InWorld - Output Model: AI Impact Assessment Proposal

Pete Trujillo

Introduction:

It is critical for InWorld to conduct an AI Impact Assessment to evaluate and understand all potential risks and impacts of their product the "Output Model." The Output Model's use of generative AI (GenAI) offers exciting innovations but has many risks that must be understood and mitigated. Conducting an AI Impact Assessment would help identify risks, ensure regulatory compliance, and ensure the ethical use of GenAI.

- **Product:** InWorld's Output Model, simulates highly realistic character performances. The Output Model utilizes GenAI algorithms, including Neural Text to Speech, to simulate human voices, and Large Language Models to simulate human dialogue (Choi, 2024; Inworld Team, 2024).
- **Risks:** The Output Model has risks associated with GenAI technologies. Consumers may be confused about whether the **performance is synthetic or human** (*PAI's Responsible Practices for Synthetic Media*, n.d.). Likeness to existing people or deep fakes may cause **copyright/liability** issues (Kahveci, 2023; Krouse et al., 2024). **Negative stereotypes or biases** could be reinforced by synthetic performances (UNESCO, 2024). Humanlike NPCs could manipulate players with **deceptive practices**, such as coercing them into purchasing digital items (Petrovskaya & Zendle, 2022)**.** Synthetic performances may result in **job losses** for human voice performers, impacting InWorld's public image (Anderson, 2014). Additionally, there is growing **GenAI regulation** from state governments, which could significant regulatory impact on InWorld (EU Artificial Intelligence Act, 2024).

Proposed Framework:

The current software build of the Output Model is V2.0, placing its product lifecycle into the **Operations and Monitoring** phase. The NIST AI Risk Management Framework (AI RMF) is the recommended for the Output Model's AI impact assessment.

- **Rationale:** AI RMF provides a flexible framework for identifying risks at all lifecycle stages, including the operation and monitoring phase. AI RMF offers reference libraries of potential AI risks, developed by the AI research community. AI RMF offers a specific GenAI "profile" template, which lists risks that are unique or exacerbated by GenAI (Nist, 2024).
- **Benefits:** Using the AI RMF framework for an impact assessment would be an investment towards a longer-term risk management governance strategy. AI RMF has been endorsed by tech companies, industry associations, research institutions, and advocacy groups *(Perspectives about the NIST AI RMF, 2023)*. Additionally, the AI RMF is updated semi-annually to keep pace with AI innovations.
- **Limitations**: Risk categories may be too technical and do not address social risks (Wesen, 2022). Framework implementation is time-consuming and requires significant investment. AI RMF provides a framework process, but not procedures for managing risks or addressing regulations.

Existing Oversight and Governance:

Company policies, state laws/guidance, best practices, actors' feedback, and the GenAI profile would be synthesized into a cohesive list. Items are mapped into risk categories and topics. Each

risk topic is assigned a tolerance score and potential harm. Prioritization would be determined using risk tolerance, potential harm, legal requirements, and company priorities.

- **Company Principles:** InWorld has not published an AI Impact Assessment, but provides Terms of Service and Safety guidelines on their website. These include prohibitions against harmful or unauthorized characters, and emphasize avoiding bias and stereotypes (*Inworld – Terms*, 2023; *Safety Policies | Inworld AI*, n.d.).
- **Industry Standards**: The organization Partnership on AI released a set of guiding principles for Synthetic Media (*PAI's Responsible Practices for Synthetic Media*, n.d.).
- **State Laws:** InWorld is based in California, USA; however, InWorld must consider state laws in markets where their software is sold and where their clients sell video games, including North America, South America, Europe, Asia, and Africa. ***European Union (EU)*** The Artificial Intelligence Act (AI Act) (2024) stipulations include data transparency and AI does not causing harm. ***California USA*****:** Protections include the California Consumer Privacy Act; Protections for voice performers; and GenAI model Transparency. (*AB 2013*, 2024; *AB 2602*, 2024; *CCPA*, 2024). ***China*****:** The Interim Measures for the Management of Generative Artificial Intelligence requires companies to implement quality measures for GenAI training data (*China*, 2023).
- **State Guidance:** In addition to state laws, some states have issued or endorsed guiding AI principles until more formal laws can be enacted. ***International*****:** Hiroshima AI Process Comprehensive Policy Framework provides guiding principles for trustworthy AI systems and is endorsed by over 50 states, including states in Asia, South America, and Africa regions (Hiroshima AI Process, n.d.). ***United States*****:** The Blueprint for an AI Bill of Rights (2022) recommends potential AI standards including, privacy and AI explainability. ***Canada***: The Artificial Intelligence and Data Act (AIDA) (2023) contains voluntary AI guidelines for companies to use until AIDA becomes state law.

Actors Required for AI Impact Assessment:

The AI Impact assessment requires a diverse team of actors who will provide a unique perspective on goals, benefits, costs, and risk tolerances. ***External auditors*** will compile the assessment to ensure impartiality and an external viewpoint. ***Organizational management*** identifies company strategy and economic risks, benefits, and risk tolerances. ***Legal*** identifies regulatory/legal risks, costs, and tolerances. ***Product managers*** identify goals, risks, costs, and benefits using client feedback. ***QA testers*** identify benefits and risks from personal product experience. ***Developers/AI Researchers*** determine risks, benefits, and goals using technical knowledge of the Output model. ***Clients*****,** provide feedback on product experiences.

Conclusion:

Impact assessments help companies identify and prioritize risks against company goals. Other companies, such as Workday, have reported positive results with using the AI RMF to map and measure their risks (Perspectives about the NIST AI RMF, 2023). For InWorld, an AI Impact Assessment would help identify risks, address legal liabilities, increase transparency, address regulatory requirements, and tackle ethical issues. Although investment in AI RMF-based impact assessments may seem high, the benefits outweigh the costs. This work provides a foundation for a long-term AI RMF governance strategy, which is essential for a Generative AI company such as InWorld. In summary, investing in AI RMF-based impact assessment is a strategic move towards sustainable and transparent AI governance.

Citations/References:

AB 2013. (2024, September). CA Legislature. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013

AB 2602. (2024, September 17). CA Legislature. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2602

AI Bill of Rights: Blue Print. (2022, October 4). The White House. https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/

Anderson, A. S. and J. (2014, August 6). AI, Robotics, and the Future of Jobs. *Pew Research Center*. https://www.pewresearch.org/internet/2014/08/06/future-of-jobs/

Artificial Intelligence and Data Act. (2023, September 27). Innovation, Science and Economic Development Canada. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

CCPA. (2024, March 13). California Consumer Privacy Act. https://oag.ca.gov/privacy/ccpa

China: Generative AI Measures Finalized. (2023, July). Library of Congress, Washington, D.C. 20540 USA. https://www.loc.gov/item/global-legal-monitor/2023-07-18/china-generative-ai-measures-finalized/

Choi, J. (2024, May). *Developing LLM benchmarks for conversational realism in lifelike AI agents* [Software]. InWorld. https://inworld.ai/blog/developing-llm-benchmarks-for-conversational-realism-in-lifelike-ai-agents

EU Artificial Intelligence Act. (2024). https://artificialintelligenceact.eu/

Hiroshima AI Process. (n.d.). Retrieved October 6, 2024, from https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html

Inworld – Terms. (2023, July). https://inworld.ai/terms

Inworld Team. (2024, May). *Speech synthesis: The path to creating expressive text-to-speech*. InWorld. https://inworld.ai/blog/speech-synthesis

Kahveci, Z. Ü. (2023). Attribution problem of generative AI: A view from US copyright law. *Journal of Intellectual Property Law & Practice*, *18*(11), 796–807. https://doi.org/10.1093/jiplp/jpad076

Krouse, S., Seetharaman, D., & Flint, J. (2024, May 24). Inside Scarlett Johansson's Battle With OpenAI --- Dispute shows ongoing collision between AI and Hollywood. *Wall Street Journal, Eastern Edition*, A.1.

Nist, G. M. (2024). *Generative Artificial Intelligence Profile* (No. NIST AI NIST AI 600-1; p. NIST AI NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

PAI's Responsible Practices for Synthetic Media. (n.d.). Partnership on AI - Synthetic Media. Retrieved October 4, 2024, from https://syntheticmedia.partnershiponai.org/

Perspectives about the NIST AI RMF. (2023, January 23). NIST. https://www.nist.gov/itl/ai-risk-management-framework/perspectives-about-nist-artificial-intelligence-risk-management

Petrovskaya, E., & Zendle, D. (2022). Predatory Monetisation? A Categorisation of Unfair, Misleading and Aggressive Monetisation Techniques in Digital Games from the Player Perspective. *Journal of Business Ethics*, *181*(4), 1065–1081. https://doi.org/10.1007/s10551-021-04970-6

Safety Policies | Inworld AI. (n.d.). Retrieved October 4, 2024, from https://inworld.ai/docs/resources/safety/

UNESCO. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models—UNESCO Digital Library*. https://unesdoc.unesco.org/ark:/48223/pf0000388971

Wesen, R. (2022, May 2). Recommendations to NIST on the AI Risk Management Framework Initial Draft—CLTC UC Berkeley Center for Long-Term Cybersecurity. *CLTC*. https://cltc.berkeley.edu/2022/05/02/nist-ai-rmf-recommendations/