TRUSTWORTHY AI GUIDELINES

Trustworthy AI means AI systems that are safe, fair, transparent, secure, and aligned with human values.
These guidelines apply to all AI systems across the organisation.

1. Accountability

Purpose: Someone must always be responsible for AI decisions.

Guidelines:

- Every AI system must have a named Model Owner.
- Human oversight must exist for all high-impact decisions.
- Decision-making boundaries must be documented (what AI can/cannot decide).
- Audit trails must record who reviewed and approved each stage.
- No AI system is allowed to run without an accountable person/team.

2. Transparency

Purpose: Users must understand how AI works and where it is being used.

Guidelines:

- Users must be informed when they interact with AI.
- AI limitations must be clearly stated.
- Provide model explanations where required (why a decision was made).
- Documentation (Model Card, Data Card, Risk Assessment) is mandatory.
- All AI decisions should be traceable.

3. Fairness & Non-Discrimination

Purpose: AI must not unfairly disadvantage any group.

Guidelines:

- Evaluate fairness metrics across demographics.
- Remove or limit use of harmful sensitive attributes.
- Prevent historical or human bias from entering training data.
- Perform fairness testing before deployment and regularly after.
- Document fairness limitations and mitigation steps.

4. Safety & Reliability

Purpose: AI should work as intended and avoid harmful outcomes.

Guidelines:

- Define acceptable performance thresholds.
- Conduct stress, robustness, and failure testing.
- Include fallback mechanisms for model unavailability.
- High-risk outputs must have Human-In-The-Loop review.
- Block unsafe outputs automatically where needed (LLM guardrails).

5. Privacy & Data Protection

Purpose: AI must protect personal data and follow legal requirements.

Guidelines:

- Use only the minimum data required.
- Apply anonymization, masking, or pseudonymisation.
- Validate legal basis for using personal data (PDPA/GDPR).
- Prevent use of sensitive data unless absolutely necessary.
- Protect training data from being recreated through model attacks.

6. Security & Robustness

Purpose: AI systems must resist adversarial attacks.

Guidelines:

- Protect against input attacks (prompt injection, adversarial inputs).
- Secure training data from poisoning.
- Prevent model extraction and model theft (rate limits, watermarking).
- Monitor suspicious API access and anomalies.
- Conduct red-teaming and adversarial testing regularly.

7. Explainability

Purpose: AI decisions must be understandable to humans when needed.

Guidelines:

- Use SHAP/LIME/Feature Importance for structured models.
- Provide simple explanations for end users where required.
- High-risk decisions require explainability by default.
- Document what can/cannot be explained.
- Maintain consistency in explanations across versions.

8. Human Oversight & Control

Purpose: Humans must remain in charge, especially for critical decisions.

Guidelines:

- Define when manual review is mandatory.
- Provide override mechanisms at all times.
- Enable human intervention for safety incidents.
- Train staff on interpreting AI outputs.
- No fully autonomous decisions for legal, financial, or life-impacting areas.

9. Monitoring & Incident Response

Purpose: AI must be continuously monitored during its lifecycle.

Guidelines:

- Monitor bias, drift, performance, security anomalies.
- Maintain an AI Audit Log for all incidents.
- Reassess risk when data or model changes.
- Define thresholds for alerts and shutdown triggers.
- Conduct quarterly reviews; recertify annually.

10. Environmental & Social Responsibility

Purpose: AI must contribute positively and avoid societal harm.

Guidelines:

- Evaluate environmental cost (energy, compute).
- Avoid applications that can be misused for harm.
- Prevent misinformation, manipulation, or unsafe content output.
- Consider broader societal impact during design.

11. Third-Party & Vendor AI Governance

Purpose: External AI tools must meet the same standards.

Guidelines:

- Conduct vendor AI risk assessments.
- Request documentation (Model Card, security controls).
- Verify privacy, security, and ethical safeguards.
- Restrict high-risk third-party AI without approval.
- Monitor vendor compliance regularly.

12. Lifecycle Governance

Purpose: Trustworthy AI is not a one-time activity; it requires continuous governance.

Guidelines:

- Apply governance at every stage:
    - Intake
    - Assessment
    - Development
    - Testing
    - Approval
    - Deployment
    - Monitoring
- Update controls as model evolves.
- Re-evaluate high-risk models annually.