Microsoft's Guidelines for Human-AI Interaction

Best Practices for Interactive AI Behavior

Based on Microsoft Research CHI 2019 Paper

Executive Summary

Microsoft Research has developed a comprehensive set of 18 evidence-based guidelines for designing AI systems that interact appropriately with users. These guidelines address critical aspects of human-AI interaction across four key phases: initial interaction, regular interaction, error handling, and long-term adaptation. The guidelines synthesize insights from multiple research disciplines including human-computer interaction, machine learning, and user experience design.

Introduction

As artificial intelligence becomes increasingly integrated into everyday applications and services, the quality of human-AI interaction has become paramount to user adoption and satisfaction. Microsoft's research team conducted extensive studies to identify patterns of successful AI behavior and common failure modes, resulting in these 18 comprehensive guidelines.

The guidelines are designed to help developers, designers, and product managers create AI systems that are not only functionally effective but also intuitive, trustworthy, and user-friendly. They represent a synthesis of empirical research, user studies, and real-world deployment experiences.

The 18 Guidelines for Human-AI Interaction

Phase 1: Initially (Upon First Interaction)

G1: Make Clear What the System Can Do

- **Principle**: Clearly communicate the AI system's capabilities and scope of functionality
- **Implementation**: Provide upfront disclosure of what tasks the AI can perform well
- **Example**: A voice assistant should explicitly state it can help with weather, music, smart home controls, etc.
- **Rationale**: Sets appropriate user expectations and prevents frustration from unmet assumptions

G2: Make Clear How Well the System Can Do What It Can Do

- **Principle**: Communicate the AI system's confidence levels and accuracy expectations
- **Implementation**: Provide information about reliability, typical performance metrics, or uncertainty ranges
- **Example**: A translation app might indicate "90% accuracy for common phrases, lower for technical terms"
- **Rationale**: Helps users calibrate their trust and usage patterns appropriately

Phase 2: During Interaction (Regular Use)

G3: Show Contextually Relevant Information

- **Principle**: Display information that is pertinent to the current user context and task
- **Implementation**: Filter and prioritize information based on user goals, location, time, and previous interactions
- **Example**: A calendar AI showing only today's meetings when asked "what's my schedule?"
- **Rationale**: Reduces cognitive load and improves task efficiency

G4: Match Relevant Social Norms

- **Principle**: Behave in ways that align with established social conventions and cultural expectations
- **Implementation**: Adapt communication style, politeness levels, and interaction patterns to social context
- **Example**: A customer service AI using formal language in professional contexts, casual language in personal apps
- **Rationale**: Creates natural, comfortable interactions that feel socially appropriate

G5: Mitigate Social Biases

- **Principle**: Actively work to reduce unfair treatment based on protected characteristics
- **Implementation**: Regular bias testing, diverse training data, fairness-aware algorithms
- **Example**: Resume screening AI that doesn't discriminate based on names suggesting gender or ethnicity

- **Rationale**: Ensures equitable treatment and maintains user trust across diverse populations

G6: Support Efficient Invocation

- **Principle**: Make it easy and quick for users to engage the AI system
- **Implementation**: Provide multiple input modalities, keyboard shortcuts, and streamlined interfaces
- **Example**: Voice activation, gesture recognition, or single-click access for common tasks
- **Rationale**: Reduces friction in human-AI interaction and encourages regular use

G7: Support Efficient Dismissal

- **Principle**: Allow users to easily disengage from or stop the AI system
- **Implementation**: Clear exit options, cancellation commands, and override mechanisms
- **Example**: "Stop," "Cancel," or "Never mind" commands that immediately halt AI processing
- **Rationale**: Gives users control and prevents frustrating experiences when AI misunderstands intent

G8: Support Efficient Correction

- **Principle**: Enable users to quickly and easily correct AI mistakes or misunderstandings
- **Implementation**: Simple correction interfaces, learning from corrections, undo functionality
- **Example**: "I meant [correction]" or thumbs up/down feedback that immediately improves responses
- **Rationale**: Maintains user productivity even when AI makes errors

G9: Scope Services When in Doubt

- **Principle**: When uncertain about user intent, ask clarifying questions rather than guessing
- **Implementation**: Present options, ask for confirmation, or request additional information
- **Example**: "Did you mean weather in Seattle, WA or Seattle, OR?" instead of assuming
- **Rationale**: Prevents errors and demonstrates the AI's understanding of its limitations

G10: Do Not Take Actions Beyond Scope

- **Principle**: Avoid performing actions outside the AI's defined capabilities or user permissions
- **Implementation**: Clear boundaries on what actions are automated vs. require user confirmation
- **Example**: An AI assistant asking permission before sending emails or making purchases
- **Rationale**: Maintains user trust and prevents unintended consequences

Phase 3: When Wrong (Error Handling)

G11: Make Clear Why the System Did What It Did

- **Principle**: Provide explanations for AI decisions and actions, especially when they seem unexpected
- **Implementation**: Transparency in reasoning, decision logs, or simplified explanations of AI logic
- **Example**: "I recommended this restaurant because it matches your preference for Italian food and has high ratings"
- **Rationale**: Builds understanding and trust, enables users to provide better feedback

G12: Remember Recent Interactions

- **Principle**: Maintain context from recent conversations and interactions
- **Implementation**: Session memory, conversation history, and contextual awareness
- **Example**: Understanding pronouns like "it" or "that" referring to previously mentioned items
- **Rationale**: Creates more natural, coherent interactions that feel conversational

G13: Learn from User Behavior

- **Principle**: Adapt and improve based on individual user patterns and preferences
- **Implementation**: Personalization algorithms, preference learning, behavioral adaptation
- **Example**: Learning that a user prefers brief responses and adjusting communication style accordingly
- **Rationale**: Improves user experience over time through personalization

G14: Update and Adapt Cautiously

- **Principle**: Make changes to AI behavior gradually and with user awareness
- **Implementation**: Incremental updates, user notifications of changes, opt-in for major modifications
- **Example**: Notifying users when a recommendation algorithm has been updated and why
- **Rationale**: Maintains user trust and prevents confusion from sudden behavioral changes

Phase 4: Over Time (Long-term Adaptation)

G15: Encourage Granular Feedback

- **Principle**: Provide multiple ways for users to give specific, actionable feedback
- **Implementation**: Rating systems, specific correction options, detailed feedback forms
- **Example**: Allowing users to specify what aspect of a response was good or bad (accuracy, tone, completeness)
- **Rationale**: Enables targeted improvements and gives users sense of agency in AI development

G16: Convey the Consequences of User Actions

- **Principle**: Help users understand how their inputs and feedback will affect future AI behavior
- **Implementation**: Clear explanations of how feedback is used, preview of changes
- **Example**: "Rating this response as helpful will make me more likely to give similar answers in the future"
- **Rationale**: Empowers users to actively shape their AI experience

G17: Provide Global Controls

- **Principle**: Offer system-wide settings and preferences that users can adjust
- **Implementation**: Privacy controls, interaction preferences, feature toggles
- **Example**: Settings to adjust response length, formality level, or types of proactive suggestions
- **Rationale**: Gives users control over their overall AI experience

G18: Notify Users About Changes

- **Principle**: Inform users when the AI system's capabilities, behavior, or policies change
- **Implementation**: Change logs, notifications, explanations of updates
- **Example**: "We've updated our language model to better understand technical questions"
- **Rationale**: Maintains transparency and helps users adapt to system evolution

---

Implementation Framework

Assessment Questions for Each Guideline

For Developers and Designers:

1. **Capability Communication**: Does your AI clearly explain what it can and cannot do?
2. **Performance Transparency**: Do users understand the AI's accuracy and limitations?
3. **Contextual Relevance**: Is information filtered appropriately for user context?
4. **Social Appropriateness**: Does the AI behave according to social norms?
5. **Bias Mitigation**: Are there systems in place to prevent discriminatory behavior?
6. **Efficient Access**: Can users easily invoke the AI when needed?
7. **Easy Exit**: Can users quickly disengage from the AI?
8. **Simple Correction**: Is it easy to fix AI mistakes?
9. **Uncertainty Handling**: Does the AI ask for clarification when unsure?
10. **Scope Adherence**: Does the AI stay within appropriate boundaries?
11. **Explainability**: Can the AI explain its reasoning?
12. **Context Memory**: Does the AI remember recent interactions?
13. **Learning Capability**: Does the AI adapt to user preferences?

14. **Cautious Updates**: Are changes made gradually with user awareness?
15. **Feedback Mechanisms**: Can users provide specific, actionable feedback?
16. **Action Consequences**: Do users understand how their input affects the AI?
17. **User Control**: Are there global settings and preferences available?
18. **Change Notification**: Are users informed about system updates?

Design Process Integration

Phase 1: Design and Development

- Use guidelines G1-G2 to design initial user onboarding
- Implement G3-G10 as core interaction patterns
- Build G11-G14 into error handling and learning systems
- Plan G15-G18 for long-term user relationship management

Phase 2: Testing and Validation

- Test each guideline against user scenarios
- Measure user satisfaction and task completion rates
- Identify guideline conflicts and resolve through user research
- Validate accessibility and inclusivity across diverse user groups

Phase 3: Deployment and Iteration

- Monitor adherence to guidelines through analytics
- Collect user feedback on guideline effectiveness
- Iterate on implementations based on real-world usage
- Update guidelines based on new research and user needs

Case Studies and Applications

Virtual Assistants

- **G1-G2**: Clear communication of supported commands and accuracy rates
- **G6-G7**: Wake words for invocation, clear dismissal commands
- **G12-G13**: Contextual awareness and preference learning
- **G17**: Privacy and interaction preference controls

Recommendation Systems

- **G3**: Contextually relevant suggestions based on time, location, and activity
- **G5**: Bias mitigation in content recommendations
- **G11**: Explanations for why specific items were recommended

- **G15**: Granular feedback (like/dislike, too expensive, wrong genre, etc.)

Automated Customer Service

- **G4**: Appropriate formality and professionalism
- **G9**: Clarifying questions when customer intent is unclear
- **G10**: Escalating to human agents for complex issues
- **G16**: Explaining how feedback improves service quality

Content Creation AI

- **G8**: Easy editing and revision of AI-generated content
- **G14**: Cautious updates to writing style and capabilities
- **G18**: Notifications about new features or model improvements

---

Measuring Success

Key Performance Indicators (KPIs)

User Satisfaction Metrics

- **Trust Rating**: User-reported trust levels in AI system
- **Ease of Use**: Task completion rates and user effort scores
- **Error Recovery**: Success rate of error correction attempts
- **Long-term Engagement**: Retention and usage growth over time

Technical Performance Metrics

- **Accuracy**: Task completion accuracy and error rates
- **Response Time**: Speed of AI responses and user feedback processing
- **Personalization Effectiveness**: Improvement in recommendations over time
- **Bias Mitigation**: Fairness metrics across different user groups

Behavioral Indicators

- **Correction Frequency**: How often users need to correct AI mistakes
- **Feature Discovery**: Rate at which users discover and adopt new capabilities
- **Feedback Quality**: Specificity and actionability of user feedback
- **Setting Utilization**: Usage of control and customization features

Evaluation Methods

Quantitative Assessment

- **A/B Testing**: Compare implementations with and without specific guidelines
- **Analytics Tracking**: Monitor user interaction patterns and success rates
- **Longitudinal Studies**: Track user satisfaction and behavior over time
- **Benchmark Comparisons**: Measure against industry standards and competitors

Qualitative Assessment

- **User Interviews**: Deep dive into user experiences and pain points
- **Observational Studies**: Watch how users interact with AI in natural settings
- **Focus Groups**: Gather feedback on specific guideline implementations
- **Expert Reviews**: Have UX professionals evaluate against guidelines

---

Common Implementation Challenges

Technical Challenges

Explainability vs. Simplicity

- **Challenge**: Providing explanations that are both accurate and understandable
- **Solution**: Layered explanations with simple summaries and detailed optional information
- **Example**: "Recommended because of your preferences" with option to see detailed reasoning

Real-time Adaptation

- **Challenge**: Learning from user behavior without disrupting ongoing interactions
- **Solution**: Background learning with periodic, gentle updates to behavior
- **Example**: Gradual adjustment of response style based on user feedback patterns

Cross-platform Consistency

- **Challenge**: Maintaining consistent AI behavior across different devices and interfaces
- **Solution**: Centralized preference management and synchronized learning
- **Example**: Voice assistant maintaining same personality and knowledge across phone, smart speaker, and car

User Experience Challenges

Expectation Management

- **Challenge**: Users may expect human-level understanding and capabilities
- **Solution**: Clear, upfront communication about limitations with concrete examples

- **Example**: "I can help with basic math, but complex calculus may require specialized tools"

Feedback Fatigue

- **Challenge**: Users may become tired of providing feedback
- **Solution**: Intelligent feedback requests and passive learning from user behavior
- **Example**: Only asking for explicit feedback when confidence is low, learning from clicks and time spent

Privacy Concerns

- **Challenge**: Balancing personalization with user privacy
- **Solution**: Transparent data practices and granular privacy controls
- **Example**: Clear explanation of what data is stored and how it's used, with easy deletion options

---

Future Considerations

Emerging Trends

Multimodal Interaction

- Integration of voice, gesture, and visual inputs
- Consistent guideline application across modalities
- Cross-modal context maintenance

Emotional Intelligence

- Recognition and appropriate response to user emotional states
- Empathetic communication while maintaining boundaries
- Cultural sensitivity in emotional expression

Collaborative AI

- Multiple AI systems working together transparently
- Clear attribution of actions and decisions
- Coordinated learning and adaptation

Research Directions

Advanced Personalization

- More sophisticated user modeling
- Context-aware adaptation
- Predictive user needs assessment

Improved Explainability

- Better techniques for making AI decisions understandable
- Domain-specific explanation strategies
- Visual and interactive explanation methods

Bias Detection and Mitigation

- More sophisticated bias detection methods
- Real-time bias correction
- Inclusive design practices

Conclusion

Microsoft's 18 Guidelines for Human-AI Interaction provide a comprehensive framework for creating AI systems that are not only functionally effective but also intuitive, trustworthy, and user-centered. These evidence-based guidelines address the full lifecycle of human-AI interaction, from initial encounter through long-term relationship development.

Successful implementation of these guidelines requires careful consideration of user needs, technical constraints, and ethical implications. Organizations should view these guidelines not as rigid rules but as flexible principles that can be adapted to specific contexts and use cases.

The guidelines emphasize the importance of transparency, user control, and continuous improvement in AI systems. By following these principles, developers and designers can create AI experiences that enhance human capabilities while respecting human autonomy and dignity.

As AI technology continues to evolve, these guidelines will likely need updates and refinements. However, the core principles of clear communication, appropriate behavior, graceful error handling, and thoughtful adaptation will remain fundamental to successful human-AI interaction.

The ultimate goal is to create AI systems that feel like helpful, reliable partners rather than unpredictable tools. By implementing these guidelines thoughtfully and consistently, we can build AI experiences that truly serve human needs and foster long-term trust and satisfaction.

References and Further Reading

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
- Microsoft Research. (2019). Guidelines for Human-AI Interaction. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf
- Norman, D. A. (2013). The design of everyday things: Revised and expanded edition. Basic books.
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human-Computer Studies, 100, 102118.
- Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 159-166).