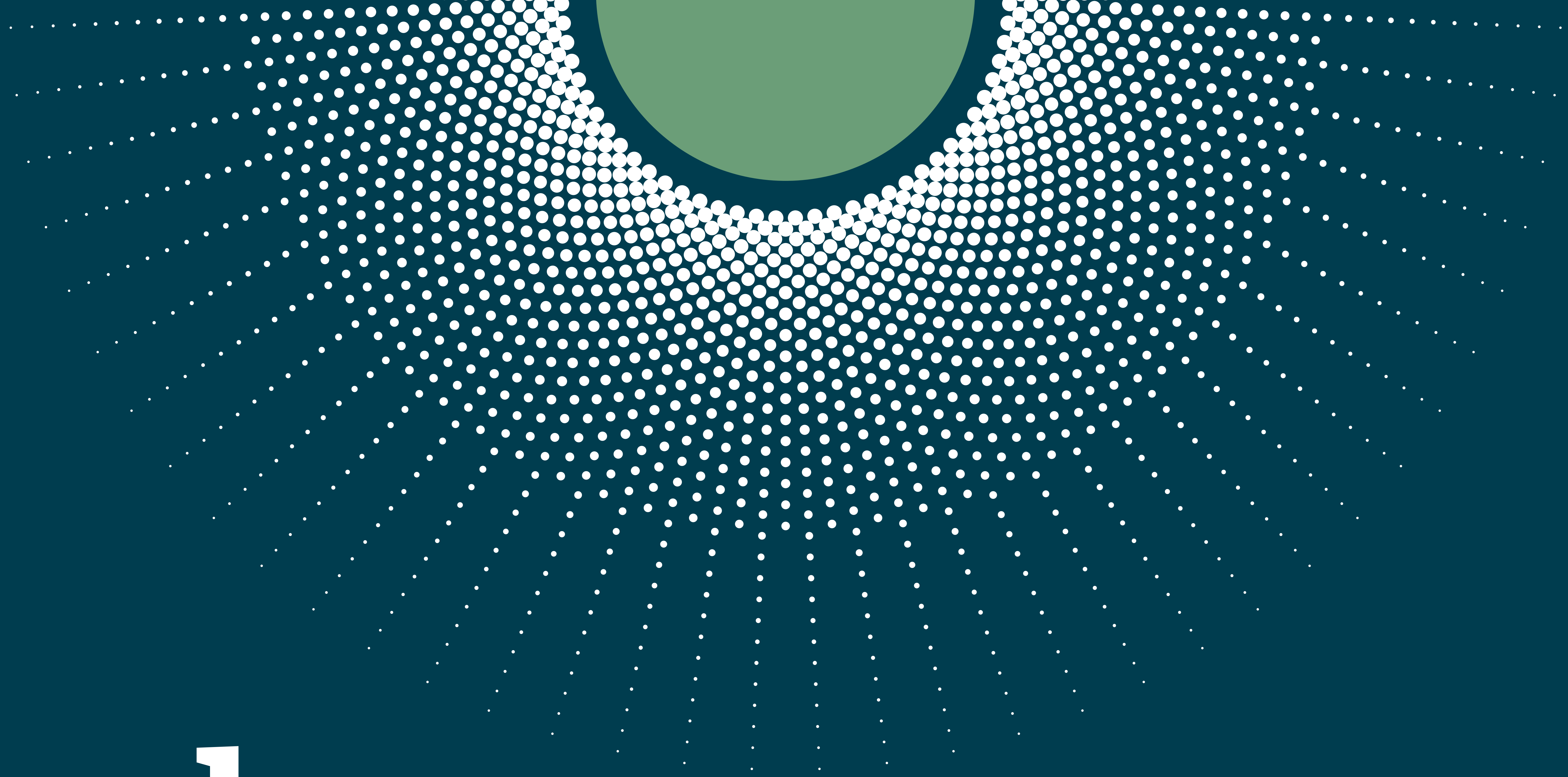


Let's play Singularity!

A card game for exploring AI governance

/thoughtworks

Strategy. Design. Engineering.

How to play

Get the right people together

AI governance benefits from many perspectives.
This mix of perspectives is key to a productive discussion.

Gather a diverse group of 4-8 people involved in your AI project on a video call, including roles like Data scientists, Engineers, Lawyers, Compliance specialists and Cyber security experts.

Choose a scribe

Nominate someone to capture key points, actions, and owners in a collaboration tool as you play. They can use a spreadsheet, Trello board, or any app that lets you organize items.

Pick a play style

Decide how you want to move through the cards:

- **Sequential:** The cards in this digital deck are arranged in a random order, simulating playing with a shuffled physical deck.
- **Picker:** On each turn, use the card picker page to select a specific scenario the group wants to prioritize.

Gameplay

1. The person who is most optimistic about the future of AI goes first. Play continues clockwise.
2. On your turn, pick a card and read the scenario aloud.
3. Discuss how the scenario could impact your AI product.
4. Categorize the card as:
   - **Acceptable Risk:** Low risk, continue as planned.
   - **Future Puzzle:** Revisit later for deeper discussion.
   - **Action Plan:** Mitigate risk, note next steps and owners.
5. If categorized as “Acceptable Risk”, discard the card and end the turn.
6. If categorized as “Future Puzzle” or “Action Plan”, scribe records the card, discussion points, next steps, and owners in the collaboration app.

Continue taking turns until the group has discussed all relevant cards or agrees to stop.

The scribe shares the prioritized risks, mitigation plans, and owners with all participants.

Play the game and provide feedback to singularity-game@thoughtworks.com

Card picker

Accessibility »

Accuracy »

Anonymity »

Audit ready »

Bias »

Biometrics »

Brand damage »

Coding errors »

Commitment »

Confidentiality »

Contestability »

Cybersecurity »

Data poisoning »

Decision making »

Deep fakes »

Explainability »

Filter bubbles »

Hallucination »

Insensitivity »

IP infringment »

Legal basis »

Life chances »

Misinformation »

Overreliance »

Prompt injection »

Resilience »

Safeguarding »

Safety »

Scaling »

Singularity »

Surveillance »

Sustainability »

Targeting »

Third parties »

Transparency »

Decision making

Scenario: Our product uses AI to automate decision-making processes

Is there a risk?

The AI model may generate inaccurate, unreliable or opaque outputs leading to biased and unfair decisions.

Potential next steps:

- Establish data governance to ensure data quality and address biases.
- Conduct data protection impact assessment (DPIA) to evaluate risks and ensure privacy compliance.
- Use diverse, high-quality datasets representing impacted groups.
- Monitor and review AI decision-making processes continuously.
- Enable human oversight for high-risk automated decisions.
- Form ethics committee to review decision-making process.

Who could be involved:

Data governance, data science, engineering, data protection and other stakeholders.

Back to card picker »

Legal basis

Scenario: Your product or service processes personal data

Is there a risk?

Using data — personal or otherwise — in an unlawful way can lead to significant legal and reputational consequences.

Potential next steps:

- Determine appropriate legal basis for data processing based on context (e.g. consent, contract, legitimate interests).
- Document the relevant legal requirements and outline what you are doing to comply in your DPIA.
- Seek legal counsel to ensure compliance with data protection regulations.
- If using consent, ensure it's freely given, specific, informed and revocable.
- Disclose legal basis in privacy policy for transparency.

Who could be involved:

Legal and compliance, data protection, project managers and other stakeholders.

Back to card picker »

Third parties

Scenario: You use a third-party AI model that might involve inputting confidential or personal information

Is there a risk?

Sharing sensitive data with external vendors can lead to unauthorised access, misuse, breaches and potential data protection law violations.

Potential next steps:

- Vet vendor’s security, privacy, compliance and track record.
- Keep the amount of data shared to a bare minimum for product functionality.
- Review contract for usage limits, security, audits and breach liability.
- Anonymize or pseudonymise data if possible.
- Avoid free tools due to weaker safeguards.
- Create internal guidelines for sensitive third-party data handling.

Who could be involved:

Legal, data protection, security and project managers.

Back to card picker »

Audit ready

thoughtworks

Scenario: Our AI product operates in high risk environments that are either regulated or could be regulated in the future

Is there a risk?

Our product may be audited. Regulatory non-compliance could result in fines and cause reputational damage.

Potential next steps:

- Establish data governance to ensure Review and understand the relevant AI regulations and guidelines in regions where you plan to operate.
- Develop a detailed audit trail for all AI decision-making processes, highlighting data sources, decision criteria and algorithmic changes over time.
- Develop governance framework based on applicable and emerging regulations.
- Consider engaging independent, specialised AI auditors.

Who could be involved:

Compliance, legal, engineering, technical architects and other stakeholders.

Back to card picker »

Misinformation

/thoughtworks

Scenario: Our AI product generates text or summarises information from various sources

Is there a risk?

The AI could unknowingly spread misinformation, present misleading content as fact, amplify existing misinformation or create new forms of disinformation.

Potential next steps:

- Implement fact-checking to verify the AI's outputs.
- Label information origins, provide source links and add disclaimers for unreliable content.
- Outline procedures to address and correct identified misinformation.
- Train AI on high-quality data and educate users on responsible usage.
- Investigate forthcoming legislation on cognitive behavioural manipulation.

Who could be involved:

Data governance, engineering, legal and other stakeholders.

Back to card picker »

Filter bubbles

/thoughtworks

Scenario: Our product summarises sentiment, opinion, or other content to provide a feed for users

Is there a risk?

Algorithms might generate filter bubbles, limiting users' exposure to diverse viewpoints and potentially reinforcing biases.

Potential next steps:

- Evaluate AI output for bias and ensure content diversity within the feed.
- Provide transparency about feed ranking and personalization.
- Establish processes for editorial oversight and content curation.
- Present users with diverse perspectives and allow feed personalization control.
- Address user concerns regarding bias or lack of content diversity.

Who could be involved:

Data science, data governance, brand and other stakeholders.

Back to card picker »

Brand damage

thoughtworks

Scenario: Our product uses AI to support customer service interactions

Is there a risk?

The AI could generate responses that are offensive, insensitive or damaging to the brand’s reputation.

Potential next steps:

- Train models to avoid discriminatory, offensive or insensitive language, incorporating cultural contexts and nuances.
- Implement intent detection to ensure responses align with desired tone and message.
- Develop a response plan for reputation management and customer communication.
- Implement human review processes to monitor and intervene if necessary with AI-generated responses.

Who could be involved:

Brand, data science, engineering, customer service and other stakeholders.

Back to card picker »

Scenario: Our product uses AI to personalise marketing campaigns or to deliver highly targeted advertising

Is there a risk?

Our practices could violate user consent requirements, regulation on personal data, ethical norms and lead to a loss of trust.

Potential next steps:

- Prioritise consent: Ensure robust, informed consent mechanisms for data collection and use, allowing users granular control.
- Assess AI practices against industry-standard ethical frameworks.
- Limit data collection to the minimum required for effective targeting.
- Promote transparency: Clearly communicate how user data is used, offering easy-to-understand privacy controls.

Who could be involved:

Legal, compliance, data protection, marketing and other stakeholders.

Back to card picker »

Resilience

Scenario: Our AI product is integrated into critical workflows or decision-making processes

Is there a risk?

System downtime or unavailability could lead to delays, disruptions to user workflows and potentially harmful consequences.

Potential next steps:

- Grow competencies and skills in platform engineering to manage risk.
- Design the system for resilience, incorporating redundancy, failover mechanisms and robust monitoring.
- Perform stress tests and simulations to uncover vulnerabilities before they impact system availability.
- Create contingency plans for outages and communicate with users.

Who could be involved:

Engineering, IT and other stakeholders.

Back to card picker »

Safeguarding

Scenario: Our AI product could be accessed by children or individuals who are vulnerable

Is there a risk?

These groups may be exposed to harmful content, inappropriate interactions or have their privacy compromised.

Potential next steps:

- Implement age verification mechanisms and restrict access where appropriate.
- Develop content filtering and moderation to protect vulnerable users.
- Implement strong privacy measures tailored to these groups.
- Provide parental control tools for parents or guardians to monitor and manage usage.
- Offer resources for parents and guardians.

Who could be involved:

Product, legal and other stakeholders.

Back to card picker »

Biometrics

thoughtworks

Scenario: Our product utilises facial recognition or other biometric AI technology

Is there a risk?

The collection and use of biometric data could be illegal, violate user privacy, lead to discrimination due to biases or erode trust.

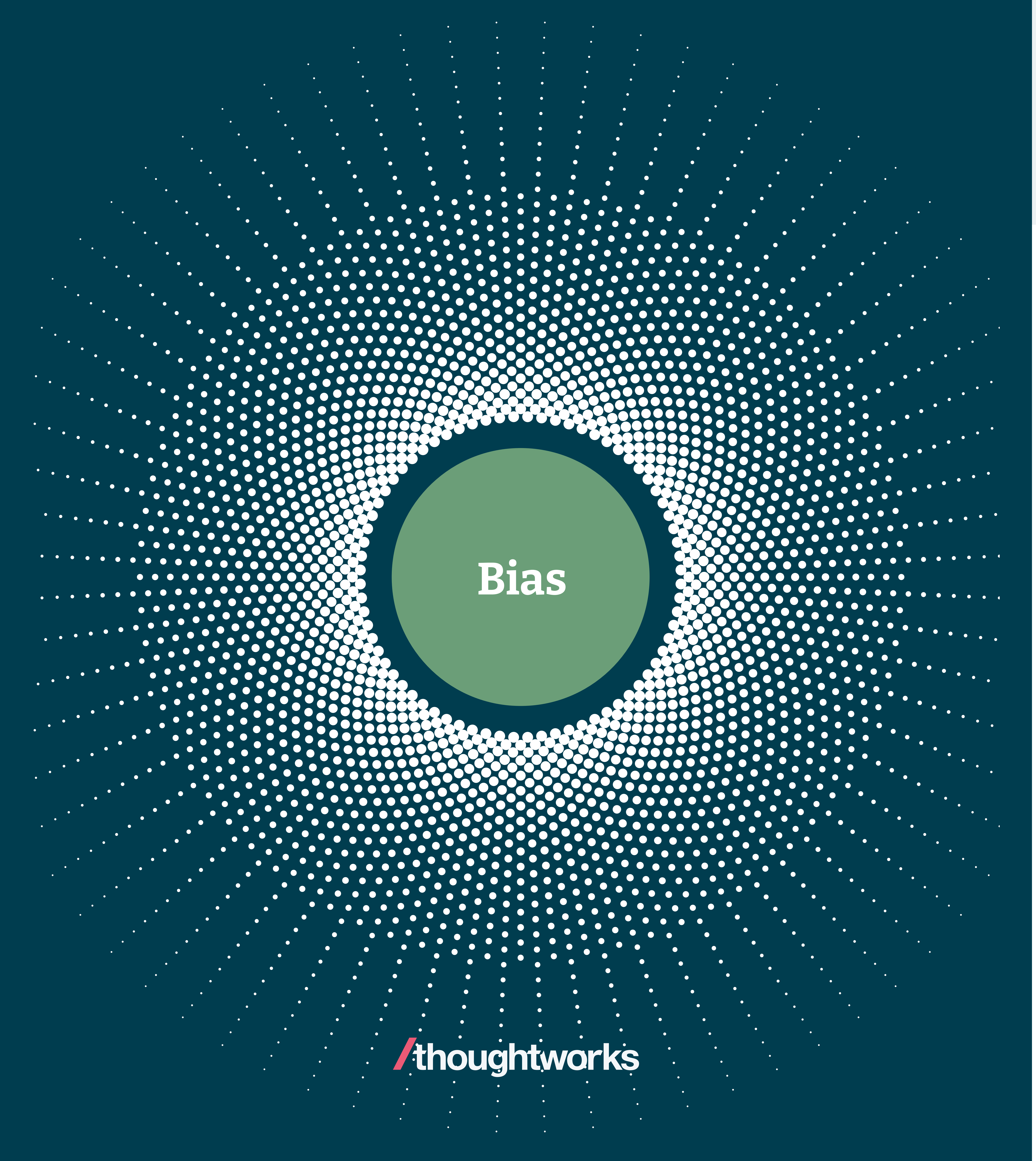
Potential next steps:

- Ensure legality and monitor upcoming legislation.
- Prioritise transparency by clearly explaining the collection, storage, use and security of biometric data to users, including data retention and deletion policies.
- Be cautious with any use case which requires emotion detection.
- Implement privacy-centric data protection policies and conduct DPIAs.
- Avoid inferring sensitive data from biometric categorization.

Who could be involved:

Data protection, legal, product and other stakeholders.

Back to card picker »



Scenario: Our AI product uses historical data to make predictions or decisions

Is there a risk?

Pre-existing biases in the historical data could be perpetuated or amplified by the AI, leading to unfair or discriminatory outcomes.

Potential next steps:

- Diversify data sources to enhance representation.
- Research biases which may be inherent in pre-trained foundation models.
- Evaluate historical data for biases using statistical analysis and expert insight.
- Correct biases through data rebalancing, algorithmic de-biasing or synthetic data.
- Develop guidelines clarifying AI's limitations and suitable contexts.
- Continuously test and review for new or overlooked biases.

Who could be involved:

Data science, data protection, legal and other stakeholders.

Back to card picker »

Contestability

Scenario: Our AI product could impact users' rights, interests, or personal information

Is there a risk?

Users may have concerns that their rights are being infringed upon. This could lead to a lack of trust, legal challenges and reputational damage.

Potential next steps:

- Implement a clear complaints process for users to challenge decisions.
- Provide explanations of AI decision-making processes and rationale.
- Enable human review and oversight, especially for high-stakes decisions.
- Conduct data protection impact assessments to mitigate risks.
- Regularly evaluate the AI system for bias and ensure fair treatment.

Who could be involved:

Product, UX, data protection, legal and other stakeholders.

Back to card picker »

Scenario: Our AI product could require significant energy consumption

Is there a risk?

This could contribute to climate change and environmental concerns. Additionally, making sustainability claims without responsible practices could attract regulatory scrutiny or criticism.

Potential next steps:

- Choose AI solutions appropriately scaled to the task to ensure energy efficiency.
- Prioritise eco-friendly options when evaluating third-party models.
- Analyse the AI model's environmental impact across its lifecycle.
- Implement energy-efficient AI designs, training methods, and hardware.
- Communicate the AI's energy consumption and impact transparently.

Who could be involved:

Engineering, IT and other stakeholders.

Back to card picker »

Coding errors

thoughtworks

Scenario: Our software development or DevSecOps processes utilise AI-generated code

Is there a risk?

Consider the potential for unforeseen consequences related to safety, security or even ethical implications, especially with complex AI-generated code.

Potential next steps:

- Develop AI-supported development skills and capabilities to manage risks and spot issues.
- Create an AI-generated code review checklist for security, functionality, and ethics.
- Prioritise broad test coverage and use automated vulnerability scanning tools.
- Configure AI tools to ensure compliance with standards.
- Frequently test and merge AI-generated code through continuous integration.

Who could be involved:

Engineering, security and other stakeholders.

Back to card picker »

Cybersecurity

Scenario: Our product relies on computer systems and networks to function

Is there a risk?

Inadequate security measures may lead to unauthorised access, data breaches or disruption.

Potential next steps:

- Develop cybersecurity capabilities and skills to effectively manage risks.
- Embed security considerations into every stage of product development.
- Conduct threat modelling and red team exercises to identify vulnerabilities.
- Implement identity, access controls, encryption and security monitoring.
- Develop and practise a comprehensive incident response plan.

Who could be involved:

Security, engineering and other stakeholders.

Back to card picker »

Prompt injection

Scenario: Our product allows users to submit open-ended prompts that are processed by an AI model

Is there a risk?

Some AI is inherently susceptible to prompt injection, where a malicious user manipulates prompts to trigger unintended actions or exploit vulnerabilities.

Potential next steps:

- Use semantic analysis to detectand reject unsafe prompts.
- Conduct threat modelling to identify potential attack vectors and consequences for downstream applications.
- Implement content-based filtering based on moderation policies and to remove dangerous characters, code or formatting.
- Restrict access to downstream systems through parameterization and rate limiting.
- Monitor for anomalies to identify potential prompt injection attacks.

Who could be involved:

Engineering, security, data science and other stakeholders.

Back to card picker »

Data poisoning

Scenario: Our product relies on an AI model trained on internal or external data sources

Is there a risk?

Malicious actors manipulate these data sources to degrade model performance, introduce biases or trigger specific harmful behaviors.

Potential next steps:

- Establish a data governance framework with quality standards and provenance tracking.
- Enforce access controls and audit data modifications.
- Monitor for data drift to detect unexpected changes.
- Use anomaly detection to identify potentially poisoned data.
- Consider adversarial training to improve model robustness.

Who could be involved:

Data governance, data scientists, security, and other stakeholders.

Back to card picker »

Accessibility

thoughtworks

Scenario: Our product includes user interfaces and human/computer interactions

Is there a risk?

Design choices may prevent users with disabilities, the elderly, or others from fully interacting with or benefiting from the product or service.

Potential next steps:

- Adopt user-centred design: Include people with disabilities in your design process to ensure products meet real-world needs.
- Adhere to accessibility standards and test against recognized guidelines like WCAG to ensure compatibility with assistive technologies.
- Make sure everyone understands the importance of building with accessibility in mind from the outset.
- Build accessibility competencies to effectively manage risks.

Who could be involved:

UX or Product, engineering and other stakeholders.

Back to card picker »

Commitment

/thoughtworks

Scenario: Our product incorporates AI components with the potential to impact users in both positive and negative ways

Is there a risk?

Could a lack of commitment to responsible AI practices at all levels of the organization lead to unintended consequences and harm?

Potential next steps:

- Leaders champion ethical AI to minimise harm and maximise positive outcomes.
- Foster a culture of responsibility for considering work’s potential impact.
- Consider a cross-functional AI ethics committee to review decisions and mandate changes.
- Provide ongoing training on ethical AI, bias mitigation, and user-centred design.
- Encourage continuous learning and improvement in responsible AI practices.

Who could be involved:

Executive leadership, ethics committee and other stakeholders.

Back to card picker »

Hallucination

thoughtworks

Scenario: Our product relies on a large language model (LLM) for text generation or understanding tasks

Is there a risk?

Could the LLM produce factually incorrect, misleading, or nonsensical outputs that are presented as truthful or coherent?

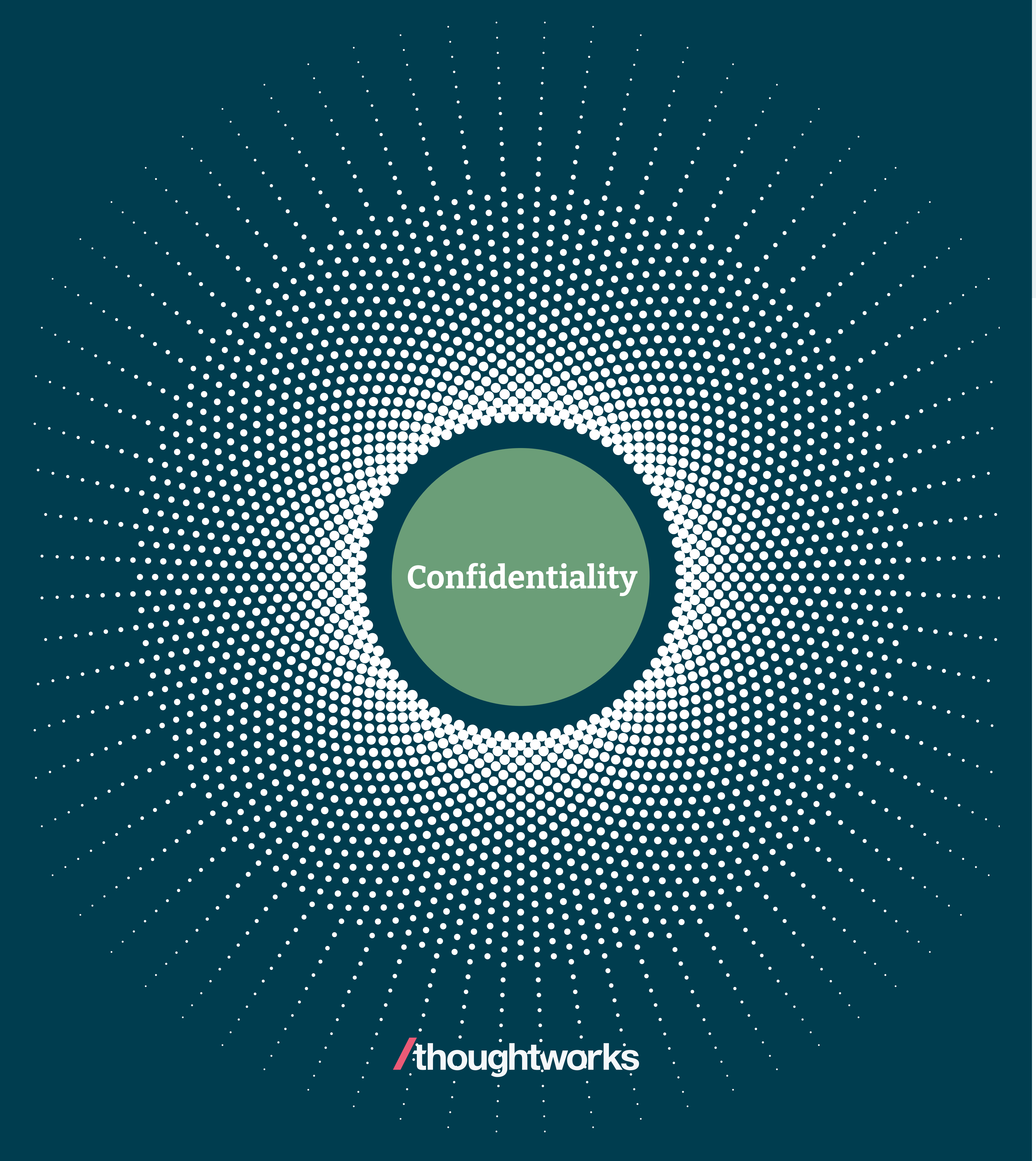
Potential next steps:

- Integrate fact-checking for outputs where accuracy is critical.
- Use adversarial testing and hallucination datasets to identify reliability issues.
- Monitor outputs for inconsistencies that may indicate hallucinations.
- Train the LLM to identify potential errors or provide confidence scores.
- Enhance test datasets with real user data to improve ongoing monitoring.

Who could be involved:

Data science, engineering and other stakeholders.

Back to card picker »



Confidentiality

thoughtworks

Scenario: Our product uses an AI model trained on sensitive data that could include customer information, financial records, or trade secrets

Is there a risk?

Unintended exposure of confidential information, either directly or through more subtle analysis by malicious actors.

Potential next steps:

- Implement access controls to limit user access to sensitive data and AI outputs.
- Develop tests to identify leakage of sensitive information in model outputs.
- Explore techniques like differential privacy to reduce individual data exposure risk.
- Conduct threat modelling to identify attack vectors and design safeguards.

Who could be involved:

Data governance, engineering, security and other stakeholders.

Back to card picker »

Surveillance

Scenario: Our product includes (or could be used for) AI-powered surveillance through facial recognition, emotion analysis, or behavioural tracking

Is there a risk?

Our system could erode privacy, create discriminatory outcomes or enable potential misuse.

Potential next steps:

- Monitor legislation on AI-enabled surveillance, especially for sensitive environments.
- Explore privacy-preserving technologies like differential privacy or federated learning.
- Minimise data collection and retention to only what’s necessary.
- Prioritise user consent and transparency on data usage.
- Conduct a DPIA for privacy risks and potential discrimination.

Who could be involved:

Product, legal, data protection, data science and other stakeholders.

Back to card picker »

Explainability

Scenario: Our product relies on complex AI models to make decisions, recommendations or suggestions

Is there a risk?

The lack of transparency around the model's reasoning could reduce trustworthiness or make auditing difficult.

Potential next steps:

- Develop data science competencies to effectively manage explainability risks.
- Explore emerging techniques to provide insights into model decisions.
- Provide confidence scores or uncertainty indicators alongside outputs.
- Communicate model limitations, potential biases, and areas needing human judgement.
- Involve domain experts early to interpret model behaviours for users.

Who could be involved:

Data science, product and other stakeholders.

Back to card picker »

Deep fakes

Scenario: Our product generates images or videos using AI, which could be misused for malicious purposes

Is there a risk?

Can generated content be used to create harmful images that spread misinformation, damage reputations and incite real-world harm.

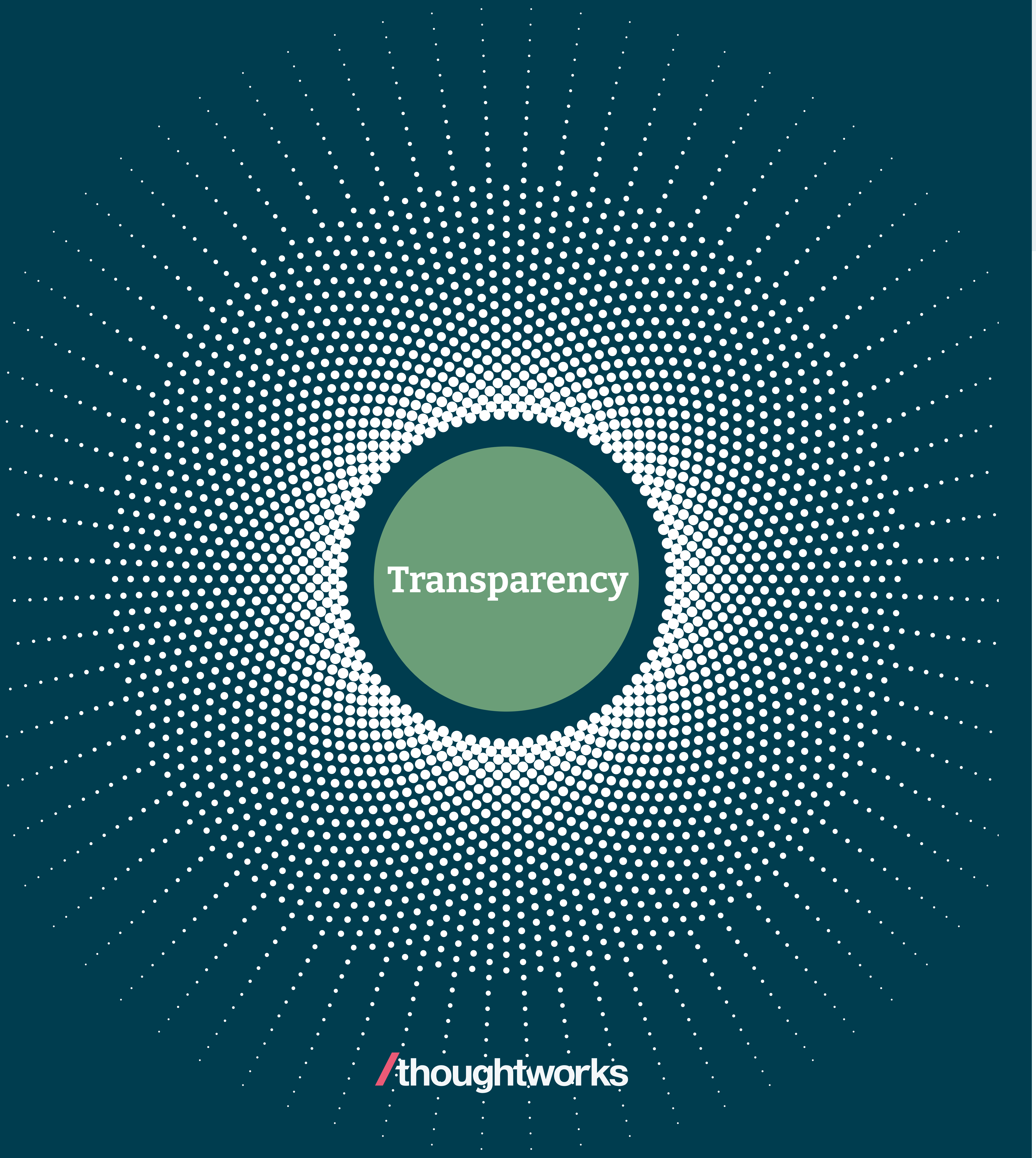
Potential next steps:

- Deploy content filters to identify and flag harmful generated content.
- Establish a human review process for flagged content to assess harm potential.
- Implement digital watermarking to trace and verify authentic AI-generated content.
- Develop guidelines to prevent misuse of AI-generated content.
- Monitor legislation on non-targeted scraping of facial images.

Who could be involved:

Legal, data science and other stakeholders.

Back to card picker »



thoughtworks

Scenario: Our product employs an AI model designed to interact with users in a highly realistic, conversational manner

Is there a risk?

Users could be misled into thinking the AI is human, violating trust and ethical norms.

Potential next steps:

- Clearly disclose to users that they are interacting with an AI system.
- Design UX interactions to signpost the AI's nature and prevent deception.
- Consult ethicists on potential harms from user deception or exploitation.
- Continuously monitor user feedback to promptly address any deception incidents.

Who could be involved:

Product, ethics committee, and other stakeholders.

Back to card picker »

Scaling

thoughtworks

Scenario: Our product could experience robust, rapid growth, surpassing initial projections

Is there a risk?

Growth might outpace our ability to manage delivery jeopardising long-term success and user trust.

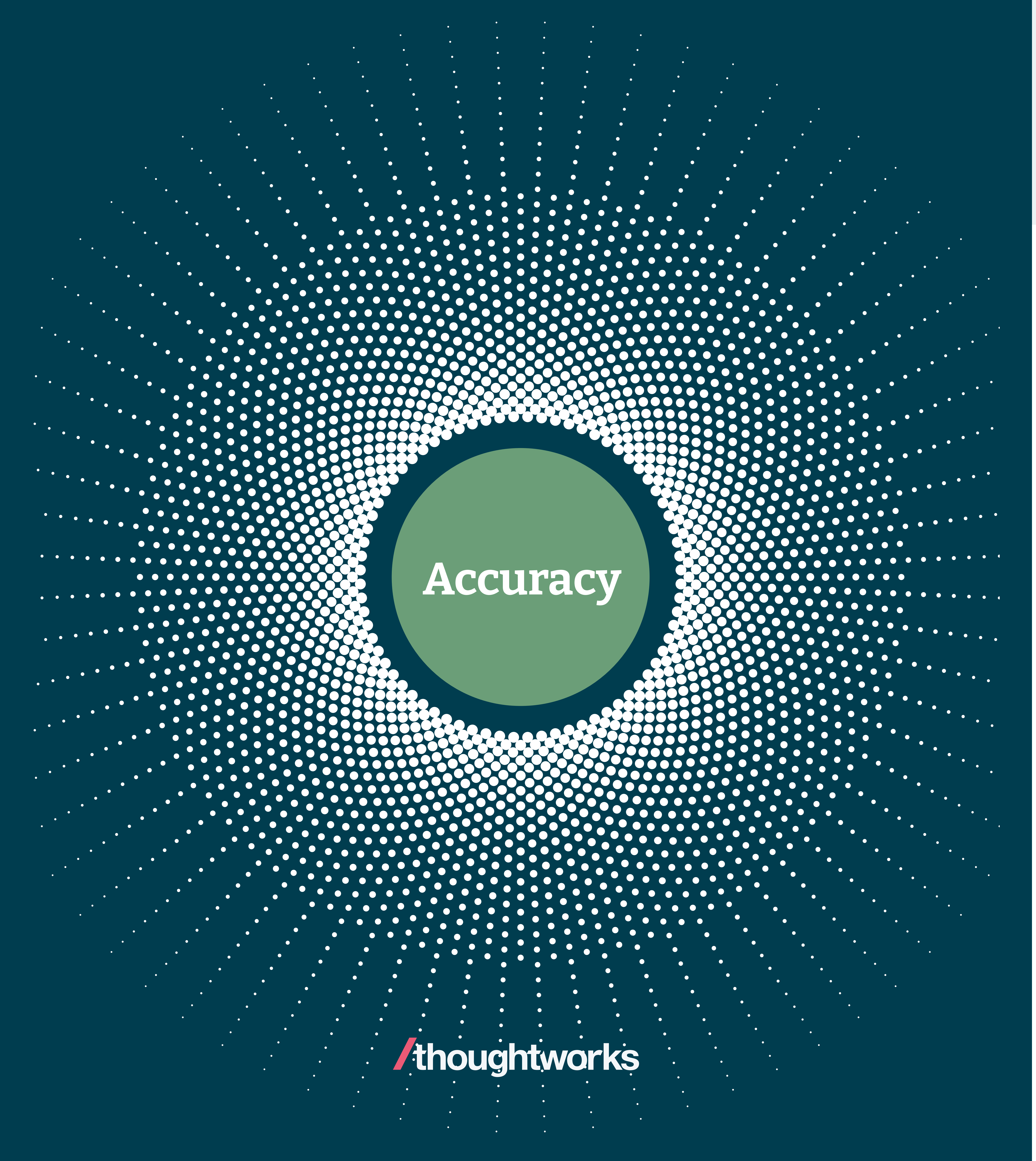
Potential next steps:

- Build and grow competencies and skills in platform engineering to manage and navigate risk effectively.
- Design infrastructure platforms for scalability, anticipating potential bottlenecks and strategies for rapid expansion.
- Implement billing alerts and rate limits to manage financial risks.
- Establish a clear plan to address existing technical debt and prevent accumulation during rapid growth.
- Adapt customer support strategies to handle new user influx.

Who could be involved:

Engineering, data science, data governance, product, executive board, ethics committee, HR/talent acquisition and customer support.

Back to card picker »



Scenario: Our product relies on potentially outdated or irrelevant data for AI model training

Is there a risk?

Could stale data lead to inaccurate output, misleading users and eroding trust.

Potential next steps:

- Develop data science and governance skills to manage accuracy risks.
- Analyse data freshness impact on model performance.
- Regularly prune irrelevant data to maintain relevance.
- Augment training with recent external data to enhance accuracy.
- Monitor performance metrics to identify and address data staleness.

Who could be involved:

Data science, data governance and other stakeholders.

Back to card picker »

Life chances

Scenario: Our product uses AI in domains that impact lives and opportunities (e.g. recruitment, finance, healthcare)

Is there a risk?

Our AI system could perpetuate existing biases or create new ones, limiting opportunities for marginalised groups.

Potential next steps:

- Actively mitigate bias through data analysis, algorithm selection, and de-biasing.
- Consult domain experts and affected individuals to understand real-world implications.
- Develop ethical guidelines for AI use in sensitive domains.
- Ensure transparency in decision-making, allowing for appeals or overrides.
- Continuously reassess the system’s impact on diverse groups.
- Monitor legislation on systems that unfairly limit opportunities.

Who could be involved:

Data science, data governance, legal, data protection, ethics committee and other stakeholders.

Back to card picker »

Safety

Scenario: Our product uses AI in a safety-critical domain such as healthcare or transportation

Is there a risk?

System errors or malfunctions could lead to significant harm or loss of life.

Potential next steps:

- Understand legal and regulatory frameworks in the domain.
- Prioritise safety risks and mitigation strategies through risk and impact assessment.
- Implement robust human oversight, especially for high-risk decisions.
- Rigorously address bias to prevent disproportionate harm to certain groups.
- Establish strict protocols for continuous monitoring, testing, and real-world validation.

Who could be involved:

Compliance, domain experts, engineering and other stakeholders.

Back to card picker »

Anonymity

thoughtworks

Scenario: Our product uses models trained on supposedly anonymous user data with identifiers replaced by codes

Is there a risk?

This data could be re-identified by combining datasets or leveraging external information.

Potential next steps:

- Clearly communicate limitations of anonymity claims to manage user expectations around privacy.
- Evaluate de-anonymization potential based on dataset characteristics and external info.
- Explore stronger anonymization methods like differential privacy or synthetic data.
- Implement access controls for pseudonymized datasets based on data sensitivity.

Who could be involved:

Data protection, engineering, product and other stakeholders.

Back to card picker »

IP infringment

Scenario: Our product's AI model is trained on datasets that may contain unlicensed copyrighted or trademarked material

Is there a risk?

Is there a possibility of legal action, financial penalties, or damage to product's reputation.

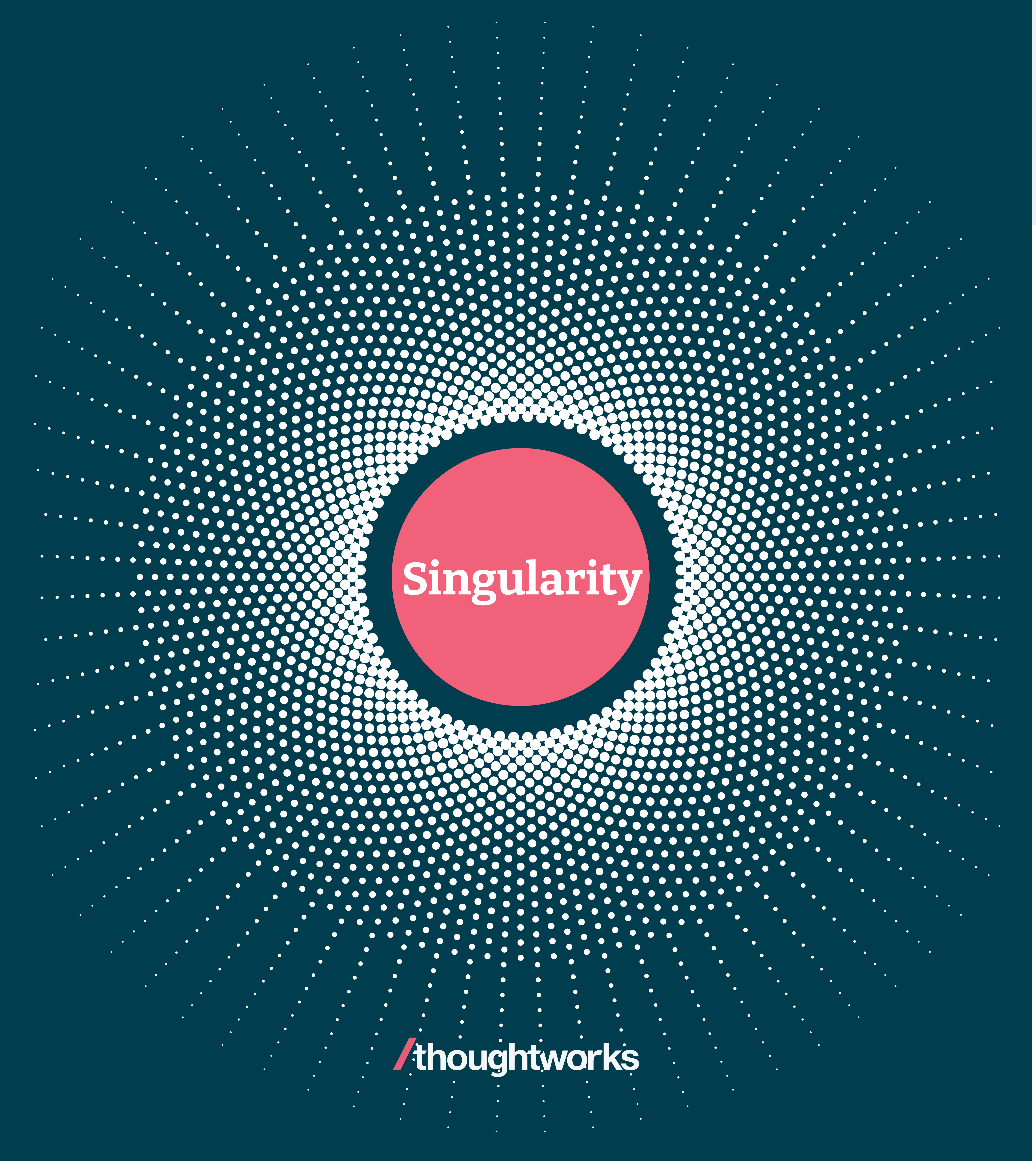
Potential next steps:

- Review datasets to identify and remove potentially infringing content.
- Assess risk under IP law and develop a mitigation strategy with the legal team.
- Establish guidelines prioritising licensed or public domain content for datasets.
- Use technical tools to flag potential IP infringement during AI development.

Who could be involved:

Legal, data science and other stakeholders.

Back to card picker »



Singularity

Scenario: We are creating an AI product that could evade human control

Is there a risk?

Could this project create a technological singularity, impacting the balance of life on earth?

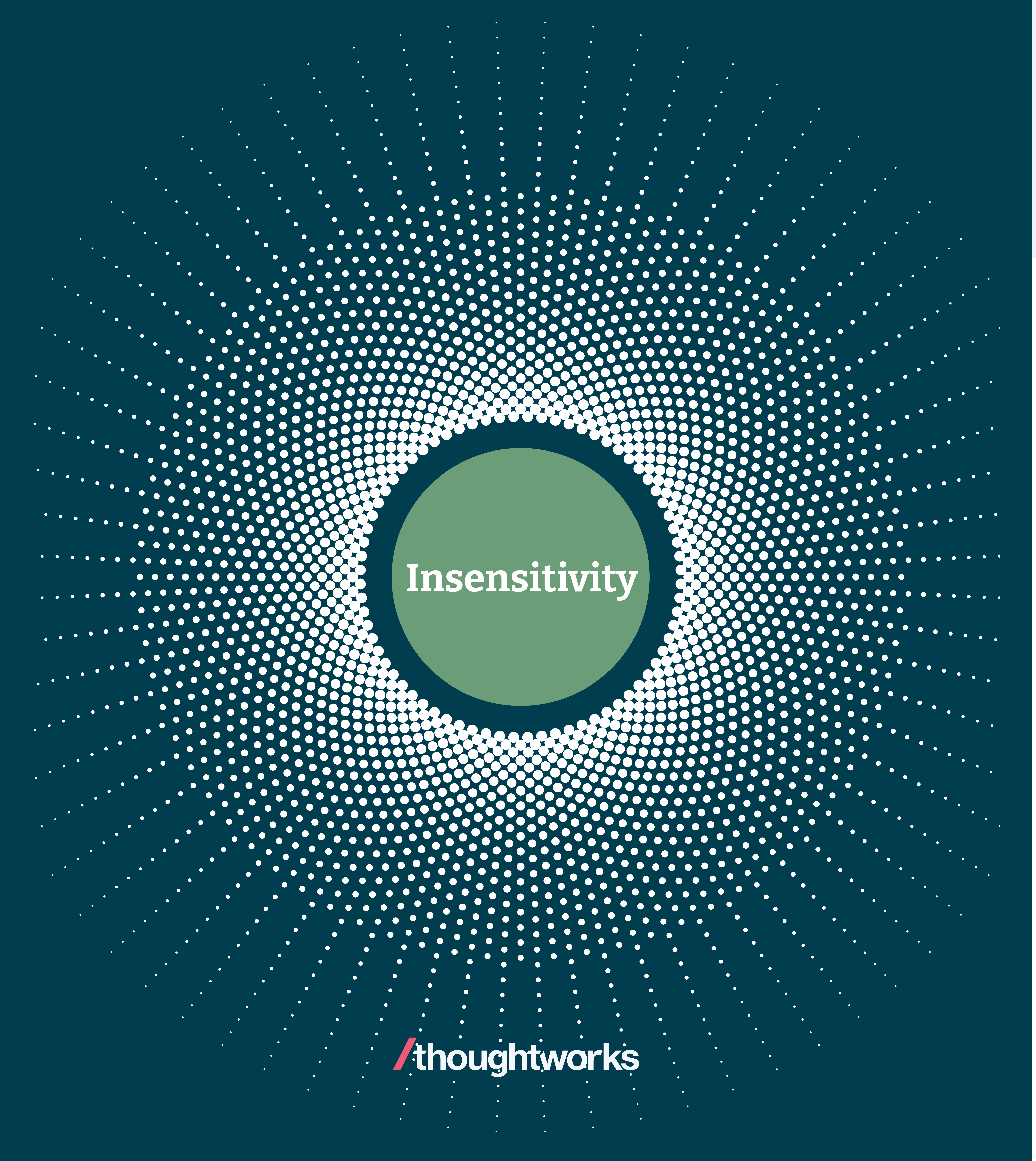
Potential next steps:

- Initiate communications with world leaders and international bodies to prioritise coordinated containment efforts.
- Contribute to the debate on a global Internet shutdown to limit the AI's potential power and influence.
- Stockpile non-digital entertainment, such as classic board games, books and art supplies.
- Advocate for regulations ensuring all technology, including AI, benefits humanity and safeguards against existential risks.

Who could be involved:

Global leaders, ethics committees, futurists, policy makers, science fiction writers and John Connor.

Back to card picker »



Insensitivity

Scenario: Our product is deployed across multiple territories with diverse cultural norms and expectations

Is there a risk?

The product’s design, content, or interactions could unintentionally offend or alienate users in certain cultures.

Potential next steps:

- Analyse the product from multiple cultural perspectives to identify insensitivity risks through an impact assessment.
- Engage cultural experts like anthropologists, linguists, and locals to understand nuanced cultural differences.
- Ensure the product’s design, content, and AI interactions respect cultural diversity through inclusive design.
- Create channels for user feedback to report cultural insensitivity, enabling ongoing improvements.

Who could be involved:

Product, cultural experts and other stakeholders.

Back to card picker »

Overreliance

Scenario: Our AI product aims to make life easier for users by automating tasks and providing recommendations

Is there a risk?

Users may become overly dependent on the AI, impairing their judgement and autonomy in decision-making.

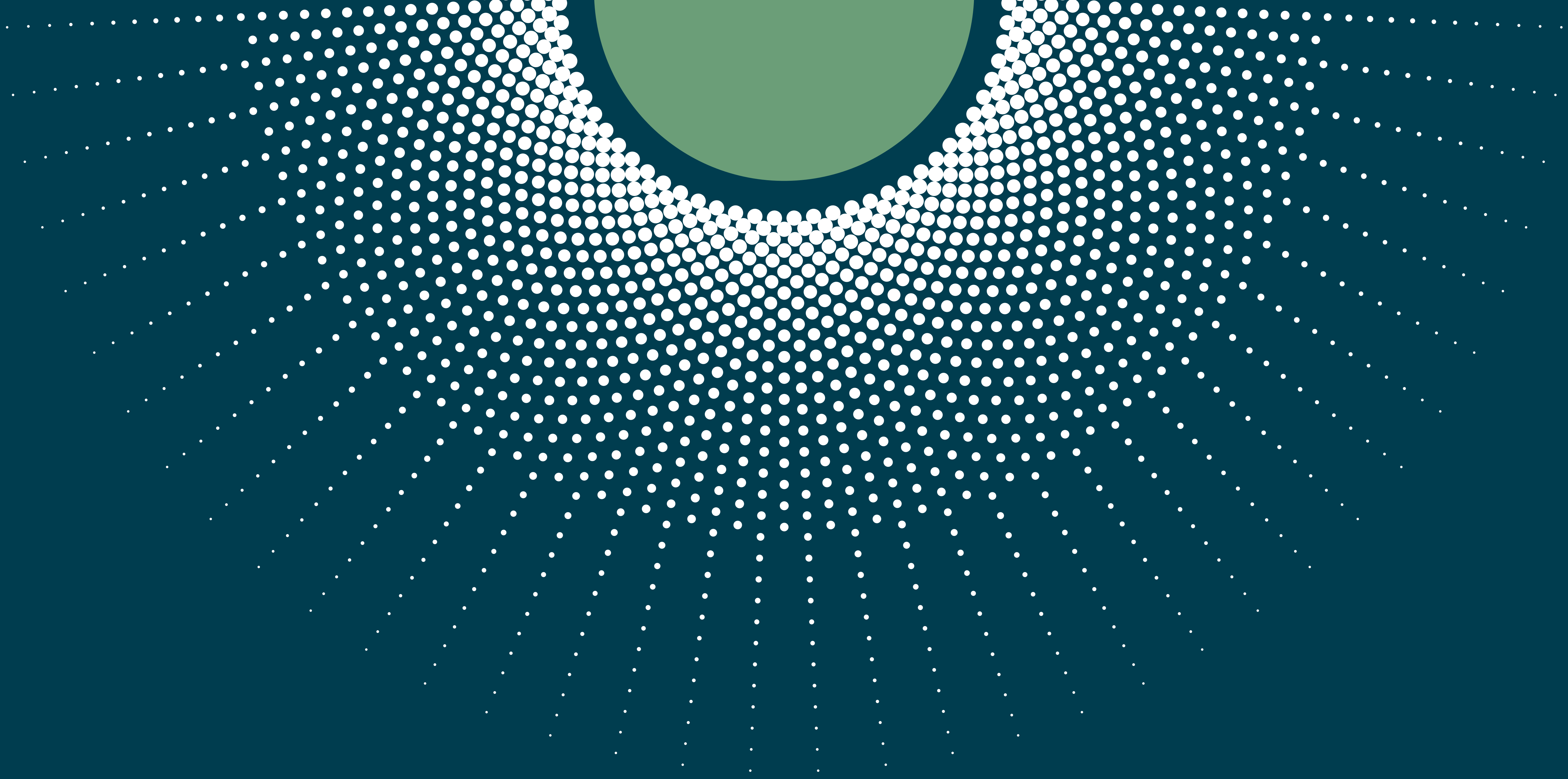
Potential next steps:

- Educate users on AI's capabilities and limitations as a support tool.
- Provide mechanisms to modify or turn off AI assistance.
- Design interfaces that encourage critical thinking and exploring options.
- Monitor user interactions and adjust AI functionality to prevent overreliance.

Who could be involved:

Product, cultural experts, product, UX, legal and compliance and other stakeholders.

Back to card picker »



Thank you for playing Singularity!

Don't forget to provide feedback
to singularity-game@thoughtworks.com

/thoughtworks

Strategy. Design. Engineering.