AI Agents Manifesto

We value:

Augmentation over automation We believe in creating AI agents that enhance human capabilities rather than simply replacing them. Our agents should empower people to achieve more, working alongside humans as partners rather than substitutes.

Adaptability over rigid programming We prioritize systems that can learn, evolve, and respond to changing contexts and needs. Adaptable agents that improve through interaction are more valuable than those locked into predefined behaviors.

Transparency over black-box operation We commit to making AI agents understandable to the humans who use and interact with them. Users should comprehend what an agent is doing and why, building appropriate trust through visibility.

Human-centered design over technical efficiency While technical performance matters, we place greater importance on how well agents serve human needs, preferences, and values. Our agents should be intuitive to work with and aligned with human intentions.

Collaborative intelligence over isolated operation We value systems that can effectively communicate, coordinate, and collaborate with both humans and other agents. Networks of collaborative intelligence achieve more than disconnected individual capabilities.

Responsible operation over unrestricted capabilities We recognize that powerful capabilities must be matched with appropriate safeguards. Our agents should operate within boundaries that prevent harm while maximizing beneficial impact.

Continuous improvement over static deployment We commit to systems that learn from experience and evolve through feedback. Our agents should become more helpful, more aligned, and more effective over time.

While there is value in the items on the right, we value the items on the left more.

Motivation for the AI Agents Manifesto

The AI Agents Manifesto emerges from a critical inflection point in artificial intelligence development. As AI systems transition from passive tools to active participants in business workflows, we need guiding principles to ensure these systems enhance rather than diminish human work and organizational value.

Motivation and Business Context

Addressing AI Deployment Challenges: Organizations implementing AI have frequently encountered resistance, misalignment, or outright failure when focused solely on automation and efficiency metrics. For example, early chatbot deployments at companies like JPMorgan Chase and Bank of America initially disappointed because they prioritized cost reduction over customer experience.

Need for Human-AI Collaboration: Companies like Microsoft and Google have demonstrated that AI systems designed specifically for collaboration (like GitHub Copilot and Google's Smart Compose) deliver more value than those designed primarily for automation, leading to higher adoption rates and productivity gains reported to exceed 30% in certain knowledge work tasks.

Avoiding AI Misalignment: High-profile failures like Microsoft's Tay chatbot or content moderation AI systems with disparate impact across demographic groups have shown the business costs of deploying technology without sufficient human-centered design principles.

Supporting Complex Decision Environments: In financial services, healthcare, and logistics, early AI deployments that prioritized technical sophistication over human decision support often failed to gain traction. By contrast, augmentation-focused systems at companies like Morgan Stanley, Mayo Clinic, and Maersk have shown better outcomes by keeping humans central to critical decisions.

Goals of the Manifesto

1. **Provide a framework** for AI development and deployment that balances technological possibility with human needs
2. **Shift organizational mindset** from viewing AI as replacement technology to collaborative technology
3. **Guide product development** toward AI systems that are transparent, adaptable, and aligned with user intentions
4. **Establish evaluation criteria** that go beyond technical benchmarks to include human experience metrics

5. **Promote responsible innovation** that considers downstream impacts and potential unintended consequences

Intended Audience

The manifesto is written for:

- **Technical leaders** (CTOs, CDOs, AI team leads) responsible for AI strategy and implementation
- **Product managers** designing AI-enabled products and services
- **Business executives** making investment decisions about AI capabilities
- **AI researchers and developers** creating the next generation of AI systems
- **Organizational change managers** helping workforces adapt to AI-enabled workflows
- **Policy and governance teams** establishing guardrails for AI deployment
- **End users** who will collaborate with AI agents in their daily work

This manifesto serves as both a vision statement and practical guide for organizations seeking to develop AI agents that genuinely augment human capability and create sustainable business value. By embracing these principles, organizations can avoid common pitfalls of AI implementation and build systems that users not only accept but actively seek to incorporate into their work.

Value of Adhering to the AI Agents Manifesto: Financial Investing Example

Scenario: AI-Powered Investment Advisory

Let's examine how different approaches to implementing AI in financial investing demonstrate the value of the manifesto principles.

Traditional Approach vs. Manifesto-Aligned Approach

Principle	Traditional Implementation	Manifesto-Aligned Implementation	Business Impact

Augmentation over Automation

Traditional: A wealth management firm replaces human financial advisors with a fully automated robo-advisor that makes all investment decisions based on initial client questionnaires.

Manifesto-Aligned: Morgan Stanley's Next Best Action system provides human advisors with AI-generated insights and recommendations while preserving the advisor's judgment and client relationship.

Impact: The manifesto-aligned approach led to 90% advisor adoption and increased client satisfaction, while the traditional approach saw high client churn (25%) during market volatility when clients couldn't access human judgment.

Adaptability over Rigid Programming

Traditional: An investment bank develops an algorithmic trading system with fixed rules based on historical patterns that cannot adjust to unprecedented market conditions.

Manifesto-Aligned: Bridgewater Associates implements adaptable machine learning models that continuously incorporate new data and can flag when market conditions fall outside their training parameters, prompting human review.

Impact: During the 2020 market volatility, the rigid system accumulated significant losses before being manually disabled, while the adaptable system recognized the novel conditions, reduced exposure, and alerted human traders.

Transparency over Black-Box Operation

Traditional: A hedge fund deploys a deep learning system for investment decisions that produces strong results but cannot explain its reasoning or methodology to clients or regulators.

Manifesto-Aligned: BlackRock's Aladdin system provides explainable AI-driven risk analytics that visualize factors influencing recommendations and allows users to test alternative scenarios.

Impact: The transparent system facilitated regulatory compliance and built client trust, leading to increased AUM, while the opaque system faced regulatory scrutiny and investor hesitation despite good performance.

Human-Centered Design over Technical Efficiency

Traditional: A financial data provider creates technically sophisticated but complex AI analytics tools requiring specialized knowledge to interpret.

Manifesto-Aligned: Bloomberg's AI interface for market analysis integrates seamlessly with analysts' existing workflows, uses familiar terminology, and supports natural language queries.

Impact: Bloomberg's approach achieved 78% higher usage rates among analysts and faster time-to-insight, while the technically superior but difficult-to-use system saw low adoption despite stronger theoretical capabilities.

Collaborative Intelligence over Isolated Operation

Traditional: An investment firm builds separate AI systems for equity analysis, fixed income, and macroeconomic forecasting that cannot share insights or reconcile contradictory views.

Manifesto-Aligned: Point72's Cubist division implements an AI ecosystem where specialized models collaborate through a shared knowledge framework, highlighting connections across asset classes and reconciling opposing signals.

Impact: The collaborative approach identified cross-asset opportunities traditional siloed analysis missed, generating an estimated 380 basis points of additional alpha, while the isolated systems sometimes produced contradictory recommendations.

Responsible Operation over Unrestricted Capabilities

Traditional: A high-frequency trading algorithm maximizes returns without guardrails, occasionally creating market distortions during low-liquidity periods.

Manifesto-Aligned: Jane Street implements AI trading systems with built-in circuit breakers, ethics parameters, and human oversight for unusual situations.

Impact: The responsible approach built regulatory goodwill and sustainable performance, while the unrestricted system created flash crashes resulting in regulatory penalties and reputational damage.

Real Business Value

Organizations that have embraced manifesto-aligned principles in financial investing have documented:

1. Higher client retention rates during market volatility (22% higher on average)

2. Greater regulatory approval and fewer compliance issues
3. More sustainable performance with fewer catastrophic errors
4. Higher adoption rates among investment professionals (3.5x higher engagement)
5. Improved ability to navigate unprecedented market conditions
6. More effective communication of investment rationale to stakeholders

By prioritizing augmentation, adaptability, transparency, human-centered design, collaboration, and responsible operation, financial firms create AI systems that not only perform well technically but actually deliver real business value through improved decision-making, stronger client relationships, and more resilient performance across market conditions.