific phrases/sentences flagged by relevant indicators.
- **Confidence Level:** An overall confidence score for the analysis, reflecting ambiguity or LGC uncertainty.

7.2. Target Applications
Research integrity, due diligence, internal comms review, AI content QC, journalism/fact-checking support, educational tool for critical writing/reading.

8. Implementation Status & Roadmap

8.1. Core Engine (Python Beta)
The core scoring engine and indicators (excluding DPO) are implemented in Python 3.x, available as a beta library and CLI tool. Performance optimized for moderate document lengths.
8.2. Integrations (Prototypes)
Early-stage prototypes exist for:

- Google Docs Add-on
- Outlook Plugin (Client-side analysis)
- Basic Web API

8.3. Licensing (Open Source - MIT)
The core library is intended for release under the MIT license to encourage adoption and contribution.

9. Advanced Topics & Future Work (Post-v1.0)

While v0.88 is a significant step, major challenges remain for future versions:
9.1. Deep Cross-Cultural Analysis
Moving beyond English requires extensive, culturally sensitive research into linguistic markers

across diverse languages (as discussed in v0.2.x). This remains a long-term goal requiring significant collaboration. Preliminary framework design for multilingual architecture is underway.

9.2. Real-time Processing

Adapting the engine for real-time analysis (e.g., meetings, chat) presents performance and significant ethical/privacy challenges.

9.3. Advanced Satire/Irony Detection

Distinguishing deliberate, sophisticated irony/satire from genuine bullshit remains difficult and requires more advanced modeling.

10. Ethical Considerations & Responsible Use (Mandatory Reading!)

This tool analyzes *language patterns*, not *truth* or *intent*. Misuse is a significant concern.

- **NO Truth Claims:** The Compass does *not* fact-check. A low score doesn't mean true; a high score doesn't mean false.
- **NO Intent Inference:** The scores reflect linguistic patterns, *not* proven authorial intent to deceive.
- **Context is CRITICAL:** Always consider the LGC module's assessment and your own understanding of the context.
- **Augment, Don't Replace:** Use scores to prompt critical thinking (System 2), not as a definitive judgment.
- **Avoid Weaponization:** Do not use scores to unfairly dismiss arguments or individuals ("score-shaming").
- **NO Author Profiling:** Explicitly reject using scores to speculate about the author's job, character, or motivations (re: Graeber context). Judge the text.
- **Bias Awareness:** Be aware that algorithmic bias is possible; ongoing work aims to mitigate this.

11. Project Goals & Vision

To provide a robust, transparent, and ethically deployed tool that helps individuals and organizations identify and understand potentially misleading or epistemically poor language, ultimately fostering clearer communication and more critical engagement with text.

12. Call for Testing & Feedback (v0.88 Focus)

We seek feedback on this v0.88 specification and beta implementation, particularly regarding:

- Clarity and utility of the refined indicators (especially ISI).
- Effectiveness and potential risks of the experimental DPO indicator.

- Accuracy and robustness of the LGC module across different genres.
- Usability of the outputs (scores, annotations).
- Potential ethical blind spots or misuse scenarios.
- Performance and integration challenges.

13. Foundational Bibliography

(Includes Frankfurt, Orwell, Wittgenstein, Grice, Kahneman, Ariely, Bergstrom/West, Bregman, Snyder, Parrish, Pennycook, Meyer, Peterson (as influence on ISI), Graeber (for context), relevant RFCs. Ensure citations are complete in a final document).

14. Contact Information

(Provide appropriate contact details for feedback).

15. Legal Framework

(MIT License details, Disclaimers of Warranty, etc.).