Appendix E

This appendix is an independent analysis based on the framework presented in:
A dynamical model of subjectivity integrating affective gain, cognitive bias, and self-regulation
doi:10.31234/osf.io/8dbft_v2

A Mathematical Interpretation of Adam Smith's "The Theory of Moral Sentiments"

Takeo Imaizumi
Independent Researcher, Kyoto City, Japan

June 17, 2025

Abstract

This appendix provides a mathematical re-interpretation of Adam Smith's classic work, *The Theory of Moral Sentiments*, using the G–μ–S framework. We demonstrate profound parallels between Smith's philosophical constructs and the model's computational components. Core concepts such as the "Impartial Spectator" are mapped to the Self-System (S), "Virtue" to the Ideal Dynamical Equilibrium (IDE), and the development of a "Learned Moral Character" to the optimization of the meta-parameter ($\boldsymbol{\theta}_{\mathrm{S}}$). This analysis suggests that 18th-century moral philosophy can be seen as a surprisingly robust blueprint for 21st-century models of subjectivity and artificial agency.

The surprising relevance of 18th-century moral philosophy

Can 18th-century philosophy be modeled by 21st-century mathematics? The modern scientific quest for a formal, universal model of the mind often looks forward, yet surprisingly robust clues can be found by looking back. Adam Smith's foundational work, *The Theory of Moral Sentiments* (1759), a cornerstone of the Scottish Enlightenment, offers more than historical wisdom; it presents a stunningly detailed blueprint for a modern computational model of the mind. This appendix illustrates the profound parallels between Smith's remarkably prescient philosophy and the dynamical model of subjectivity.

His work was a significant departure from the reason-centric moral theories of his time. Smith placed "sympathy"—the capacity to imaginatively share in the feelings of others—at the heart of morality. Equally groundbreaking was his theory of the "Impartial Spectator," which explained that we make moral judgments by evaluating our own conduct from the perspective of an informed but unbiased third party. This internal simulation and self-regulation mechanism, once considered purely philosophical, finds a precise new language in the computational framework.

Mapping Smith's Philosophy to the G–μ–S Model

The following table details the striking correspondence between Smith's philosophy and the mathematical model.

The Theory of Moral Sentiments	**The Mathematical Model**
Impartial Spectator	Self-System (S)
Virtue / Propriety	Ideal Dynamical Equilibrium (IDE)
Self-Command	Dynamic Adjustment of G & μ
Pursuit of Virtue & Happiness	Optimization of Objective Function ($\mathcal{L}$)
General Rules of Morality / Learned Moral Character	Meta-parameter ($\boldsymbol{\theta}_S$) that governs the Self-System

The Spectator as a Computational Self-System

The most direct parallel lies between Smith's "Impartial Spectator" and the model's Self-System (S). The Spectator is the internalized, unbiased viewpoint from which we moderate our passions and judge our own conduct to achieve what Smith called "propriety." Computationally, the Self-System performs an analogous function: it is a higher-order meta-controller that adaptively regulates affective gain (G) and cognitive bias (μ) to achieve a state of balanced responsiveness, or what the model terms the Ideal Dynamical Equilibrium (IDE).

Moral Learning as Meta-Parameter Optimization

Even more profoundly, the process Smith describes for developing a "Learned Moral Character" through experience finds a direct mathematical analogue in learning the model's meta-parameter ($\boldsymbol{\theta}_S$). For Smith, we derive "general rules of morality" from repeated observations of what is approved or disapproved of by our inner Spectator. This is a lifelong learning process. In the model, $\boldsymbol{\theta}_S$ represents precisely these learned, underlying rules that shape the Self-System's regulatory policy. Smith's "Pursuit of Virtue & Happiness" is computationally framed as the continuous, goal-directed optimization of the objective function ($\mathcal{L}$), which in turn refines and solidifies the meta-parameter ($\boldsymbol{\theta}_S$).

Conclusion: Onko-chishin as a Blueprint for Future Science

This re-examination is a form of "Onko-chishin"[1]. It demonstrates that Smith's 18th-century masterpiece is not merely a historical artifact but a rich, structured theory of the human mind that, when viewed through a modern computational lens, is revived as a stunningly relevant blueprint for understanding subjectivity, self-regulation, and even for designing the ethical architecture of future artificial agents.

[1] Onko-chishin is a Japanese proverb derived from the Analects of Confucius. It translates to "learning from the a past" or, more deeply, "discovering new insights by studying the old."