# Основной процесс разработки AI/ML сервисов Yandex Cloud

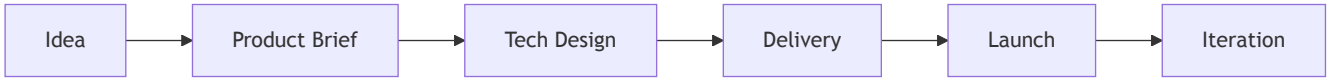
Обновлено 1 ноября 2025, 18:22 Сергей Золотов

**Владельцы (A):** noxoomo
**Соавторы (R/C):** hotckiss, eranik, sergeyzolotov

# Executive Summary

Этот процесс предназначен для разработки AI Studio, а также любых SaaS продуктов в зоне ответственности AI/ML сервисов Yandex Cloud.

Он соединяет продуктовую и техническую проработку через понятные вехи (milestones), чтобы каждая идея проходила одинаковую воронку:




Решение принимается на **Product Review** (ценность) и **Architecture Review** (реализация).

- **Product Review** решает «зачем» и «что». При необходимости на Product Review приглашаются технические коллеги для совместной проработки.
- **Architecture Review** решает «как» и «насколько надёжно».
- Все решения фиксируются (approve / reject / pivot / defer).
- **Product Brief** описывает цели, метрики, ресурсную модель, экономику.
- **Tech Design Doc** описывает архитектуру, Cloud/On-prem, Data/Control Plane, эксплуатацию, Change Management.
- R&D отделён от production: только прототипы без SLA, которые при успехе идут в Product Brief/Tech Design Doc.
- Capacity распределяется по бакетам (Security, Must-do, Сквозные, R&D Product, R&D Tech).
- ML-фичи развиваются итеративно (feedback → retrain → rollout).
- Lean startup принципы: быстрые артефакты, эксперименты, fail fast.

Ключевая ценность: **процесс создаёт прозрачность и баланс между продуктовой скоростью и технической надёжностью**, снижает риски хаоса и выгорания команд, даёт управляемость портфелю инициатив.

# Основные процессы

- **Product Review** — зачем и что делаем (ценность, приоритет, RICE). Решение: go/no-go на основе **Product Brief**.
- **Architecture Review** — как именно делаем (архитектура, NFR, безопасность, эксплуатация). Решение: go/no-go на **Tech Design Doc**.
- **Tech Proposal** — обсуждаем breaking/совместимость и интерфейсы ДО разработки; окно ревью 7–14 дней; обязательны sign-off'ы.
- **Decision Note** — фиксируем принятое архитектурное решение и trade-offs.

## Облачный процесс запуска фичей и сервисов

**Tech Preview → Preview → GA**

- **Tech Preview:** ранняя версия для тестирования концепции, без SLA
- **Preview:** функционал готов, идёт стабилизация, ограниченная доступность
- **GA (General Availability):** полная готовность, SLA, поддержка, биллинг

## Security

- **SecReview** — перед началом разработки или при изменении архитектуры
- **SecAudit** — после запуска (до GA) и при существенных модификациях

# How we decide

Решения принимаем на двух основных синках

## Product Review

- Отвечает за продуктовые решения и приоритеты.
- Решает:
  - запуск идеи в Product Brief,
  - включение фичей в продуктовый pipeline и квартальный план,
  - go/no-go на запуск (с бизнесовой стороны).
- Состав: Product Leads, TPMs, представители Marketing, Sales, Finance, Tech команды.
- Главный вопрос: **"Стоит ли делать это с точки зрения ценности для пользователей и бизнеса?"**

## Architecture Review

- Отвечает за технические решения и архитектурное качество.

- Решает:
  - согласование Tech Design Doc,
  - технические trade-offs (tactical vs RnD, Cloud vs On-prem),
  - готовность к production с точки зрения надежности, security, ops.
- Состав: техлиды платформенных команд, представители Security, Infra, Applied ML, Core ML Research.
- Главный вопрос: **"Сможем ли мы это сделать правильно и масштабируемо?"**

### Совместная работа

- Продуктовая и технические команды работают в паре: продукт отвечает за **зачем**, техническая — за **как**.
- В спорных вопросах решения принимаются на уровне CTO/CPO AI/ML сервисов.

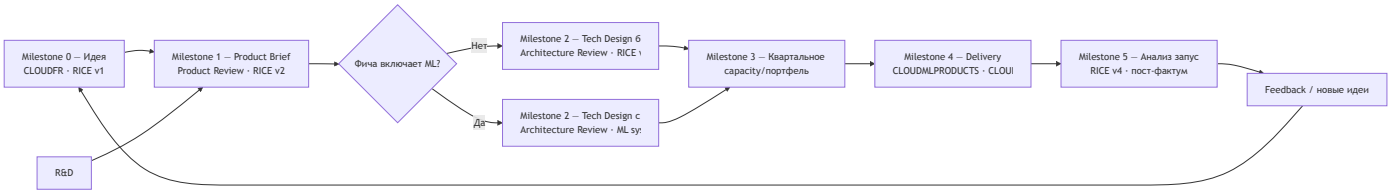
---

# 📌 Процесс работы с фичами

### R&D

Идеи, требующие предварительных исследований, проходят R&D-трек.
Он включает короткие и длинные циклы (ML Research и Product R&D).
Подробно см. https://wiki.yandex-team.ru/cloud/mlcloud/process/rnd-processes/

### Milestones (ML и non-ML)




**Ready Checklist для входа в квартал:** Product Brief + Tech Design Doc (Architecture Review) + актуальный RICE v3 + согласованные зависимости.

---

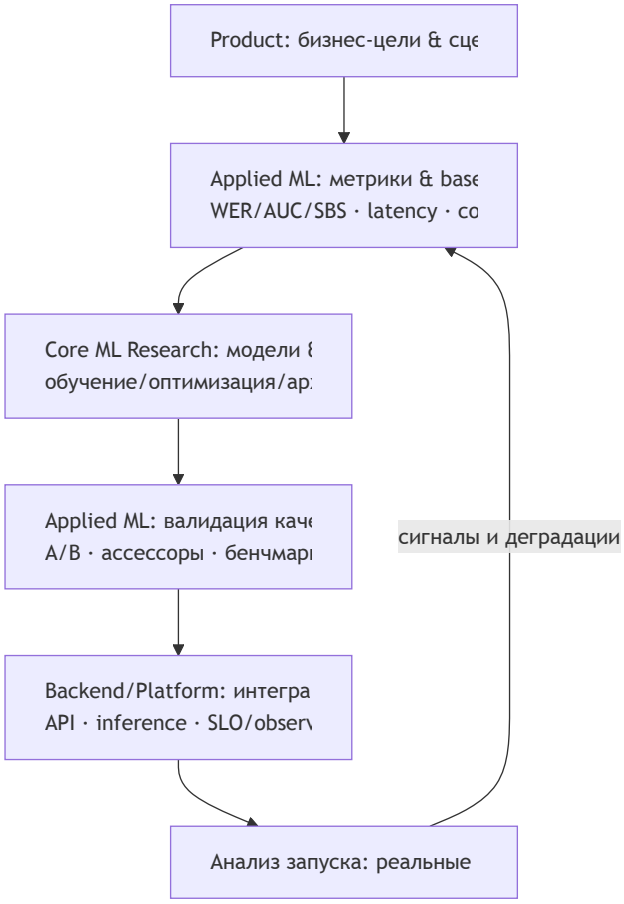
### Post-Launch Review (чеклист)

- Метрики факта: KPI/ML KPI, usage, COGS, SLO/ошибки.
- Выводы: что сработало/что нет; гипотезы и причины.
- Backlog решений: оптимизации, масштабирование, R&D, deprecation/pivot/kill.
- Решение: продолжать/масштабировать/перезаписать план.

Ответственные: Product Lead (A) · TPM (R) · Backend/ML/SRE (C) · Marketing/Sales/Support (C/I).

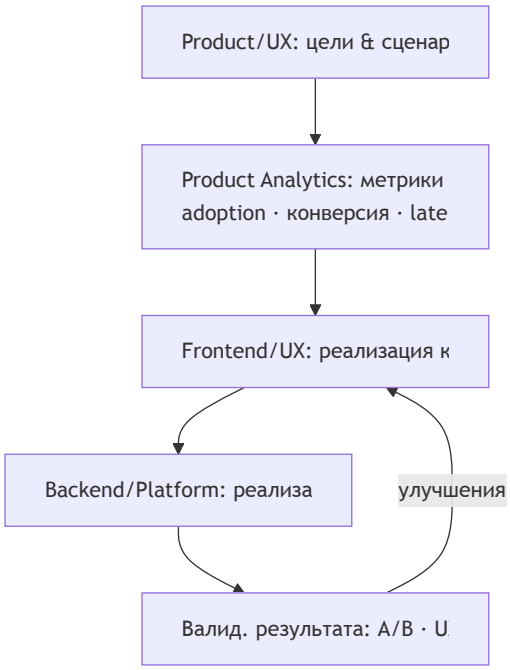
---

### Итеративная ML-петля




---

### Non-ML фичи (упрощённая петля)




---

## Определение RICE (+ Strategic Value)

Формула: `RICE = (Reach × Impact × Confidence) / Effort`
Расширение для платформы: `Platform RICE = RICE × StrategicValue`

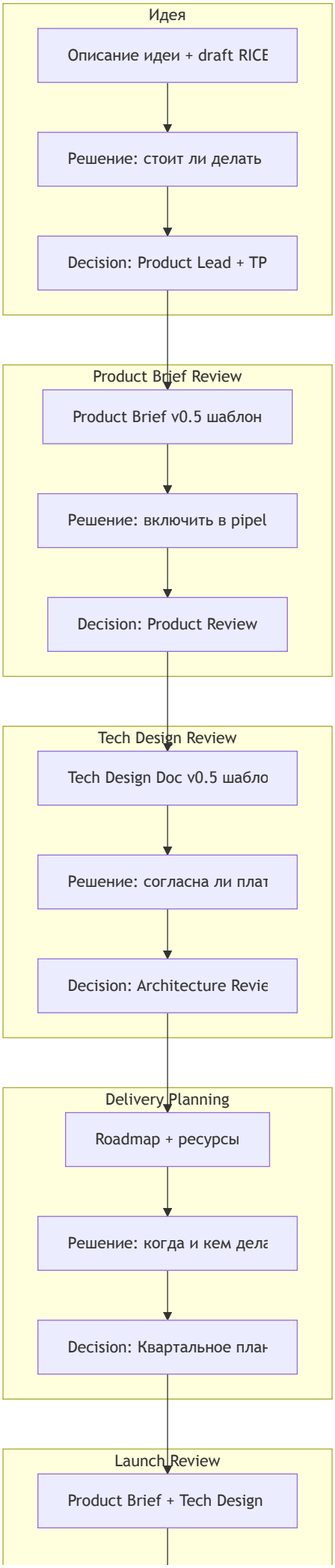
- **Reach**: охват (пользователи, разработчики, API calls).
- **Impact**: эффект (доход, удержание, TTM, GPU consumption).
- **Confidence**: уверенность (0.2–1.0).
- **Effort**: инженерно-человеко-месяцы.
- **Strategic Value (SV)**: мультипликатор (0.8–1.5) — насколько фича открывает новые продукты/классы задач.

## RACI по ключевым стадиям

| Этап / Артефакт | Product | Product Analytics | Applied ML | Core ML Research | Backend / Platform | Frontend | Designer | M |
|---|---|---|---|---|---|---|---|---|
| **Milestone 0: Идея** | **R/A** | **C** | C (для ML) | C (для ML) | C | C | C | C |
| **Milestone 1: Product Brief** | **R/A** | **R** | C (ML KPI/ метрики) | C (feasibility) | C | C | R (UX прототипы) | C |
| **Milestone 2: Tech Design + Architecture Review** | A | C | **R** (ML system design) | **R** (PoC, модели) | **R** (API/ архитектура/NFR) | R (клиентские части) | C (UX макеты) | I |
| **Milestone 3: Планирование** | **R/A** | R | C | C | C | C | C | C |
| **Milestone 4: Delivery** | A | C | **R** (валидация ML) | **R** (обучение /релиз моделей) | **R** (интеграция) | **R** (UI) | **R** (UX тесты) | R |
| **Milestone 5: Анализ запуска** | **R/A** | **R** | **R** (качество/ деградации) | C | C | C | C | C |

## Команды и зоны ответственности

- **Product Team** — владелец ценности, формулирует цели и KPI. Команда Артура.
- **Product Analytics** — данные, кастдевы, конкурентка, KPI. Команда Никиты.
- **ML Analytics** — метрики ML, разметка, оценка качества, бенчмарки. Команда Саши Кондратенко.
- **ML R&D** — разработка/оптимизация моделей, встраиваемых в продукты. Команда Антона.
- **Backend / Platform** — API, интеграция, observability, SLO. Команды Андрея Кириленко, Андрея Путрина, Вити.
- **Frontend** — UI, SDK, клиентские библиотеки. Команда Жени.
- **TPM** — процесс, планирование, capacity, соблюдение процесса работы. Команда Сережи Золотова.
- **Designer** — UX-прототипы, макеты, usability-тесты. Дима.
- **Marketing** — go-to-market, позиционирование, коммуникации.
- **Sales** — сигналы клиентов, пилоты, фидбек.
- **Support** — входящие запросы, проблемы клиентов, обратная связь.
- **Docs** — документация SDK, API, гайды.
- **Legal** — проверка на соответствие законам, лицензионные риски.
- **Finance** — бюджеты, cost-эффективность, модели монетизации.
- **Security** — требования безопасности, приватность, комплаенс.


Идея
Описание идеи + draft RICE
Решение: стоит ли делать
Decision: Product Lead + TP
Product Brief Review
Product Brief v0.5 шаблон
Решение: включить в pipel
Decision: Product Review
Tech Design Review
Tech Design Doc v0.5 шабло
Решение: согласна ли плат
Decision: Architecture Revie
Delivery Planning
Roadmap + ресурсы
Решение: когда и кем дела
Decision: Квартальное план
Launch Review
Product Brief + Tech Design

# Portfolio / Capacity Management

Capacity в квартале распределяется по бакетам:

- **Security ≈ 10%**
  - задачи от команды безопасности (уязвимости, комплаенс, аудит).
  - приоритизация согласуется между продуктовой командой и Security.
- **Обязательные задачи ("нельзя не сделать") ≈ 10%**
  - поддержка, обновления, критический техдолг.
- **Сквозные проекты (10–15%)**
  - общие инициативы по платформе, влияющие на все сервисы (например, унификация resource model, интеграция в IAM/Billing/Monitoring).
- **R&D Product (35–40%)**
  - новые продуктовые инициативы, идущие через Product Brief/Tech Design Doc процесс.
- **R&D Tech (35–40%)**
  - инициативы технических команд: оптимизации, миграции, улучшение пайплайнов, снижение COGS.

## Guidance

- Security и «нельзя не сделать» не фиксированы первыми по дефолту, приоритет определяется совместно с Security/TPM.
- Сквозные проекты идут от CTO/CPO облака
- Остаток capacity балансируется между R&D Product и R&D Tech. Мы стремимся соблюдать баланс как в целом по всей команде, так и внутри подкоманд.
- Приоритизация внутри бакета:
  - R&D Product → по RICE + Strategic Value.
  - R&D Tech → по техценности (COGS, скорость команд, риски).