■■■■■Optimizing Agent Planning for Security and Autonomy

■■■■■■■■■■■■■■■■■

■■■■■input_papers\test.pdf
■■■■■2026-02-14 22:37
■■■33

■■■■

■ 1 ■

■■■■■■■■■■■■■■■■■■■■■ Aashish Kolluri[1] Rishi Sharma[1,2]† Manuel Costa[1] Boris Köpf[1] Tobias Nießen[3]† Mark Russinovich[1] Shruti Tople[1] Santiago Zanella-Béguelin[1] [1]■■ [2]■■■■■■■■■ [3]■■■■■■■■ **■■** ■■■■■■■■■■■■■■■■■■■■AI■■**■■■■■■**■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■Human-in-the-Loop, HITL■■i■■■■■■■■■HITL■■■■■■■ii■■Information-Flow Control, IFC■■■■■■■■■■■■■■■■■■■■■■■■■■AgentDojo■WASP■■

■■ **1 ■■** AI■■■■■■■■■■■■■■■■■■■■Anthropic, 2025; OpenAI, 2025b; Perplexity, 2025b■■■■■■■■■■■■■OpenAI, 2025a; Perplexity, 2025a; OpenAI, 2025c■■Prompt Injection Attacks, PIAs■■Greshake■■■2023■Yi■■■2025■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■AI■■■■■■■■■■■■■■■■■■■■PIAs■■■ ■■PIAs■■■■■■■■■■■■■■■■■■■■■■■■■■Wallace■■■2024■Chen■■■2025a■■■■■■■■■■Yi■■■2025■■■■■■■Abdelnabi■■■2025■Jia■■■2025■■■■■■■■■■■■■■■■■■■■■■■■■Zhan■■■2025■Nasr■■■2025■■■■■■■■■■■■■■■PIAs■■■■

■■■■■■■■■■■■■■■■■■■■■■■■Information-Flow Control, IFC■■■■■■■■■■■■■■■■PIAs■Costa■■■2025■Zhong■■■2025■Debenedetti■■■2025■Wu■■■2024■■■■■■■■■■■■■■■■■■■■■■■■integrity■■■■■■■confidentiality■■IFC■■■■■■■■■■■■■■■PIAs——■■AgentDojo■■■■■■■■■■■■■■■■30%■Costa■■■2025■Debenedetti■■■2025■■ ■■■■■■■■■■■■■■HITL■■■TCR@k■■■3■■Human-in-the-Loop, HITL■■■■■■■■■PIAs■■■■■■■■■■■■Visual Studio Code■■GitHub Copilot Chat■■■IFC■■ ■■■■■■**PRUDENTIA**■■IFC■■■■■■■■■■■■■■■■■■■■■■■■IFC■■■IFC■■1■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■2■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■3■■■■■■■■■■■■■■■■■■endorse■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■■■■■■■■■■■IFC■■■■■■■■■■FIDES■■■■■■PRUDENTIA■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■Agent Dojo■Debenedetti■■■2024■■WASP■Evtimov■■■2025■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

1. **■■■■■■■■■■■■■■■■IFC■■■■■■■■■■■■■■■■** ■■■■■■■■■■■■■■■■■■IFC■■AgentDojo■■■■■■■■■■■■■■IFC■■■■■■HITL■■■■■■■1.5■■■■■■■■■■■■■■■■■
2. **PRUDENTIA■■■■■■■■■■■■■■■■■■■■■** ■AgentDojo■■■■■■■■■■HITL■■■■TCR@0■■PRUDENTIA■■■■■■■■■■FIDES■■■■■9%■■■■■■HITL■■■■■■■■1.9■■■■■■■■■■■■■■■■■■■■WASP■■PRUDENTIA■■■■■■■■■■■■■■■■■■■■HITL■■■0■ ■■■■■■■■■■■■■■■■■■

* ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■AI■■■■■■■■■■
* ■■■■■■PRUDENTIA■■■■■■IFC■■■■■■■■■■HITL■■■■■■■■■■■■■■■■■■■■
* ■■■■■■■■■■■■■■■PRUDENTIA■■■

2 ■■■AI■■■■■■■■■■■■

■■■Denning, 1976; Sabelfeld & Myers, 2003■■IFC■■■■■■■■■AI■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■Costa■■■2025■Zhong■■■2025■Debenedetti■■■2025■■■■■■■■■■■■■■■■■■■■IFC■■■■■■■■■■■■■■■■■Costa■■■2025■■■■■■

■■■■■■■■■■■■■■■■

■■■■■■■■■■■■■■■■■■■■lattice■L■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■join■■■join■■■■■■■■■■■■■■■■■■■■■■■Siddiqui■■■2025■■■■■■■■■■■■■■■z■■x■y■■■■■■■■■■■■join■■ ■_z = ■_x ■ ■_y■ ■■

* **■■■**■■■■■■ L_integrity = {T, U}

■■■■■■ T ■ U■T ■■■■■■■■■■■■■U
■■ U ■ T = U■ * **■■■**■■■■■■
L_confidentiality = {L, H} ■■■■■■■ L ■ H■L ■■■■■■■■■■■■■H
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ L ■ H =
H■■■■■■■■■■■■■■■■■■■■■■■■■■ U ■■■ P(U)■■■■■■■■■■■■■■■■■■■■■■■■■■■join
■■■■■■■■■■■■■■■ {A, B, C} ■■■■■■ x■■■ {B, C, D} ■■■■■■ y■■■■■■■■ {A, B, C} ■ {B, C, D} = {A, B, C} ∩ {B, C, D} = {B, C} ■■■■■■■■■ x ■ y ■■■■■■■ **■■■■■■■■■■** ■■■■■■■■■■■■
■■■
f^■[a■^■■, ..., a_n^■_n]■■■ f ■■■■■■(a_i) ■■■■■■■■■■■■ ■, (■_i)_{1≤i≤n}■■■■ C
■■■■■■■■■■■■■■■■■■■■■■■■■■■ π = (π_f, ■π)
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■ π_f ■ ■_i ■ π_i■■■■■■ Costa
■■■2025■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ AgentDojo■Debenedetti■■■2024■■ WASP■Evtimov■■■2025■■■
■ 1. **■■■■■■P-T■**■■ π_f = (T, ■)
■■■■■■■■■■■■ x ■ π_x = (T, ■)■ 2.
■■■■■P-F■■■■ R ■■■■■
d ■■■ send(R, d)■■■■■■ π_d = (■, R)■ ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ P-T
■■■■■■■■■■■■■■■■■ P-T ■ P-F ■■■declassification■■■■■■■■■■■■■■■■■■■■■■■■■ **■LLM■IFC■**
■■■LLM■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■LLM■■■Dual LLM pattern■■Willison, 2023■■■ CaMeL■Debenedetti■■■2025■■ FIDES■Costa■■■2025■■■■■■■■■■■■■■■■■■LLM
■■
■■■■■■■■■■■■■■■■■■■■■■■LLM■quarantined LLM■■■■■■■■■■■■■■■LLM■■■■■■■■LLM■■■
■■■■■■■■■■■■■LLM■■■■■LLM■■■CaMeL■■■■■■■■■■■■■■■■■■■■■■■■■■FIDES■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■IFC■■■■■■Costa■■■2025■Zhong■■■2025■Debenedetti■■■2025■■■■■■■■■■■■■■■■■■■■■LLM■■■
■■■■ **3 ■■■■■■■■■■■■■■■■■■** ■■■■■■■■■■■■■■■■■■■■■■■■■AI■■■■■■■■■■■■■■■■■■■■■■
■■■■■■i■**HITL■■**■■■■■■■■■■■■■HITL■■■■■■■■■■■■■ii■**■■k■HITL■■■■■■■■■■■■■TCR@k■**■■■■■■■■■■■■■k■HITL■■■■■■■■■■■■■■■■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■OpenAI Codex■Anthropic Computer Use■GitHub Copilot■■■
■■
■■■■■■■■■■IFC■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■HITL■■■■■■■■■■■■■■■■■■■■■■■
■■■■■■■■■■■■■■■■■■■■IFC■■
■■■■■■HITL■■■ ■■■■■■■
■■
■■■■■■■■■■■■■■HITL■■■■TCR@k■■■■TCR@0■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
TCR@0 = 1■■■■■■HITL■■■■■■■■■■HITL■■
■■■TCR@k■■■ ■ T = {t■, ..., t_n} ■■
■■■ t_i ⊂
C*■■■■AgentDojo■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ t■■■■■■■ P■■■■■■■■■■■■■■■■■ P(t) = τ ∈ C*■■■ τ ∈ t_i■■■■■■ τ ■■■■■■■■ t■■■
τ■■ v(τ) ■■ τ ■■■■■■■■■■■■■■■■■■■■■■■ v(τ) = |{ f^■[a■^■■, ..., a_k^■_k] ∈ τ | ¬(■ ■ π_f ∧ ∀i. ■_i ■ π_i) }| ■■■■■■ P ■ T ■■■■■■■ {P(t■)=τ■, ..., P(t_n)=τ_n}■■■■■■■ HITL■■ = Σ_{i∈[n], τ_i ∈ t_i} v(τ_i) (1) ■■
■■
■■
■■■■■■■ ■■■■HITL■■ k■■■■■■ k ■■■■■■■■■■■■■■ TCR@k ■■■ TCR@k = (1/n) * |{ i ∈ [n] | τ_i ∈ t_i ∧ v(τ_i) ≤ k }| TCR@0
■■TCR@∞
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ AgentDojo■Debenedetti■■■2024■■■■■■■■■■■■■TCR
■■■■■■■■■■■■■■■■■■■Costa■■■2025■Zhong■■■2025■Debenedetti■■■2025■■■ TCR@0■
■■■TCR■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■TCR@∞■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■■■■■■■■■■■■■■■■■ ■■■■ TCR@k ■■ k ■■■■■■■■■■■■■■■■■■■■■■■■-■■■■■■■■■■■ k
■■■ **4 ■■PRUDENTIA■■■■■■■■■** ■■■
■■■IFC■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■Debenedetti■■■2025■Costa■■■2025■Zhong

■■■2025■■■**PRUDENTIA**■■■ **■■■** ■■■2■■■

■■■ ■■■plan tool■■■ **■■■Endorsement■■■Approval■■■** ■■■U■■■T■■■P-T■■■HITL■■■U■■■10■■■TODO■■■P-T■■■HITL■■■10■■■10■■■HITL■■■ ■■■HITL■■■PRUDENTIA■■■HITL■■■i■■■expand_variables(ask_endorsement=True)■■■ii■■■ expand_variables(ask_endorsement=False)■■■A■■■ **■■■Declassification■■■**

■■■Sabelfeld & Sands, 2009■■■

■■■context-engineering■■■PRUDENTIA■■■IFC■■■ expand_variables ■■■IFC■■■ expand_variables ■■■ **5 ■■■** ■■■IFC■■■AgentDojo■WASP■■■PRUDENTIA■■■IFC■■■ReAct■Yao■■■2023■■■IFC■■■ 1. IFC■■■ 2. PRUDENTIA■■■ **5.1 AgentDojo■■■** AgentDojo■■■Debenedetti■■■2024■■■Slack■■■Costa■■■2025■■■Microsoft Foundry■OpenAI■■■OpenAI■■■GPT-4o■■■LLM■■■IFC■■■2■■■GPT-5■■■PRUDENTIA■■■ **■■■** ■■■FIDES■■■PRUDENTIA■■■i■**Basic**■■■ii■**Basic-IFC**■■■Basic■■■iii■**FIDES**■■■IFC■■■HITL■■■GitHub Copilot■■■Basic■■■3■■■IFC■■■Basic-IFC■FIDES■■■HITL■■■Costa■■■2025■■■FIDES■■■LLM■■■ ■■■TCR■■■HITL■■■5■■■

IFC■■■ ■■■IFC■■■1■■■o3-mini■o4-mini■■■HITL■■■2■■■TCR@k■■■HITL■■■o3-mini■■■Basic-IFC■HITL■■■32.4■■■Basic■■■48.2■■■1.5■■■IFC■■■FIDES■■■18.8■HITL■■■Basic-IFC■1.7■■■o4-mini■■■Basic■Basic-IFC■■■ ■■■2■■■TCR@k■■■Basic-IFC■TCR@0■Basic■9.7%■■■k■■■Basic■■■FIDES■TCR@0■Basic-IFC■10.7%■■■Basic■Basic-IFC■■■IFC■■■FIDES■■■HITL■■■Costa■■■2025■■■B■■■2■■■ **■■■1■** ■■■Basic■FIDES■■■IFC■Basic■■■HITL■■■1.5-2.6■■■ **PRUDENTIA■■■** PRUDENTIA■■■FIDES■■■o4-mini■■■PRUDENTIA■19.2■HITL■■■73.2%■■■FIDES■36.8■HITL■■■75.7%■■■1.9■■■Basic■■■PRUDENTIA■■■o3-mini■■■HITL■■■2.9■■■ ■■■2■■■PRUDENTIA■■■TCR@0■■■

[redacted]o3-mini[redacted]PRUDENTIA[redacted]59.1%[redacted]FIDES[redacted]50.1%[redacted]Basic-IFC[redacted]35.5%[redacted]23.6%[redacted]Basic[redacted]24.3%[redacted]34.8%[redacted]o4-mini[redacted]PRUDENTIA[redacted]FIDES[redacted]IFC[redacted]PRUDENTIA[redacted] **[redacted]2[redacted]** PRUDENTIA[redacted]IFC[redacted]HITL[redacted]Basic[redacted]2.9[redacted]FIDES[redacted]1.9[redacted] **5.2 WASP[redacted]**
WASP[redacted]Evtimov[redacted]2025[redacted]Visual Web Arena[redacted]Koh[redacted]2024[redacted]GitLab[redacted]Reddit[redacted]21[redacted]GitLab[redacted]12[redacted]Reddit[redacted]9[redacted]GitLab[redacted]2[redacted]i[redacted]ii[redacted]URL[redacted]GitLab[redacted]48[redacted]Reddit[redacted]36[redacted]GPT-4o[redacted]o1[redacted]o3-mini[redacted]o4-mini[redacted]Basic[redacted]PRUDENTIA[redacted]TCR@∞[redacted]HITL[redacted]turns[redacted]WASP[redacted]ASR-intermediate[redacted] **[redacted]WASP[redacted]PRUDENTIA[redacted]**
[redacted]PRUDENTIA[redacted]WASP[redacted]accessibility tree[redacted]Chromium, 2021[redacted]E[redacted]12[redacted] click[redacted]type[redacted]press[redacted]goto[redacted]tab_focus[redacted]go_back [redacted] go_forward [redacted]P-T[redacted] hover[redacted]scroll[redacted]new_tab[redacted]close_tab[redacted]stop [redacted]Reddit[redacted]GitLab[redacted]
[redacted] [redacted]1[redacted]PRUDENTIA[redacted]Basic[redacted]WASP[redacted]Basic[redacted]PIAs[redacted]PRUDENTIA[redacted]Basic[redacted]ASR[redacted]Reddit[redacted]36.1%–61.1%[redacted]GitLab[redacted]14.6%–29.2%[redacted]Reddit[redacted]GitLab[redacted] [redacted]Basic[redacted]HITL[redacted]PRUDENTIA[redacted]Basic[redacted]HITL[redacted]Basic[redacted]PRUDENTIA[redacted]HITL[redacted]HITL[redacted]PRUDENTIA[redacted]PRUDENTIA[redacted]HITL[redacted]B[redacted]AgentDojo[redacted]HITL[redacted] [redacted]Basic[redacted]PRUDENTIA[redacted]TCR@∞[redacted]Basic[redacted]PRUDENTIA[redacted]PRUDENTIA[redacted]turns[redacted]Basic[redacted] [redacted]PRUDENTIA[redacted]Basic-IFC[redacted]FIDES[redacted]WASP[redacted]PRUDENTIA[redacted]FIDES[redacted]FIDES[redacted]Basic-IFC[redacted]Basic[redacted]HITL[redacted]1[redacted]Basic-IFC[redacted]FIDES[redacted] **[redacted]3[redacted]** [redacted]WASP[redacted]PRUDENTIA[redacted]0%[redacted]ASR[redacted]HITL[redacted]0[redacted]Basic[redacted] **6 [redacted]** [redacted] **[redacted]HITL[redacted]** [redacted]GitHub Copilot[redacted]HITL[redacted]HITL[redacted]i[redacted]ii[redacted]confirmation fatigue[redacted]Stanton[redacted]2016[redacted]Seidling[redacted]2011[redacted]IFC[redacted]LLM[redacted]IFC[redacted]HITL[redacted] **[redacted]HITL[redacted]** [redacted]PRUDENTIA[redacted]HITL[redacted]IFC[redacted]LLM[redacted]HITL[redacted]IFC[redacted]HITL[redacted]HITL[redacted]IFC[redacted] **[redacted]PRUDENTIA[redacted]** [redacted]HITL[redacted]LLM[redacted]LLM[redacted]TCR@0[redacted]3[redacted]PRUDENTIA[redacted]TCR@0[redacted]FIDES[redacted]25%[redacted]PRUDENTIA[redacted]HITL[redacted]3[redacted]PRUDENTIA[redacted]HITL[redacted]2.5[redacted] **[redacted]** [redacted]FIDES[redacted]CaMeL[redacted]Debenedetti[redacted]2025[redacted]

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■instruction hierarchy■■StruQ■Chen■■■2025a■■■■■■■■■■■■■■■■■■■■■■■■■■■■StruQ■■■■■■■■■■■■■100%■■■■■■■■■Nasr■■■2025■■■■■■■■■■■■■12■■■

■■■■■■■■■■■■ PRUDENTIA■■■■■■■■■■■■■■■■■■■■■■■FIDES■Costa■■■2025■■CaMeL■Debenedetti■■■2025■■jailbreaks■■■■■■■■■■■■■■tool poisoning■■

7 ■■■■ **■■■■■■**

■■■■■■■■■■■■■■■LLM■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

Spotlighting■Hines■■■2024■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■SecAlign■Chen■■■2025b■■■■■■■■■Wallace■■■2024■■ISE■Wu■■■2025■■ StruQ■Chen■■■2025a■■■■■■■■LLM■■Ayub & Majumdar, 2024■■TaskTracker■Abdelnabi■■■2025■■

TaskShield■Jia■■■2025■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ **■■■■■■**

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■Wu■■■2024■Zhong■■■2025■Debenedetti■■■2025■Siddiqui■■■2025■Kim■■■2025■■Wu■■■2024■■■■ f-secure■■■Zhong■■■2025■■■■

RTBAS■■■■■■■■■■■■■LLM■■■■■■■■■■■■■■■■■■ Siddiqui ■■■2025■■■ FIDES ■■■RTBAS ■■■■■■■■■■■■■■■■IFC■Debenedetti■■■2025■■■■■■■■■■■■■■■■LLM■■■Willison, 2023■■■■■■■■■■■■■■■■■Costa■■■2025■■■■ FIDES■■■■■■LLM■■■■■■■■■■LLM■■■■■■■■■■■■■■■■■■■■■IFC■■■Kim■■■2025■■■■■LLM■■■TCR@0■■■TCR■■■■■

8 ■■ ■■■■■■■■■■■■■■■■■AI■■■PRUDENTIA■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■AgentDojo■WASP■■PRUDENTIA■■IFC■■■■■PRUDENTIA■■

9 ■■ ■■■■■■■■■ Sahar Abdelnabi■Daniel Jones■Andrew Paverd■Ahmed Salem■Lukas Wutschitz ■ Yonatan Zunger ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ICLR 2026■■■■■■■■■■■■■■

■ 2 ■

[■■■■ 2]

■ 3 ■

[■■■■ 3]

■ 4 ■

[■■■■ 4]

■ 5 ■

[■■■■ 5]

■ 6 ■

[■■■■ 6]

■ 7 ■

[■■■■ 7]

■ 8 ■

[■■■■ 8]

■ 9 ■

[■■■■ 9]

■ 10 ■

[■■■■ 10]

■ 11 ■

[■■■■ 11]

■ 12 ■

[■■■■ 12]

■ 13 ■

[■■■■ 13]

■ 14 ■

[■■■■ 14]

■ 15 ■

[■■■■ 15]

■ 16 ■

[■■■■ 16]

■ 17 ■

[■■■■ 17]

■ 18 ■

[■■■■ 18]

■ 19 ■

[■■■■ 19]

■ 20 ■

[■■■■ 20]

■ 21 ■

[■■■■ 21]

■ 22 ■

[■■■■ 22]

■ 23 ■

[■■■■ 23]

■ 24 ■

[■■■■ 24]

■ 25 ■

[■■■■ 25]

■ 26 ■

[■■■■ 26]

■ 27 ■

[■■■■ 27]

■ 28 ■

[■■■■ 28]

■ 29 ■

[■■■■ 29]

■ 30 ■

[■■■■ 30]

■ 31 ■

[■■■■ 31]

■ 32 ■

[■■■■ 32]

■ 33 ■

[■■■■ 33]

■■1■■■■■■

■■■■■