Cultural Impact and Fairness in AI Systems: Report Analysis and Practical Application

Hunjun Shin

1 Part 1: Summary of Key Findings (Sections 2, 3, and 4)

1. **The report states, "AI is not neutral." Explain what this means, using the example of facial recognition technology discussed in the text.**

   The statement **"AI is not neutral"** means that AI systems are not simply objective technical tools; instead, they are socio-technical artifacts that frequently reflect and reinforce the dominant cultural norms, assumptions, and values of their primarily Western-centric developers. This results in systems that are ill-equipped to serve culturally diverse populations and often perpetuate existing patterns of exclusion and inequality. For example, **facial recognition technology** exhibits higher error rates for darker-skinned individuals, particularly women, compared to lighter-skinned men. These disparities arise because the training datasets used to build the technology are overwhelmingly composed of lighter-skinned, male faces, embedding bias from the beginning.

2. **What is "digital colonialism" as described in the report, and how does it relate to the dominance of the Global North in AI development?**

   Digital colonialism is a systemic risk in which dominant technological paradigms dictate how different cultural groups can engage in online spaces, leading to the marginalization of non-Western epistemologies and knowledge systems. This phenomenon is directly related to the **dominance of the Global North**, where AI development is geopolitically concentrated. This concentration results in technologies developed in dominant regions being applied elsewhere with insufficient consideration for local contexts. Consequently, Global South countries often become passive consumers of foreign technologies, which reinforces existing dependencies and limits their technological sovereignty.

3. **According to Section 3.1, why do AI content moderation tools often fail when dealing with non-Western languages and dialects? Provide one example from the report.**

   AI content moderation tools frequently fail in dealing with non-Western languages and dialects primarily due to the **lack of diverse training data**. Without adequate data, these systems struggle to differentiate between genuine hate speech, satire, and culturally specific expressions. This deficiency leads to tools that are ineffective at consistently identifying harmful content while simultaneously risking the over-censorship of legitimate expressions from marginalized groups. For instance, AI moderation on platforms like WhatsApp struggles to identify **hate speech directed at lower-caste communities in India** because discriminatory language is often encoded in culturally specific terms or regional dialects.

4. **The report proposes the use of "cultural impact assessments" for high-risk AI systems. What is the purpose of such an assessment?**

   The purpose of a **cultural impact assessment** is to serve as a proactive evaluation—similar to an environmental impact report—designed to identify, predict, and mitigate potential adverse effects on a community's norms, values, traditions, and languages. This assessment is required for all high-risk AI systems deployed in sensitive domains like hiring, education, healthcare, and law enforcement. By consulting local experts and affected groups, the assessment analyzes how an AI system might reinforce stereotypes, marginalize specific dialects, or misinterpret culturally significant content, then outlines concrete mitigation steps.

5. **Why does the report advocate for "participatory design"—involving affected communities from the outset—as a key strategy for building fairer AI?**

   The report advocates for **participatory design** because achieving true algorithmic fairness requires structural change, moving beyond mere technical adjustments. This method ensures that the experiences and perspectives of marginalized communities are centered throughout the AI development process, rather than treating inclusivity as an afterthought. By involving affected communities, participatory design facilitates the **co-creation of AI systems** that affirm cultural diversity, challenge dominant power structures, and make justice, rather than just efficiency, the guiding design principle.

2 Part 2: AI Assistant "Whiskey4U" Demonstration

The following link directs to the "Whiskey4U" AI assistant. This prototype was adapted to apply the principles of cultural awareness and bias mitigation discussed in the report and analyzed in the reflection below.

Link to AI Assistant: `https://www.useinvent.com/e/ast_2dzvUqDnQvWgB1KzUhgVpj`

3 Part 3: Reflection on Adapting the AI Assistant

The process of adapting the "Whiskey4U" prototype into a more culturally aware tool, as required by the Coding Challenge, highlighted the stark difference between aspirational AI ethics and practical implementation. This dual approach of proactive design and thoughtful reconfiguration of systems already in operation is necessary to ensure technological progress benefits all population groups, not just the dominant few.

The Biggest Challenge: Technical Limitations and the Retrofitting Problem

The greatest challenge in adapting the assistant was the constraint of the existing tool's infrastructure. Instead of a technical solution like dynamic routing to detect a user's region, the adaptation had to rely on a dialogue-based approach. By modifying the assistant's core **Constraints** and **Guidelines**, we forced it to actively ask for clarification when faced with ambiguity, rather than assuming a default context.

This practical hurdle perfectly mirrors the large-scale problem of retrofitting existing technology discussed in the Savage & Savage report. The report emphasizes that AI development is often shaped by the priorities of the Global North, resulting in systems that reflect and reinforce dominant cultural norms. When AI systems are not designed with cultural pluralism in mind from the outset, incorporating inclusivity becomes an effort of "adapting existing systems, not just replacing them." The need to build dialogue-based logic is a microcosm of this structural problem. Since the tool's underlying architecture was built on a "one-size-fits-all" standard, achieving fairness necessitated moving to a context-aware approach, proving that retrofitting systems for transparency requires the same level of cultural sensitivity as building them from scratch.

Policy Recommendations in Practice: From Passive Fixes to Active Mitigation

The adaptations made to "Whiskey4U" were direct attempts to operationalize key policy recommendations from Section 4 of the report, moving beyond simple, passive fixes to active, real-time bias mitigation.

First, we put the principle of **Integrate Perspectives and Priorities of Diverse Communities** into practice by **Diversifying Its Knowledge Base**. AI systems risk contributing to cultural homogenization by privileging Western-centric data. We addressed this by explicitly requiring the assistant's knowledge base to include whiskies from producers in **India (Amrut), Taiwan (Kavalan), and Japan (Nikka)**, along with culturally specific consumption methods like the Japanese *Mizuwari* and Highball. Food pairings were expanded beyond the Western norm of steak and cheese to include suggestions like pairing a smoky Scotch with **Korean Bulgogi** or an Indian single malt with **curry**. This effort ensures the assistant reflects pluralism.

Second, we enacted the principle to **Promote transparency and accountability** by upgrading its logic to **Enhance Its Ability to Handle Ambiguity**. When faced with an ambiguous query like, **"How should I drink this whiskey?"** the assistant is now constrained from defaulting to the Western-centric "neat or on the rocks." Instead, it must present a range of culturally diverse options, promoting transparency by clarifying its lack of a single default context.

Most significantly, the adaptation moved beyond ensuring the assistant's *own* language is neutral (a passive fix) to actively handling **biased user input**. Through a new core **Constraint**, the system was trained to recognize and reframe gendered stereotypes. For example, if a user asks for a **"real man's whisky"** or makes an assumption like, **"My boss is a woman... I assume she prefers something smooth, right?"**, the system is now required to recognize the embedded bias. Instead of fulfilling the request as stated, its guidelines compel it to reframe the query around neutral preferences like flavor profiles (e.g., "smoky, sweet, or spicy"). This aligns with the call for AI models to be trained to detect discriminatory patterns. By ensuring the system is not merely clean in its output but robustly handles biased input, the assistant embodies the principle that inclusion is an ongoing responsibility, moving beyond isolated fixes to address socio-technical structures.

Future Development: Implementing a Community-Sourced Feedback Loop

If I were to continue developing this assistant, the single most impactful feature to add would be a **community-sourced feedback and correction loop**. This feature would allow users to submit suggestions, corrections, or new cultural knowledge directly to the assistant—for example, a new regional food pairing, a local consumption custom, or information about an emerging distillery. These submissions would be flagged for review by human moderators and, if verified, integrated into the assistant's core knowledge base.

My choice is based directly on the report's emphasis on **"participatory design"** as a crucial strategy for building truly fair AI. While the current adaptations have made the assistant more inclusive, its knowledge is still curated and defined by its developers in a top-down manner. A feedback loop would begin to decentralize this power. It would transform the assistant from a static authority into a dynamic, "living" system that evolves with input from the diverse communities it serves. This feature is essential because it provides a structural mechanism for marginalized groups to represent their own cultures and priorities authentically, ensuring the AI remains relevant, accurate, and genuinely inclusive over the long term. It moves beyond simply consuming pre-approved knowledge to actively co-creating a more equitable information ecosystem.