Task 7: Ethical Implications of Decision Making

Purpose: To provide actionable recommendations for improving U.S. soccer performance while ensuring transparency, ethical responsibility, and reproducibility of analysis.

Executive Summary

This report builds upon the descriptive statistics and narratives generated in Task 5 and the deep fake interview clips from Task 6 to provide actionable recommendations for U.S. soccer stakeholders. The analysis of historical international football data (1872–2017) shows that:

• The game has evolved toward lower-scoring matches, with home vs away advantages narrowing over time. Goal averages declined from 3.0 (1870s) to 1.6 (2010s).

• Seasonal scoring patterns peak in April–May, coinciding with tournament calendars. Seasonal peaks in April–May: ~1.9–2.0 home goals per match.

• Post-tournament matches often see a scoring bump, though results vary by competition. Post-tournament scoring increases by up to +0.6 goals.

• World Cup performance is more challenging than friendlies; global data shows friendlies inflate win rates. World Cup win rates (e.g., Germany 69.9%) differ sharply from friendlies (Germany 52.9%).

• Rivalries like Argentina–Uruguay or England–Scotland show how repeated matchups push improvement; for the U.S., Mexico remains the defining rivalry. Argentina–Uruguay (183 matches), England–Scotland (118).

Tiered recommendations are presented:
- Low Risk: Operational adjustments, like reinforcing defensive drills in the final 20 minutes of games.
- Medium Risk: Further data collection for U.S. performance post-2017 to validate findings.
- High Risk: Strategic staff or player role changes, requiring oversight by HR/legal teams.

Ethical concerns include dataset bias, fairness, and the risks of AI-generated deep fake content. Confidence in findings is moderate to high, with statistical validation included. Next steps focus on robustness checks, fairness audits, and expanding analysis with current data.

Background & Decision Context

Stakeholder: The key stakeholders here are the coaching staff and directors of the U.S. Soccer Federation. They are responsible for decisions on training, player development, and strategy at the national level. These decisions directly affect team performance and also shape how the U.S. is seen in international football.

Decision Context: This project provides insights to guide tactical planning and long-term competitive strategy. The focus is on using data to answer practical questions like: Should more emphasis be placed on defense in the final minutes of games? How do rivalries impact performance? Are friendlies giving a false impression compared to World Cup matches? These decisions matter because they can affect results in major competitions and player development.

Risk Level: The risk level is **medium-to-high**. Some actions, like changing drills or training focus, are low risk and easy to implement. Others, like changing staff roles or player selections, carry higher stakes because they can impact careers and the team's reputation. That's why the recommendations are organized into low, medium, and high-risk categories.

Data & Methods

The analysis is based on the Kaggle International Football Results dataset, covering over 40,000 matches from 1872 to 2017. The dataset is strong but has more European and South American matches compared to other regions, which can skew results.

I used Python to calculate descriptive statistics such as goal averages, win rates, and rivalry frequencies. I also asked Large Language Models (LLMs) the same questions to compare their answers to the Python outputs. The LLMs were accurate about 85–100% of the time, though sometimes they rounded numbers.

Key examples include:

- Home goals dropped from about 3.0 in the 1870s to 1.6 in the 2010s.

- Seasonal peaks: April–May games averaged ~1.95 home goals.

- Brazil's win rate: 63.5% across 1,049 matches.

- Germany: 69.9% win rate at World Cups vs 52.9% in friendlies.

- Mauritius: longest win streak of 17 matches.

- Argentina–Uruguay rivalry: 183 matches, the most frequent in international football.

This approach gave both precise stats (Python) and a check on how well AI tools can summarize and interpret them (LLM).

Findings (with Uncertainty)

1. Evolution of the Game: Narrowing home advantage (CI excludes zero).

2. Seasonal Scoring: Significant April–May peak confirmed via bootstrap resampling.

3. Tournament Effects: Average +0.3–0.6 goals after tournaments (heterogeneous impact).

4. Rivalries: Frequent rivalries drive performance consistency. Causal link uncertain.

5. World Cup vs Friendlies: World Cup WR significantly higher ($p<0.05$ across nations).

6. Host Advantage: Gold Cup WR 76%, World Cup WR 61% (CI wide for small competitions).

Risk Matrix

Recommendation	Risk Level	Impact	Notes
Late-game defensive drills	Low	Operational (low cost)	Quick to implement, low downside.
Collect modern match data	Medium	Investigatory (moderate)	Requires resources; supports stronger validation.
Staff/player role changes	High	Strategic (high stakes)	Impacts careers; requires HR/legal oversight.

Ethical / Legal Concerns
- Provenance: Dataset public, no personal identifiers.
- Fairness: Overrepresentation of Europe/South America skews trends.
- LLM Risks: Potential hallucinations; need validation.
- Deep Fake: Task 6 showed realistic but fictional interviews; must label outputs clearly.

Next Steps & Validation Plan
- Collect and analyze post-2017 match data.
- Conduct robustness checks with varied parameters.
- Implement fairness and subgroup audits.
- Document every script, prompt, and result.

Process Documentation
The professor emphasized: 'I am less concerned with the final report and more interested in your process: how you approached the task. Document your process thoroughly.'

Process Steps:
1. Task 5: Generated descriptive stats with Python, validated with LLM.

2. Task 6: Created Gemini Veo deep fake clips; faced 8s limit.

3. Issues: Clips stylistically inconsistent, required simplification.

4. Task 7: Combined narrative + ethical framing, introduced risk tiers, uncertainty, and fairness audits.

5. Documented prompts, code, edits for reproducibility