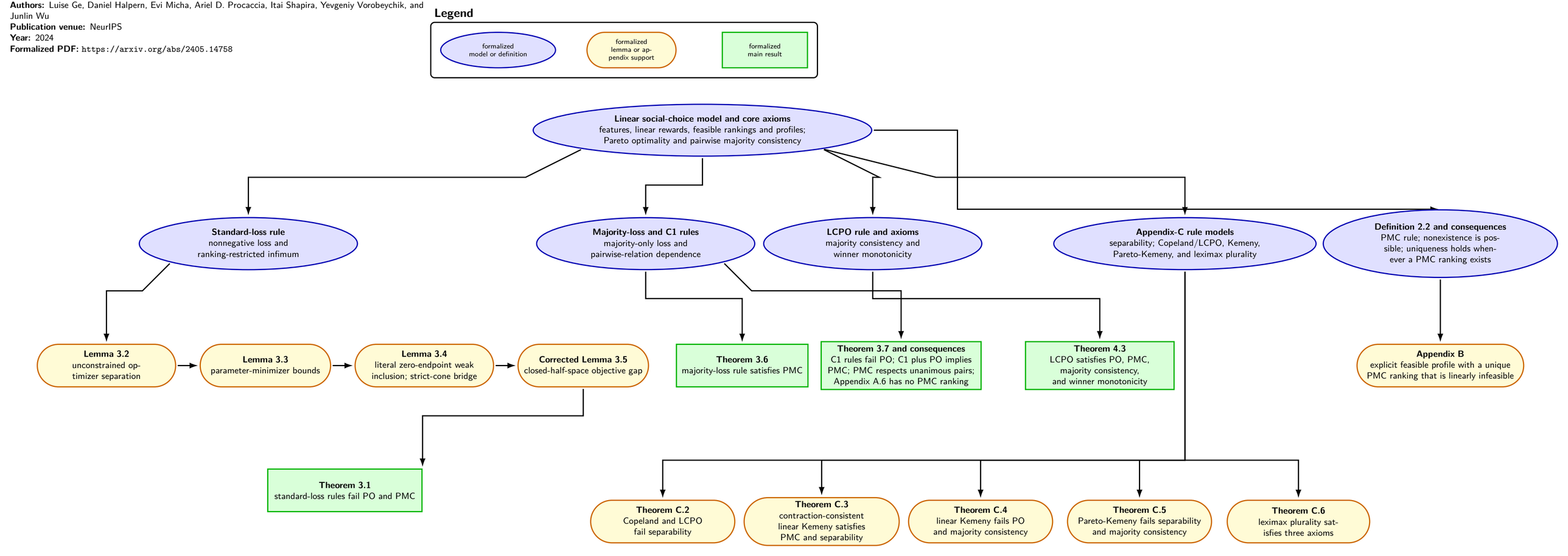Paper: Axioms for AI Alignment from Human Feedback
Authors: Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu
Publication venue: NeurIPS
Year: 2024
Formalized PDF: `https://arxiv.org/abs/2405.14758`