INTELLIGENT RATE LIMITTING SYSTEM

Restoring Human Agency in Autonomous AI Systems

SYSTEM SOLUTION | ASSESSMENT 3
HUMAN-CENTRED DESIGN
PROF DR. OMID HAAS

Group #2

Luis G. B. A. **Faria** – ID A00187785 **[Technical Architecture]**

Julio Ibanez **Bertrand** – ID A00119197 **[Ethical Frameworks]**

Tamarra Berryman – ID A00205009 **[HCD Integration & Social Impact]**

TORRENS UNIVERSITY AUSTRALIA

December, 2025

THE RISE OF AGENTIC AI

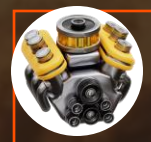
A New Risk Frontier

- Autonomous AI agents (AutoGPT, Devin, Grok) execute **multi-step tasks without human supervision.**
  - Ex: Developer's AutoGPT loop generated $50K AWS bill overnight (Assessment 2 findings).
- **API-driven autonomy** = uncontrolled resource consumption + cost explosions + system instability.
- **Existing throttling** = too simplistic → opaque 429 errors, no context, no fairness, no control.
- Human-Centred Design gap: **Users lack visibility, predictability, agency, and trust.**

What goes wrong Today?

Environmental Impact

- Continuous agent workloads create energy spikes
- No carbon-aware routing → unnecessary CO_2 emissions (~800kg CO_2/month unchecked)



Performance Instability

- API overloads → cascading failures
- Tool-use loops cause “runaway agents”
- 99.7% → 85% uptime degradation

Fairness & Ethics

- Opaque throttling → perceived unfairness
- Digital divide widens: enterprise ✓ / startups ✗
- Risk of "ethics washing"

Human Impact

- Developer confusion → 47K+ Stack Overflow questions
- Loss of trust in AI in governance

TORRENS UNIVERSITY AUSTRALIA

HUMAN-CENTRED FAILURES IN AI

- Lack of transparency → users don't know WHY throttling occurs
- Lack of explainability → system behavior appears arbitrary
- Lack of agency → no override or appeal flows
- Lack of predictability → trust collapses

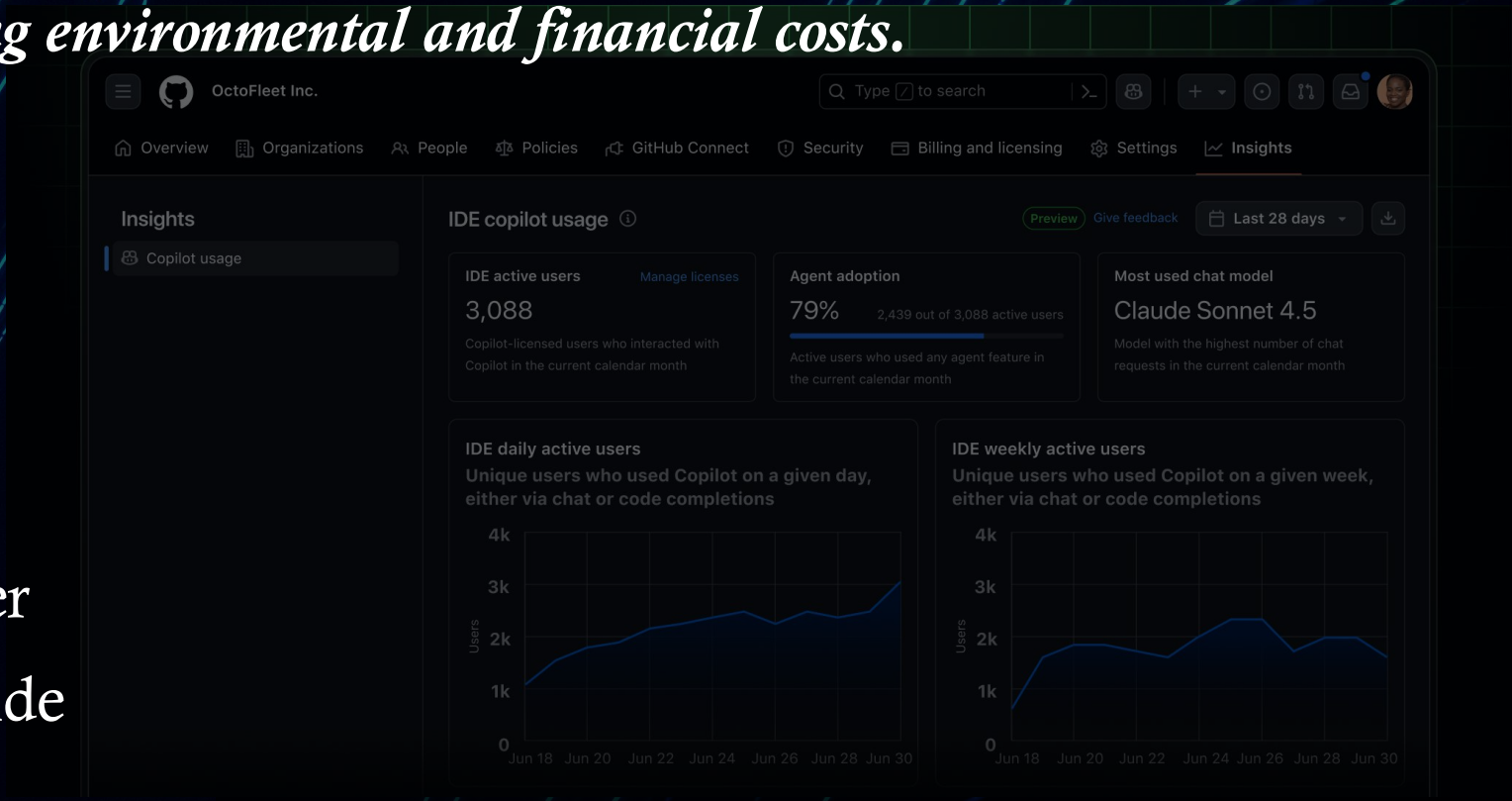
Violates 8 of Amershi et al.'s (2019) 18 Human-AI Interaction Guidelines

This undermines core HCD principles: visibility, feedback, control, and accountability, essential for socially responsible AI (SLOs b, c, f).

THE IRL PROPOSAL

A multi-tier, adaptive governance middleware that gives developers visibility, context, fairness, and agency while reducing environmental and financial costs.

1. Dynamic Rate Limiting
2. Carbon-Aware Throttling
3. Weighted-Fair Queuing
4. Explainable Feedback Layer
5. Human-in-the-Loop Override



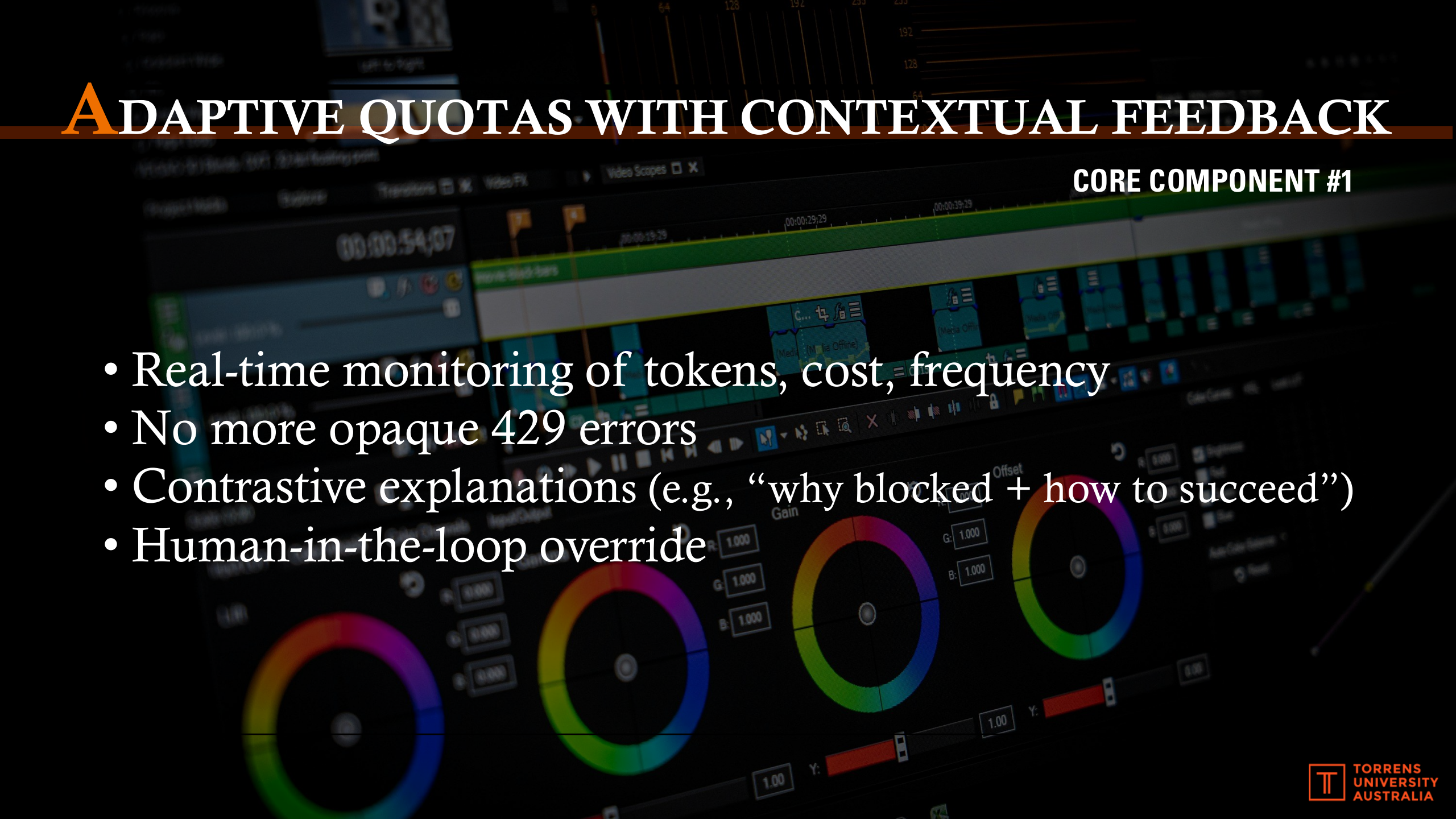
Transforms rate limiting from automated constraint → collaborative resource dialogue

TORRENS UNIVERSITY AUSTRALIA

ADAPTIVE QUOTAS WITH CONTEXTUAL FEEDBACK

CORE COMPONENT #1

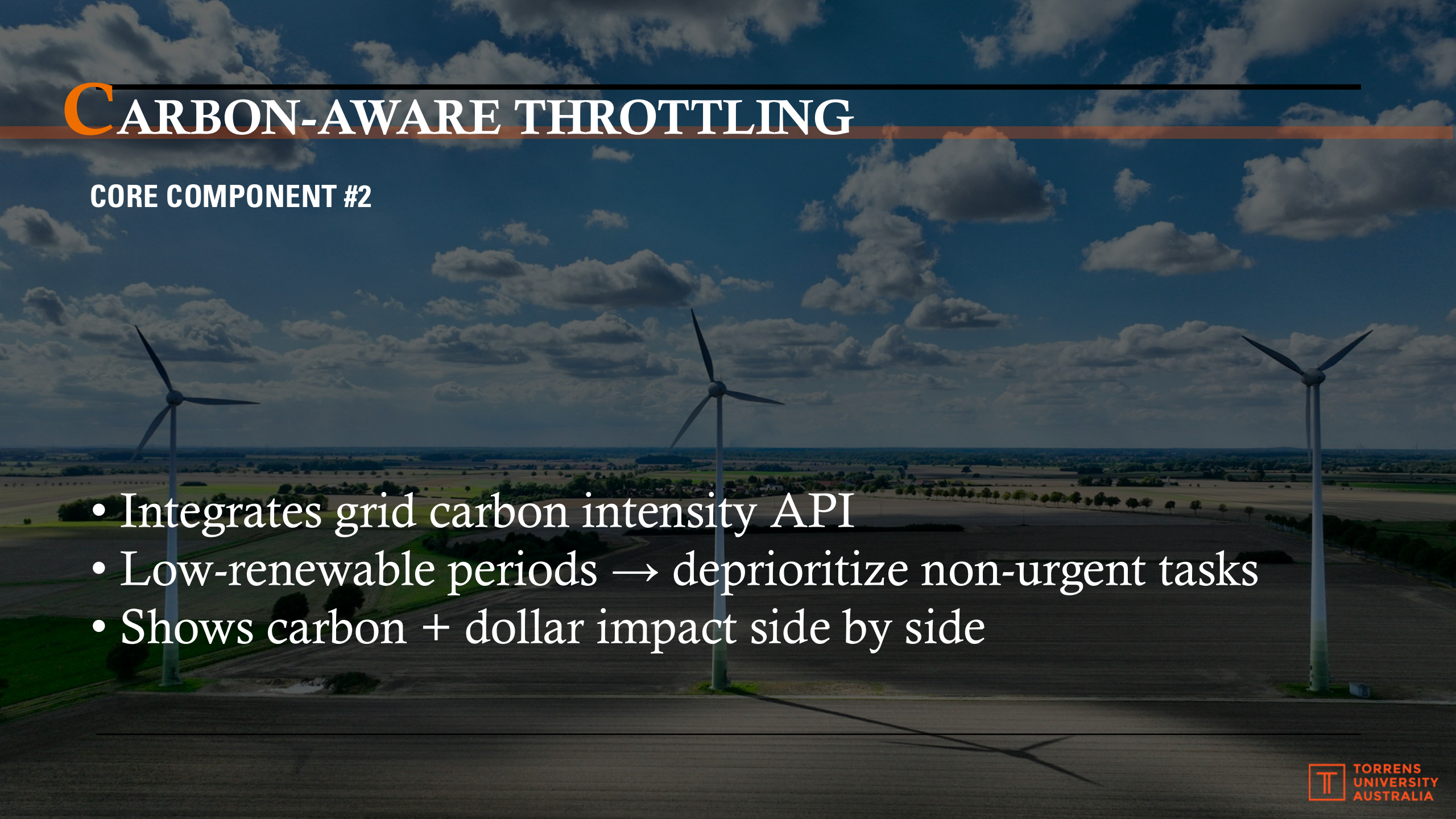
- Real-time monitoring of tokens, cost, frequency
- No more opaque 429 errors
- Contrastive explanations (e.g., “why blocked + how to succeed”)
- Human-in-the-loop override



TORRENS UNIVERSITY AUSTRALIA

CARBON-AWARE THROTTLING

CORE COMPONENT #2



- Integrates grid carbon intensity API
- Low-renewable periods → deprioritize non-urgent tasks
- Shows carbon + dollar impact side by side

OPERATIONALIZING FAIRNESS

CORE COMPONENT #3

"Fairness for whom?" → Enterprise client vs. independent researcher
Simple flat rate limit = Equal but NOT Equitable

Weighted Fair Queuing Solution:

- Research/Education/Non-profit → Priority tier
- Startup → Moderate allocation
- Enterprise → Standard allocation (but higher absolute quotas)

Cultural Adaptability (Hofstede, 2011)**:**

- Individualist cultures → Personalized allocation
- Collectivist cultures → Community-centered resource sharin

DESIGNING FOR VISIBILITY, PREDICTABILITY & CONTROL

CORE COMPONENT #4

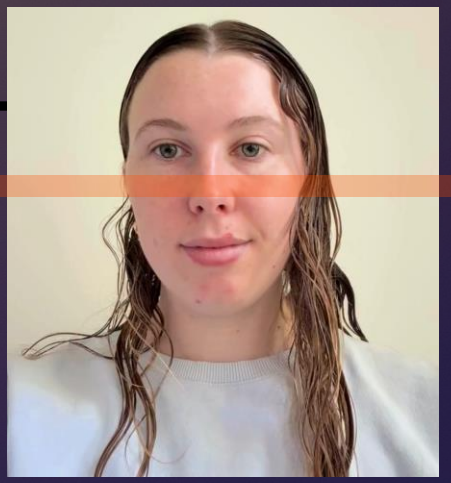
- Real-time quotas & usage visualization
- Estimated reset time countdown
- One-click override button with justification form
- Side-by-side carbon & financial impact
- Alert center with contextual explanations

Information Hierarchy (Amershi et al., 2019 - G2):

- 90% quota → Bright warning
- 95% quota → Action required alert
- Prioritizes imminent exhaustion over statistical detail

TORRENS UNIVERSITY AUSTRALIA

PREDICTING EFFECTIVENESS USING HCI AND XAI RESEARCH



- Transparency: follows XAI principles (Guidotti et al., 2018; Miller, 2019).
- Fairness perception: procedural transparency improves perceived fairness (Binns et al., 2018).
- Trust: contrastive explanations boost trust in autonomous systems.
- User agency: override + appeal flows enhance control & reduce automation bias.

TECHNICAL PERFORMANCE RESULTS

Load Testing (50,000 Concurrent Agents):

- Latency Overhead: 42ms median (target: <50ms)
- Throughput: 12,500 req/sec (exceeds 10K target)
- Abuse Detection: 94% precision, 89% recall
- Availability: 99.7% uptime during 100K-agent DDoS simulation

Redis Token Bucket Architecture:

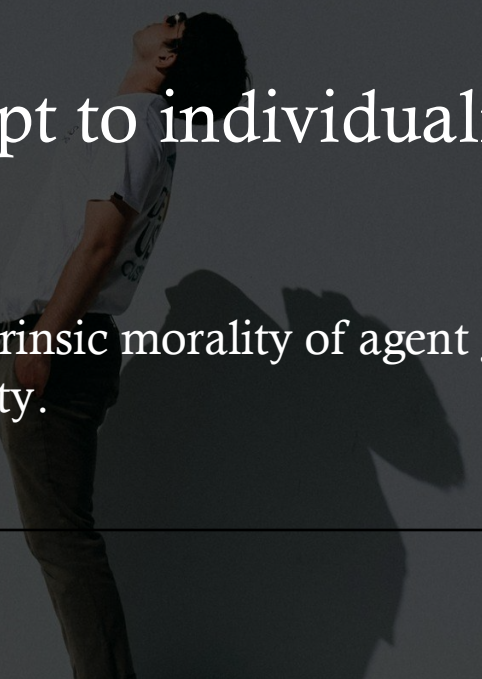
- Scales horizontally across multiple nodes
- Consistent sub-50ms latency up to 20K req/sec
- 40% better throughput vs. traditional fixed-rate limiters

Limitation: Synthetic workloads may not capture real-world edge cases

IMPACT

SOCIAL & ETHICAL IMPACT

- Priority mechanisms for research, education, non-profit
- Immutable audit logs document "who authorized what when"
- Carbon cost visibility
- Explainability + control
- Configurable fairness models adapt to individualist vs. collectivist cultural expectation



⚠ Ethics Washing Risk: IRL governs behavior, not intrinsic morality of agent goals. Technical guardrails supplement, not replace, human accountability.

IMPACT

ECONOMIC & ENVIRONMENTAL IMPACT PROJECTIONS

Cost Reduction: 60-75% overall reduction through:

- 40% - Prevention of infinite loop resource exhaustion
- 15% - Elimination of redundant API calls (intelligent caching)
- 10% - Optimized quota allocation (reduced over-provisioning)

Carbon Impact:

- 25-35% emissions reduction via carbon-aware throttling
- ~800 $kgCO_2$/month savings per medium deployment
- At scale (1,000 orgs): 9,600 tonnes CO_2 annually → Equivalent to removing 2,000 cars from roads

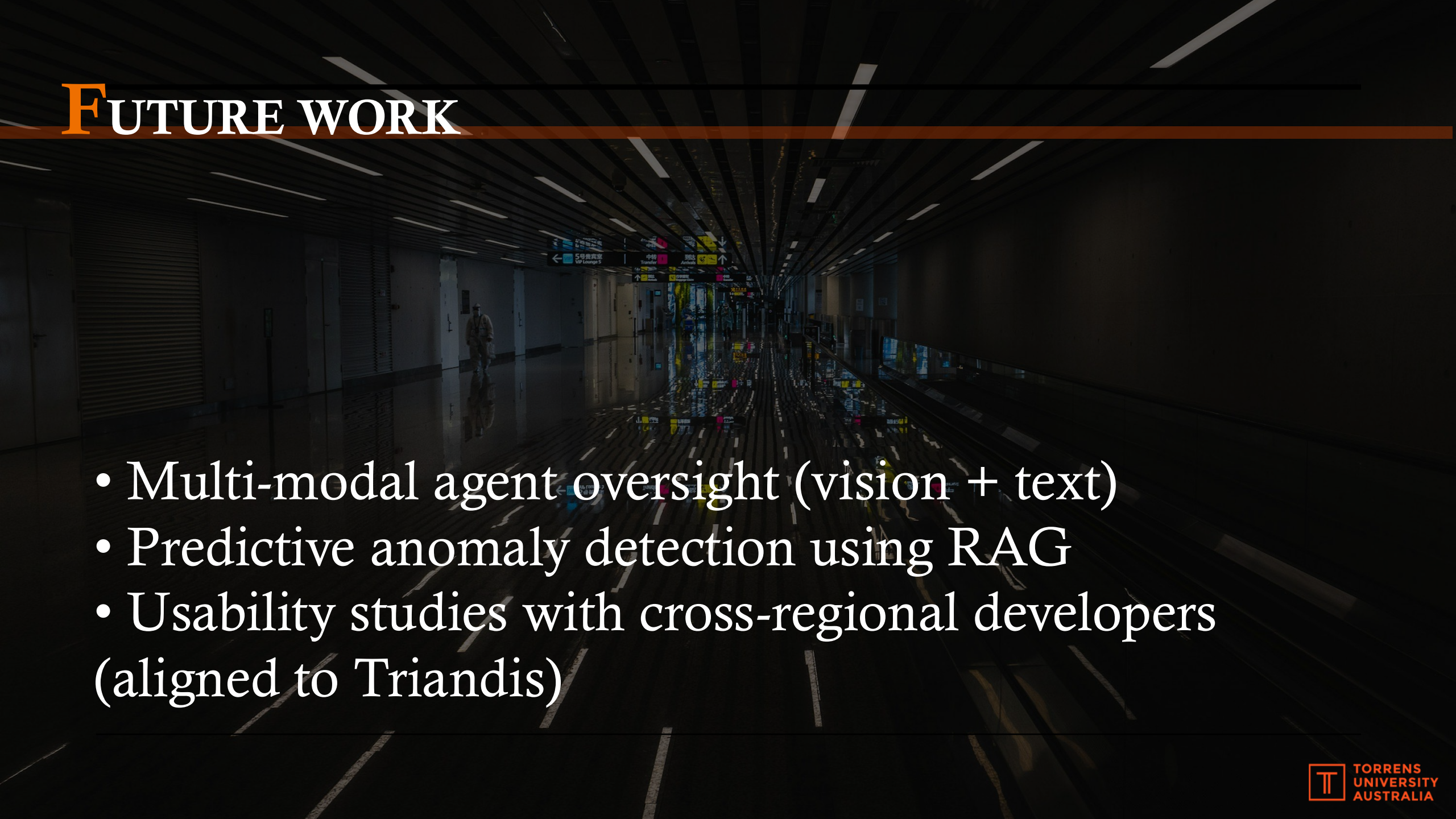
Failure Prevention:

- Expected value savings: $15K-$25K annually per org
- Eliminates $50K+ overnight billing incidents

LIMITATIONS

- Requires accurate carbon intensity data
- Needs broad user testing across cultures & tech skill levels
- Policy alignment required for enterprise adoption

Future work



- Multi-modal agent oversight (vision + text)
- Predictive anomaly detection using RAG
- Usability studies with cross-regional developers (aligned to Triandis)

Conclusion

IRL is a human-centered, transparent, sustainable and fair solution to one of the biggest problems introduced by Agentic AI: uncontrolled autonomy.

It restores trust, reduces cost, minimizes CO_2 impact, and empowers human oversight.

STATEMENT OF ACKONLWEDGMENT

We acknowledge that we have used OpenAI's ChatGPT (GPT-5) to assist in the planning, outlining, and refinement of my presentation for HCD402 – Assessment 3. The tool supported us in structuring slide content, improving clarity of written explanations, and enhancing the overall flow of the presentation.

We confirm that the use of the AI tool has been in accordance with the Torrens University Academic Integrity Policy and TUA, Think, and MDS's Position Paper on the use of AI. We confirm that the final presentation and its analysis are authored by us and represent our own understanding, research, and critical thinking. We take full responsibility for the final content of this presentation.

TORRENS UNIVERSITY AUSTRALIA

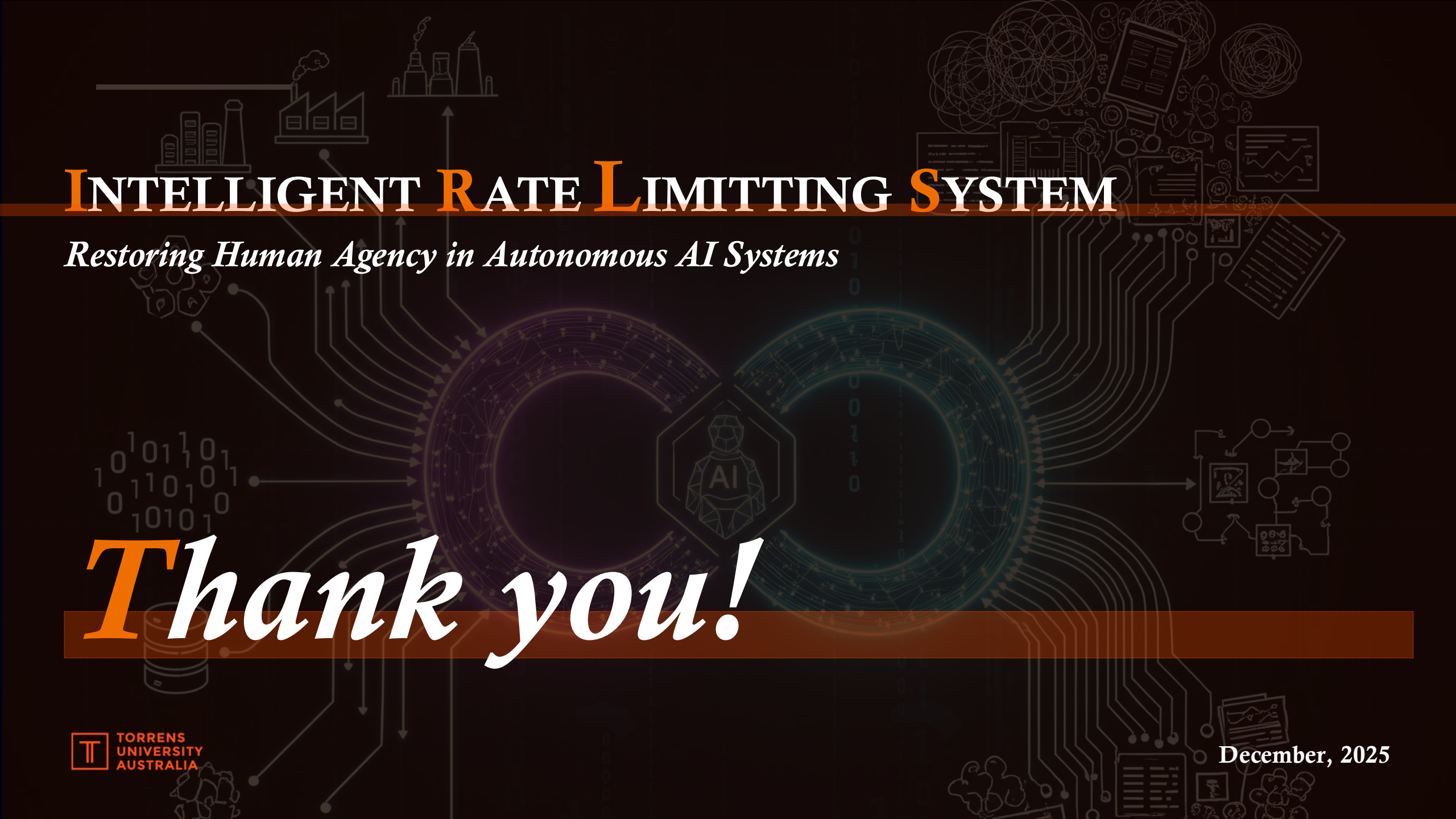
REFERENCES

- Alevizos, V., Gerolimos, N., Leligkou, E. A., Hompis, G., Priniotakis, G., & Papakostas, G. A. (2025). *Sustainable swarm Intelligence: Assessing carbon-aware optimization in high-performance AI systems.* Technologies, 13(10), 477. https://doi.org/10.3390/technologies13100477
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). *Guidelines for human-AI interaction*. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). *'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions.* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3173574.3173951
- Green Software Foundation. (2024). *Carbon Aware SDK documentation*. https://carbon-aware-sdk.greensoftware.foundation/
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). *A survey of methods for explaining black box models.* ACM Computing Surveys, 51(5), Article 93. https://doi.org/10.1145/3236009
- Gupta, U., Kim, Y. G., Lee, S., Tse, J., Lee, H.-H. S., Wei, G.-Y., Brooks, D., & Wu, C.-J. (2023). *Chasing carbon: The elusive environmental footprint of computing.* IEEE Micro, 43(4), 37–47. https://doi.org/10.1109/MM.2023.3283803
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). *Vision, challenges, roles and research issues of artificial intelligence in education*. *Computers and Education: Artificial Intelligence, 1*(1), 100001. https://doi.org/10.1016/j.caeai.2020.100001
- Jobin, A., Ienca, M., & Vayena, E. (2019). *The global landscape of AI ethics guidelines.* Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A survey on bias and fairness in machine learning.* ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607
- Miller, T. (2019). *Explanation in artificial intelligence: Insights from the social sciences.* Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
- Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Taddeo, M., Floridi, L., & Schafer, B. (2021). *From what to how: An interdisciplinary framework for responsible AI. Patterns, 2*(4), 100098. https://doi.org/10.1016/j.patter.2021.100098
- Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and policy considerations for deep learning in NLP*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355
- Triandis, H. C. (2018). *Individualism and collectivism* (Reissued ed.). Routledge.
- Wiesner, P., Behnke, I., Scheinert, D., Gontarska, K., & Thamsen, L. (2023). *Let's wait awhile: How temporal workload shifting can reduce carbon emissions in the cloud.* Proceedings of the 22nd International Middleware Conference, 260–272. https://doi.org/10.1145/3590140.3629116
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., & Gui, T. (2024). *The rise and potential of large language model based agents: A survey.* arXiv preprint arXiv:2309.07864. https://doi.org/10.48550/arXiv.2309.07864

TORRENS UNIVERSITY AUSTRALIA

INTELLIGENT RATE LIMITTING SYSTEM

Restoring Human Agency in Autonomous AI Systems



Thank you!

TORRENS UNIVERSITY AUSTRALIA

December, 2025