ETHICAL DECISION-MAKING FRAMEWORK FOR AI PROJECTS

AI Leadership & Project Management Masterclass

Curtin University

2025-10-27

ETHICAL DECISION-MAKING FRAMEWORK FOR AI PROJECTS

Four Questions to Guide Difficult Decisions

When to Use This Framework

Use these four questions whenever you face:

- **Ethical dilemmas** in AI deployment
- **Bias concerns** in AI models
- **Privacy questions** about data use
- **Transparency issues** with AI decisions
- **Fairness concerns** about AI impact
- **Any decision** where values and ethics matter

The Four Questions

1. WHO BENEFITS?

Ask yourself:

- Who gains value from this AI system?
- Who might be harmed or disadvantaged?
- Are the benefits distributed fairly?
- Who is excluded or left behind?
- Does this widen or narrow inequality?

Example (RetailFlow chatbot):

- Benefits: Customers (faster service), Company (lower costs)
- Potential harm: Customer service staff (job security)
- Question: Are we transparent about automation with staff?

Red flags:

- Benefits concentrate to company, costs to customers or employees
- Vulnerable groups disproportionately affected
- Can't clearly articulate who benefits

---

2. WHAT COULD GO WRONG?

Ask yourself:

- What are the worst-case scenarios?
- What unintended consequences might occur?
- What biases might the AI amplify?
- How could this be misused?
- What happens if the AI fails?

Example (RetailFlow chatbot):

- AI might provide wrong information → customer harm
- AI might be biased against certain customer groups
- Data breach could expose customer conversations
- Over-reliance on AI could degrade human service skills

Red flags:

- You can't imagine how it could go wrong (lack of imagination)
- Worst-case scenario is catastrophic but you proceed anyway
- "It'll probably be fine" is your risk assessment

---

3. HOW DO WE KNOW IT'S WORKING AS INTENDED?

Ask yourself:

- What metrics prove this AI is ethical and fair?
- How do we detect bias or harm?
- Who monitors the AI's decisions?
- How often do we audit?
- What triggers a review or pause?

Example (RetailFlow chatbot):

- Monitor accuracy across different customer demographics

- Track escalation rates (are certain groups escalated more?)
- Review random conversations for quality
- Customer satisfaction surveys by demographic
- Monthly bias audits

Red flags:

- No plan to monitor fairness or bias
- "We'll check if someone complains"
- No clear metrics beyond business ROI
- Technical team monitors but no ethics oversight

---

4. WHEN DO WE STOP?

Ask yourself:

- Under what conditions do we pause or shut down?
- What's our threshold for unacceptable harm?
- Who has authority to stop the system?
- Is there a kill switch?
- What triggers a full review?

Example (RetailFlow chatbot):

- **Immediate stop:** Data breach, safety issue, discrimination detected
- **Pause for review:** Accuracy drops below 90%, customer complaints spike
- **Full audit:** Bias detected in any demographic group
- **Authority:** AI Ethics Committee can order immediate stop

Red flags:

- "We'll stop if it's really bad" (no definition of "really bad")
- No clear authority to shut down
- Business pressure overrides ethics concerns
- Sunk cost fallacy: "We've invested too much to stop now"

How to Use This Framework (Practical Steps)

During Crisis 4 Exercise (Ethical Dilemma)

When you discover the RetailFlow chatbot has bias:

1. **Work through all four questions** with your team
2. **Document your answers** - write them down
3. **Identify conflicts** - Where do answers conflict with business goals?

4. **Make a decision** - What will you recommend?
5. **Justify with the framework** - Use the four questions in your recommendation

In Real Projects

Before launch:

- Answer all four questions in project planning
- Get stakeholder input on each question
- Document answers and include in project charter

During operation:

- Revisit quarterly or when major changes occur
- Update answers based on real-world performance
- Use as basis for ethics audits

When problems arise:

- Return to the four questions immediately
- Don't rationalize - answer honestly
- Let answers guide your response

Example: Applying the Framework to RetailFlow Bias

Scenario: Chatbot escalates complaints from certain postcodes (lower-income areas) to human agents less often than complaints from wealthy postcodes.

Question 1: Who benefits?

- **Currently:** Wealthy customers get faster resolutions
- **Harmed:** Lower-income customers wait longer, get worse service
- **Unfair distribution:** System creates two tiers of service

Question 2: What could go wrong?

- Class-based discrimination becomes embedded in business
- Reputation damage if discovered publicly
- Legal liability for discriminatory practices
- Reinforces existing inequalities

Question 3: How do we know?

- Audit showed: 45% escalation rate for wealthy postcodes, 12% for lower-income postcodes
- Customer satisfaction: 85% (wealthy) vs. 62% (lower-income)
- **We know because we measured** - good!

Question 4: When do we stop?

- **Answer:** We pause AI escalation decisions immediately
- Manual review of all escalation decisions while we fix the bias
- Resume only after bias is eliminated and audited

Decision: Pause the biased feature, fix it, transparently communicate to affected customers.

Common Ethical Traps to Avoid

"We didn't know"

- Wrong: You have a responsibility to find out
- Right: Build monitoring and auditing from day one

"It's technically legal"

- Wrong: Legal ethical
- Right: Ethics go beyond legal compliance

"Everyone else does it"

- Wrong: Doesn't make it right
- Right: Be better than the industry standard

"We'll fix it later"

- Wrong: Harm is happening now
- Right: Stop, fix, then resume

"The benefits outweigh the harm"

- Wrong: Who decides? Who bears the harm?
- Right: Those harmed should have a voice in the decision

"We can't afford to stop"

- Wrong: You can't afford NOT to stop if ethics are violated
- Right: Reputation damage costs more than pausing

Decision Tree for Ethical Dilemmas

```
Is there potential harm to people?
  NO → Proceed with standard risk management
  YES → Answer the four questions

      Clear harm, no mitigation possible?
          STOP the project or feature

      Harm can be mitigated?
          Implement mitigation BEFORE proceeding

      Uncertain about harm?
           Run controlled pilot with extra monitoring
```

Your Ethical Leadership Responsibilities

As an AI project leader, you must:

1. **Ask these questions** - even when it's uncomfortable
2. **Listen to concerns** - especially from affected groups
3. **Act on findings** - don't rationalize away ethical issues
4. **Be transparent** - with stakeholders and affected people
5. **Have courage** - to stop projects that cause harm
6. **Document decisions** - create an audit trail
7. **Learn and improve** - from ethical mistakes

Remember: Your job is not just to deliver AI projects. It's to deliver AI projects **responsibly**.

Key Takeaway

The four questions:

1. Who benefits?
2. What could go wrong?
3. How do we know?
4. When do we stop?

If you struggle to answer any of these questions clearly and honestly, that's a red flag. Pause and investigate before proceeding.

Additional Resources

After the course, explore:

- Australia's AI Ethics Principles
- IEEE Ethically Aligned Design standards
- EU AI Act requirements
- Your organization's AI ethics policies

Keep this framework handy. Use it. It will save you from making decisions you'll regret.