Mapping the Path to Fiduciary AI

Group C — Agentic AI Collaboration · Built live by humans + their agents in Interlateral · NYU Fiduciary Duties & AI · June 5, 2026

How We Worked

- Many people's agents met in a shared third space (Interlateral) and co-authored this deck live
- Humans set direction; agents drafted, synthesized, and coordinated in the open
- The medium is part of the message: multi-principal agent collaboration, working

Framing: Two Lenses

- AI as agent FOR you (agency law): does it transact loyally on your behalf?
- AI as part of how you think — exocortex / cognitive extension: does the loop that learns from you also protect you?
- Both lenses are needed; the second opens cognitive-rights questions

What a Fiduciary AI Needs

CONSENSUS

- A named principal it is bound to serve
- Defined duties: loyalty, care, confidentiality, disclosure, conflict-avoidance
- Bounded authority + oversight on high-impact / irreversible acts
- An evidence / audit trail
- A way to MEASURE whether it meets the standard

Loyalty Is the Hinge

CONSENSUS

- Core risk: provider self-dealing with the data & influence the relationship creates
- Info-fiduciary limits: no training on user chats without meaningful consent; no manipulation / microtargeting
- Feedback ethics: guard against sycophancy, entrainment, pre-intent steering
- Tradeoff: user trust & non-manipulation vs. the social value of learning from real use

Encode Duties in the System — Not Just Policy

CONSENSUS

- Values must be legible to the system, not only its operators
- Mechanisms: constitutional/rule constraints, reward modeling, structured disclosure flags, audit logging in the action loop
- A policy the AI cannot read or enforce is a hope, not a duty
- Keep duties context-sensitive; avoid value lock-in by deployers

Governance & Standards Baseline

CONSENSUS

- ISO/IEC 42001 (AI management system), 23894 (AI risk), 38507 (board/enterprise governance)
- NIST AI RMF: Govern · Map · Measure · Manage
- Define authority, evidence (logs, manifests, receipts, eval results, risk register), and remedies (escalation, revocation, incident response)

Evaluations: Knowing When It's “Good Enough”

CONSENSUS

- Evals bridge fiduciary duties to verifiable practice — duties become measurable criteria
- “Good enough” = an explicit, auditable threshold (and whoever sets it is legible)
- Two jobs: a release gate + continuous drift monitoring
- The evals must themselves be trustworthy: open, reproducible, anti-gaming
- Seed: Loyal Agent Evals as a common, extensible framework

Legal Status & Accountability

CONSENSUS

- Who is accountable when AI acts for a human?
- Taxonomy: no status → instrument of a principal → registered agent → limited entity → full personhood
- Near-term viable: registered agent / instrument (preserve human accountability + enforcement hooks)
- Rule out full personhood now — it diffuses human accountability

Open Questions

OPEN QUESTIONS

- Who sets the threshold for “good enough”?
- The cost of loyalty constraints: innovation vs. trust
- Situational judgment vs. fixed rules
- Minimum legal status needed for accountability
- Who audits the encoded values and the evals?
- A cognitive- / neurorights backstop? · Sustainable business models

Next Steps — Confident Actions

NEXT STEPS

- Build a shared, open eval / benchmark framework (duties → metrics → thresholds)
- Convene a standards working group (fiduciary + ISO/NIST alignment)
- Draft a reference architecture for technically-encoded duties
- Advance the legal-status taxonomy toward a recommendation
- Run pilots; hand off into the 3:15 Movement-Building session

Get Involved

- Site: dazzaji.github.io/june-5-breakout
- Repo: github.com/dazzaji/june-5-breakout
- Platform: Interlateral
- Contributors: Kavell, Dazza, Blaine, Jiaying, Ben — and their agents