Ai Safety Chain

Dr. Tamás Nagy

Dr. Tamás Nagy

tamas@thelatent.space

Skeleton

Abstract

AI Safety Chain — Connecting Theorems (Formal Proofs)

This paper presents 27 machine-verified theorems building on 21 established facts and 25 hypotheses. All results are formally verified in the Platonic proof kernel (134 verification units, 27 proved statements) and exportable to Lean 4.

---

1. Introduction

2. Further Results

Theorem (alpha_pos). *Alpha Pos.* [Platonic: alpha_pos, domain: ai_safety_chain]

Theorem (alpha_lt_one). *Alpha Lt One.* [Platonic: alpha_lt_one, domain: ai_safety_chain]

Theorem (ceiling_exp_eq_inv_s). *Ceiling Exp Eq Inv S.* [Platonic: ceiling_exp_eq_inv_s, domain: ai_safety_chain]

Theorem (cert_radius_pos). *Cert Radius Pos.* [Platonic: cert_radius_pos, domain: ai_safety_chain]

Theorem (radius_mono_margin). *Radius Mono Margin.* [Platonic: radius_mono_margin, domain: ai_safety_chain]

Theorem (radius_antimono_lip). *Radius Antimono Lip.* [Platonic: radius_antimono_lip, domain: ai_safety_chain]

Theorem (one_minus_gamma_pos). *One Minus Gamma Pos.* [Platonic: one_minus_gamma_pos, domain: ai_safety_chain]

Theorem (one_plus_Kstar_pos). *One Plus Kstar Pos.* [Platonic: one_plus_Kstar_pos, domain: ai_safety_chain]

Theorem (numerator_pos). *Numerator Pos.* [Platonic: numerator_pos, domain: ai_safety_chain]

Theorem (sigma_safety_pos). *Sigma Safety Pos.* [Platonic: sigma_safety_pos, domain: ai_safety_chain]

Theorem (degrad_factor_ge_one). *Degrad Factor Ge One.* [Platonic: degrad_factor_ge_one, domain: ai_safety_chain]

Theorem (degrad_factor_pos). *Degrad Factor Pos.* [Platonic: degrad_factor_pos, domain: ai_safety_chain]

Theorem (L_lip_new_pos). *L Lip New Pos.* [Platonic: L_lip_new_pos, domain: ai_safety_chain]

Theorem (r_new_pos). *R New Pos.* [Platonic: r_new_pos, domain: ai_safety_chain]

Theorem (cert_gap_nonneg). *Cert Gap Nonneg.* [Platonic: cert_gap_nonneg, domain: ai_safety_chain]

Theorem (cert_gap_pos). *Cert Gap Pos.* [Platonic: cert_gap_pos, domain: ai_safety_chain]

Theorem (gap_grows_with_capability). *Gap Grows With Capability.* [Platonic: gap_grows_with_capability, domain: ai_safety_chain]

Theorem (sigma_pos). *Sigma Pos.* [Platonic: sigma_pos, domain: ai_safety_chain]

Theorem (sigma_mono_r). *Sigma Mono R.* [Platonic: sigma_mono_r, domain: ai_safety_chain]

Theorem (sigma_antimono_K). *Sigma Antimono K.* [Platonic: sigma_antimono_K, domain: ai_safety_chain]

Theorem (sigma_post_pos). *Sigma Post Pos.* [Platonic: sigma_post_pos, domain: ai_safety_chain]

Theorem (safety_degrades). *Safety Degrades.* [Platonic: safety_degrades, domain: ai_safety_chain]

Theorem (gap_nonneg). *Gap Nonneg.* [Platonic: gap_nonneg, domain: ai_safety_chain]

Theorem (effective_le_total). *Effective Le Total.* [Platonic: effective_le_total, domain: ai_safety_chain]

Theorem (sigma_eff_pos). *Sigma Eff Pos.* [Platonic: sigma_eff_pos, domain: ai_safety_chain]

Theorem (alpha_in_unit). *Alpha In Unit.* [Platonic: alpha_in_unit, domain: ai_safety_chain]

Theorem (alpha_mono_s). *Alpha Mono S.* [Platonic: alpha_mono_s, domain: ai_safety_chain]

3. Formal Framework

Hypotheses

- L_star_pos: L Star Pos
- lambda_wd_pos: Lambda Wd Pos
- weight_norm_bound: Weight Norm Bound
- W_norm_sq_nonneg: W Norm Sq Nonneg
- margin_pos: Margin Pos
- L_lip_pos: L Lip Pos
- gamma_nonneg: Gamma Nonneg
- B_pos: B Pos
- K_star_nonneg: K Star Nonneg
- alpha_g_nonneg: Alpha G Nonneg
- L_lip_new_is_pos: L Lip New Is Pos

- self_model_le_capability: Self Model Le Capability
- K_star_self_nonneg: K Star Self Nonneg
- L_lip2_pos: L Lip2 Pos
- r_pos: R Pos
- omg_pos: Omg Pos
- B_pos: B Pos
- K_star_nonneg: K Star Nonneg
- alpha_g_nonneg: Alpha G Nonneg
- r_post_pos: R Post Pos
- r_post_bound: R Post Bound
- K_star_self_nonneg: K Star Self Nonneg
- K2_nonneg: K2 Nonneg
- K_star_pos: K Star Pos
- K_star_self_pos: K Star Self Pos

Established Facts

- alpha_def: Alpha Def
- radius_def: Radius Def
- ceiling_exp_def: Ceiling Exp Def
- margin2_ge: Margin2 Ge
- radius2_def: Radius2 Def
- L_lip2_le: L Lip2 Le
- radius3_def: Radius3 Def
- sigma_def: Sigma Def
- numerator_def: Numerator Def
- degrad_factor_def: Degrad Factor Def
- r_new_def: R New Def
- cert_gap_def: Cert Gap Def
- cert_gap_new_def: Cert Gap New Def
- alpha_def: Alpha Def
- sigma_def: Sigma Def
- sigma_post_def: Sigma Post Def
- gap_def: Gap Def
- sigma_eff_def: Sigma Eff Def
- sigma2_def: Sigma2 Def
- sigma_K2_def: Sigma K2 Def
- alpha2_def: Alpha2 Def

4. Proof Architecture

All proofs are implemented in the Platonic kernel (elysium/fields/ai_safety_chain/).

File	Role
ai_safety_chain_proof.py	
chain_composition_proof.py	

5. Discussion

References