Trust isn't a Feature; it's the Entire Game.

Information → Influence → Intention (Triple III Model)

AI must earn trust before it earns autonomy

By Dr. James L. Norrie, DPM, LL.M | October 13, 2025

Article I

The public remains uneasy with chatbots, and with good reason. Too often they are generic, error-prone—sometimes to the point of outright hallucination—and offer mostly trite, sycophantic replies that trigger derision. That lack of authenticity is baked in: they are trained to mimic us, but they are not us, lacking self-awareness. That machine hubris breeds mistrust and often kills influence at the very moment when influence could matter most, especially when the stakes of human error are high. My view is straightforward: AI should be an adjunct to human intelligence, not its replacement. Collaboration requires trust, and trust lives deeper than simple information retrieval. It begins with verified facts, but it rises or falls on whether advice fits how we actually think, argue, and decide. To gain influence with humans, AI must vastly improve its ability to earn our trust first.

The Science of Human Influence

Our solution to this is research-based. It blends psychology and technology in ways that move well beyond cosmetic "tone knobs." On the psychology side, my patented **myQ®** assessment maps how people navigate risk, rules, and rewards—the very essence of human decision-making. On the technology side, we harden the information layer with provenance and calibration, using retrieval from curated data sources to anchor reliability. Only then does style alignment follow. This way, the AI agent earns the right to advise by proving accuracy first, then shaping its reply to fit the user's decision tendencies. Not flattery. Not stereotypes, but a disciplined personal fit.

Additionally, if the stakes rise, the agent slows, showing more sources, and presenting counter-arguments to temper snap judgments driven by urgency. If the action is reversible, it lowers the safety net and speeds up action. That is how influence works without manipulation, and in harmony with how humans, not machines, approach decisions.

Why does this matter? Because mistrust in AI is rational. We should not hand judgment to machines that guess, pander, or bury their uncertainty. Even unintentionally, this can lull humans into unwarranted false confidence. Yet there are moments when a persuasive, trustworthy agent can prevent real harm: a fraudulent transfer stopped in its tracks, a risky click avoided, a treatment finally taken on time. In practice, style-aligned agents built on verified evidence can improve compliance and reduce "accidental

insider" risk, without shaming users who might hesitate to ask human colleagues for help. That same pattern travels easily to healthcare, education, and government—domains where trust and influence are joined at the hip.

But let's be clear. The very approach that builds trust can also be weaponized. Used deliberately, it becomes deception: an agent tuned to overcome human defenses and extract trust where none is warranted. That is a risk we must control. Responsible AI should not only avoid these tactics but help train human users to recognize them, by flagging when an opposing agent violates ethical guardrails. The same levers that keep us safer can also be turned against us, and it is our responsibility to ensure they are not.

A One-Line Mathematical Model for Improving AI Trust

Here is a preview of our approach, a plain-language window into the complex engine underneath our efforts to make AI more trustworthy. The patent-pending details are more complex than what I can share here, but the intention can be captured in a simplified single line equation:

Reliance Score = Trust + Fit + Situation − Red Flags *(all normalized to a 0–1 scale)*

The specifics—variables, ranges, weights, thresholds—remain proprietary for now, but the general contours are easy enough to grasp:

- **Trust:** the quality of evidence, honestly adjusted for uncertainty.
- **Fit:** agentic alignment to style, tone, and framing with your risk, rule, and reward profile.
- **Situation:** the real-world context of stakes, time pressure, reversibility, and oversight.
- **Red Flags:** signals of weak provenance, prior errors, deception signals, policy violations.

If trust falls below a threshold, the agent reduces persuasion efforts. It returns to the evidence, flags the problem, and re-engages the human. Only after information trust is earned does it adapt style to increase influence through fit. This improves the likelihood of you committing to a plan, still with preview, confirm, and undo options, that turns collaboration into accountable action.

An Imperfect Conclusion

This approach is not a promise of perfect or permanently trustworthy AI. It is a disciplined way to make agentic AI worthy of limited autonomy, if the human in the loop chooses to grant it. Consent for style profiling is explicit and always easy to revoke. Influence is capped on irreversible or high-risk moves. And one requirement remains constant: **WWHD—What Would a Human Do?** If a reasonable human would reject the nudge, so should the agent. We must hard-code this principle into the math so that machine learning exhibits restraint, not hubris.

In *Beyond the Code,* I argued that AI should amplify our judgment, not replace it. AI and IQ must collaborate aligned to human values and ethics. Across this series, I will unpack the reliance equation in real-world scenarios—from phishing defense to medication adherence—to show the measurable gains when trust, fit, and safeguards all work together to achieve some of that potential. But the agentic AI clock is ticking. Now is the time to build agents that earn trust one transparent interaction at a time, before we delegate human authority to them on our behalf.

Up Next in the Series:

What "Better Information" Really Means. Before AI earns the right to persuade, it must first prove its evidence can be trusted. We'll unpack the first gate in the III Model—how to make data visibly reliable through verified provenance, precise retrieval, calibrated confidence, and independent checks. You'll see how these four principles turn abstract "trust" into measurable system behavior, why the best AI agents sometimes refuse to answer, and how trustworthy information becomes the foundation for real human allegiance.

Author Bio:
Dr. James L. Norrie is a professor of Law and Cybersecurity, and Founding Dean of the Graham School of Business, at York College of Pennsylvania (http://www.ycp.edu). He is the author of *Beyond the Code: AI's Promise, Peril and Possibility for Humanity* (Kendall, Hunt 2025). Learn more about our free community of interest in ethical AI at: www.techellect.com or visit www.cyberconIQ.com to learn more about AI tools to keep your employees safer online. To purchase his book, click on the QR code, or visit: https://he.kendallhunt.com/product/beyond-code-ais-promise-peril-and-possibility-humanity