"Prototype to Production" : main points

1. **The "Last Mile" Production Gap and AgentOps:** Building an AI agent prototype is easy, but making it a trusted, production-grade system ("the last mile") is where 80% of the effort is spent. This requires a new operational discipline called **AgentOps**.
2. **Unique Challenges of Agentic Systems:** Unlike traditional software, agents are autonomously interactive, stateful, and follow dynamic execution paths, creating unique operational challenges related to dynamic tool orchestration, scalable state management, and unpredictable cost/latency.
3. **Foundation for Production:** Successfully moving to production is built on three pillars: Automated Evaluation, Automated Deployment (CI/CD), and Comprehensive Observability.
4. **People and Process:** Effective AgentOps requires a well-orchestrated team of specialists, including **AI Engineers**, **Prompt Engineers**, and traditional MLOps/DevOps roles, working together across the organization.
5. **Pre-Production Strategy: Evaluation-Gated Deployment:** The core principle for trustworthiness is that no agent version reaches users without first passing a comprehensive evaluation that proves its quality and safety. This can be enforced manually (Pre-PR Evaluation) or automatically (In-Pipeline Gate) within the CI/CD pipeline.
6. **The Automated CI/CD Pipeline:** A robust CI/CD pipeline is crucial for managing the complexity of composite systems (code, prompts, tools). It operates in three phases:
    - **Phase 1: Pre-Merge Integration (CI):** Fast checks and agent quality evaluation (shifting left).
    - **Phase 2: Post-Merge Validation in Staging (CD):** Comprehensive integration tests and load testing in a production-like replica.
    - **Phase 3: Gated Deployment to Production:** Final sign-off and safe rollout (e.g., Canary, Blue-Green).
7. **Building Security from the Start:** Agents' autonomy introduces unique risks *(Prompt Injection, Data Leakage)*. Security requires a comprehensive strategy with *three layers of defense: Policy Definition (System Instructions), Guardrails (Input/Output Filtering, HITL), and Continuous Assurance (Rigorous Evaluation, Red Teaming).*
8. **Operations In-Production: The Observe → Act → Evolve Loop:**
    - **Observe:** Gaining insight into an autonomous agent's behavior through **Logs**, **Traces**, and **Metrics** (e.g., using Google Cloud operations suite).
    - **Act:** Real-time intervention to manage system health (scaling, cost, latency) and risk (Security Response Playbook with a "circuit breaker").
    - **Evolve:** Strategic, long-term improvement by turning production data and failures into new test cases and deploying fixes rapidly via the automated CI/CD pipeline.
9. **Beyond Single-Agent Operations: Interoperability:** To scale to multiple agents, organizations need standardized protocols for collaboration.
    - **Agent2Agent (A2A) Protocol:** Designed for complex, stateful collaboration and delegating high-level goals between intelligent agents, using **Agent Cards** for discovery.
    - **Model Context Protocol (MCP):** Complementary to A2A, used for simple, stateless interactions with tools (primitives with structured I/O).
10. **Registry Architectures:** For very large ecosystems, **Tool Registries** (using MCP) and **Agent Registries** (using AgentCards) help with discovery, reuse, and centralized governance.

Sources:

- Prototype to Production.pdf