**Scott D. Blain, PhD**

Quincy, MA 02169 | (270) 287-8688 | [scottdougblain@gmail.com](mailto:scottdougblain@gmail.com) | [www.linkedin.com/in/scott-blain-phd/](http://www.linkedin.com/in/scott-blain-phd/)

**Model Behavior Architect**

***Bridging psychology, neuroscience, and AI safety to develop frameworks for reliable, aligned systems***

Cognitive scientist applying insights from human intelligence to AI safety and alignment challenges. Leveraging expertise in pattern recognition, social cognition, personality, and AI model evaluation frameworks to support development of beneficial AI systems aligned with human values. Effective communicator and collaborator; adept at translating complex concepts across specialized domains, facilitating creative approaches to novel problems, and leading interdisciplinary research and design projects.

Research Design & implementation | Statistical Analysis | Data Interpretation | Scientific Writing | AI Ethics & Governance

AI Model Evaluation & Benchmarking Design | Neural Network Analysis & Interpretation | Pattern Recognition & Anomaly Detection

Prompt Engineering & Model Behavior Testing | Interdisciplinary Thinking | Analytical Reasoning & Systematic Problem Solving

**Technical Skills**

***Programming:*** Python, R, MATLAB

***ML/AI:*** Large Language Models (LLM), Deep Learning, RLHF, Prompt Engineering, Finetuning, Evaluation

***Analysis:*** Bayesian Modeling, Diffusion Models, Structural Equation Modeling

***Experimental Design:*** System Evaluation, Behavior Testing, Psychometrics

**Experience**

**Cohere, San Francisco, CA | Snorkel, Redwood City, CA**

**Senior Data Quality Specialist (Freelance) 2024 – Present**

Design prompt engineering strategies and evaluation frameworks for LLM outputs. Implement data generation pipelines and verification systems to improve model responses. Identify edge cases and behavioral inconsistencies through systematic testing.

* Collaborated with ML researchers to enhance models while ensuring quality and reliability.
* Developed benchmarks to assess model behavior across domains (e.g., honesty, factuality).

**University of Michigan, Ann Arbor, MI | The Ohio State University, Columbus, OH**

**Postdoctoral Scholar 2023 – Present**

Lead computational psychiatry research on cognitive control systems and social cognition.

* Created evaluation frameworks assessing pattern recognition and social decision-making.
* Published findings on how social intelligence enables both cooperation and strategic deception.
* Applying artificial neural networks and deep learning to multimodal neurobehavioral data.

**Education & Professional Development**

**University of Minnesota, Minneapolis, MN**

**PhD in Psychology/Neuroscience**

*NSF Graduate Research Fellow* | *Thesis on individual differences in social cognition/behavior*

**Master of Arts (MA) in Psychology/Neuroscience**

*Thesis on apophenia and exploration-constraint balance*

**Vanderbilt University, Nashville, TN**

**Bachelor of Science (BS) in Cognitive Science**

**Global Challenges Project 2025**

*Emerging challenges in AI safety and biosecurity and interdisciplinary approaches to existential risks*

**Ethics of AI | University of Helsinki 2025**

*Ethical AI development/use and applying frameworks from moral philosophy to questions related to contemporary AI*

**AI Safety Fundamentals | BlueDot Impact 2024**

*Technical alignment studies, including inner/outer alignment, interpretability, and safety frameworks*

**Areas of Research**

**Pattern Recognition & Hallucination Analysis**

* Applied signal detection theory to quantify false-positive errors in pattern recognition.
* Mapped neural mechanisms underlying apophenia (seeing patterns where none exist).
* Published papers on human apophenia and connections to AI hallucinations.
* Developed frameworks for identifying and mitigating hallucinations in generative models.

**Social Intelligence & Alignment Research**

* Investigated how mentalizing abilities may facilitate both compassion and deception.
* Illuminated social-intelligence mechanisms through imaging and computational models.
* Developed experimental paradigms to assess social reasoning capabilities.
* Creating comprehensive evaluation suite for LLM mentalizing abilities.
* Generating prompting framework to enhance alignment-relevant mentalizing capabilities.

**Cybernetic Theories of Personality**

* Contributed to theory and empirical research conceptualizing personality as parameter variations of cybernetic systems, with relevant applications to AI behavior.
* Building human-compatible systems for personality-inspired fine tuning of LLMs.

**Professional Memberships**

Association for Psychological Science 2016 – Present

Society for Personality and Social Psychology 2018 – Present

Association for Research in Personality 2019 – Present

Society of Research in Psychopathology 2021 – Present

Hierarchical Taxonomy of Psychopathology Consortium 2021 – Present

**Selected Conference Presentations**

**Blain, S.D., et al. (2025).** Aberrant cerebellar-cortical connectivity contributes to social cognitive deficits in schizophrenia**. Society of Biological Psychiatry, Toronto, ON, Canada.**

**Blain, S.D., et al. (2021). The promise of task‑based measures in personality research. In: Unique strength of multi‑method multi‑informant approaches in personality assessment. Society for Personality and Social Psychology, Virtual Conference.**

**Blain, S.D., et al. (2020). Affiliation on the brain: Neural mechanisms of prosociality and social cognition. Society for Personality and Social Psychology, New Orleans, LA.**

**Blain, S.D., et al. (2019). Functional brain correlates of the openness‑psychoticism continuum. International Society for the Study of Individual Differences, Florence, Italy.**

**Selected Publications**

**Blain, S.D.** (2024). "Hallucinations Aren't New: What Human Psychology Can Teach Us About AI Safety." Substack.

**Blain, S.D.**, et al. (2020). "Apophenia as the Disposition to False Positives: A Unifying Framework for Openness and Psychoticism." Journal of Abnormal Psychology, 129(3), 279-292.

**Blain, S.D.**, et al. (2020). "Toward a Neural Model of the Openness-Psychoticism Dimension." Schizophrenia Bulletin, 46(3), 540-551.

**Blain, S.D.**, et al. (2022). "Activation of the Default Network during a Theory of Mind Task Predicts Individual Differences in Agreeableness and Social Cognitive Ability." Cognitive, Affective, and Behavioral Neuroscience, 22(2), 383-402.

**Blain, S.D.**, et al. (2021). "Extraversion but not Depression Predicts Reward Sensitivity." Journal of Personality and Social Psychology, 121(2), e-e18.