Digital Ethics Preliminary Screening Checklist

This checklist is meant to spark reflection and conversation among teams developing or deploying AI in clinical, education, or research settings. It adapts the World Health Organization's framework to our institutional priorities and is designed as an early-stage screen—not a full ethics review.

If you mark **No** or **Unclear** for more than two items in any section, we recommend pausing implementation and contacting the Digital Ethics Consult Service for further guidance.

Some items on the checklist are marked with an asterisk (*); these questions indicate critical safeguards. A **No** response should trigger an immediate pause and ethics consultation. To seek an ethics consultation please contact DIRECT (Digital Innovation and Research Ethics Consultation Team) at: ResearchHelp@URMC.Rochester.edu and Jonathan.herington@rochester.edu

For each item below, check the appropriate column and add notes as needed.

1. Protect Autonomy (Privacy, Consent, Control)

	Yes	No	Unclear	Notes
*Are patients/users aware when interacting with AI vs. a human?	☐	☐	☐	
*Is there a clear pathway for human override? i.e. Does AI support—not replace—clinical, academic, or administrative decision-making?	☐	☐	☐	
*Is there a clear pathway for patients/users to opt out of AI-driven decisions or data use?	☐	☐	☐	
*Do patients/users understand how their data is being used?	☐	☐	☐	*Disclosure* in an EULA may not be sufficient, and the usage of any high-risk data (i.e. PHI) should be disclosed prominently to data subjects.

Is data collection limited to only those data points required to complete the intended task(s) or project? (i.e. you are not collecting data "just because")	☐	☐	☐	Often there is a temptation to collect all feasible data points that might be useful in a secondary research project, but it is often more appropriate to limit collection to only those data points required for the primary project.
Is the data retention period defined and limited?	☐	☐	☐	
Is there potential for secondary reuse of the data from this project and, if so, did patients/users consent to this?	☐	☐	☐	*Research* projects are required to disclose the possibility of secondary reuse of data in information and consent documents.

2. Promote Well-being, Safety, and the Public Interest

	Yes	No	Unclear	Notes
* Is there a plan to formally assess the system's *safety* (i.e. risk of harm to patients/users, including cybersecurity vulnerabilities) in its deployment context?	☐	☐	☐	
* Is there a plan to formally assess the system's *efficacy* (i.e. performance at the intended task) in its deployment context?	☐	☐	☐	
Is there a plan that assesses and addresses the *risks of unintended* use? (e.g. through red-teaming)	☐	☐	☐	
Does the tool *provide medical advice* or a therapeutic intervention to patients?	☐	☐	☐	
Does the tool *predict or diagnose stigmatized behaviors* or conditions (e.g. mental health outcomes, substance abuse, and sexual health)?	☐	☐	☐	

Does the tool *target interaction by vulnerable or marginalized status* (e.g. children, older adults, race, religion, gender identity)?	☐	☐	☐	

3. Ensure Transparency, Explainability, and Intelligibility

	Yes	No	Unclear	Notes
*****Are tool outputs (e.g. predictions, decisions, chat logs) recorded and reviewed regularly?	☐	☐	☐	
Is *documentation* available from developer/vendor describing tool behavior, training data, and performance?	☐	☐	☐	
Is the *output* presented in a form understandable to its intended audience (clinicians, patients, administrators)?	☐	☐	☐	
Is there a mechanism for recording and reviewing the *explanations* for the system's outputs? (i.e. feature importance metrics or other explainability techniques)	☐	☐	☐	

4. Ensure Inclusiveness and Equity

	Yes	No	Unclear	Notes
***** Does the AI use "sensitive attributes" (i.e race, ethnicity, gender, language, ability, and socioeconomic status) to make predictions?	☐	☐	☐	
Were clinicians, patients and affected communities consulted about fairness and inclusivity during design?	☐	☐	☐	

Are potential biases in training data documented and mitigated?	☐	☐	☐	
Was the AI’s performance evaluated in relevant sub-populations? i.e. those sensitive attributes (e.g. race, gender, age, language, ability, and socioeconomic status) there is a risk for bias.	☐	☐	☐	The specific sub-population analyses required will be dependent on the tool and task. (e.g. language groups may be relevant for speech-to-text tools, insurance status for healthcare cost prediction)
Will the AI tool / deployment of AI tool be available to all regardless of race, ethnicity, gender, age, language, ability, and socioeconomics?	☐	☐	☐	

5. Foster Responsibility and Accountability

	Yes	No	Unclear	Notes
*Is accountability for AI outputs clearly assigned?	☐	☐	☐	_Is there a human who has ultimate responsibility for the use and oversight of the tool’s outputs?_
*Is there a process for logging, reporting, and acting on adverse events or unexpected behavior from the tool?	☐	☐	☐	
Have stakeholders (incl. both users _and_ primary data subjects) been consulted during the design process?	☐	☐	☐	
Are there mechanisms for ongoing community accountability—not just one-time feedback—built into the AI’s lifecycle?	☐	☐	☐	
Is there a plan for ongoing updates, recalibration, and retraining of the AI system?	☐	☐	☐	