Ethical Guidelines – Some latest ones

Virpi Roto
Aalto University, Department of Design
Lead of RAAS Ethical, Acceptability and Desirability research task force

Ethical guidelines for AI coming out NOW

Published	Organization	Name of the document	Note
25.3.2019	IEEE	Ethically aligned design - A vision for prioritizing human well-being with autonomous and intelligent systems	Global, written as enforcement
8.4.2019	EU High-Level Expert Group on AI	Ethics guidelines for trustworthy AI (Made it to headlines)	Generic, EU (not a law - yet)
1.5.2019	Microsoft	Guidelines for Human-AI Interaction (Also more generic Microsoft AI principles exist)	Practical, for AI developers

More and more of scientific publications, e.g., in the new journal of Nature (incl. ethical, social and legal issues):

- Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1.
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM.

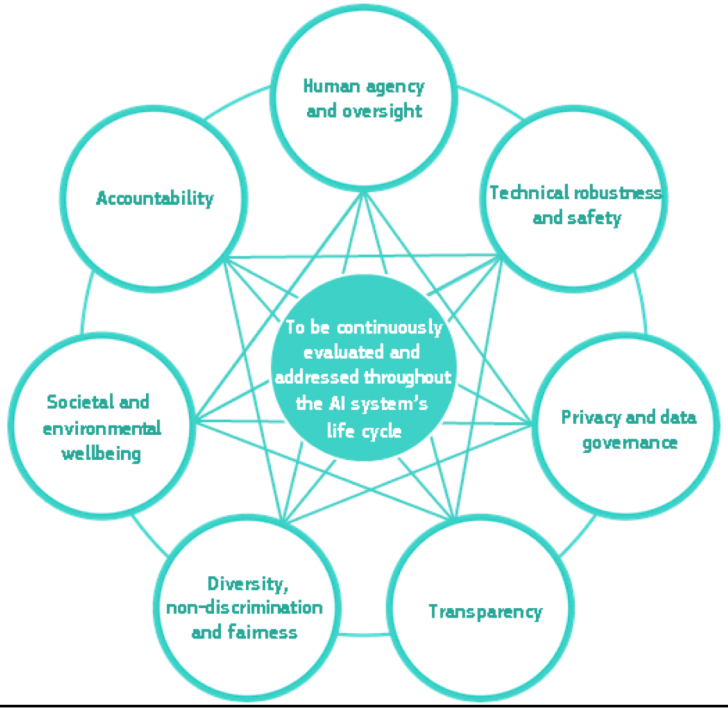
IEEE: Ethically aligned design

	Principle	Description (A/IS = Autonomous or Intelligent Systems)
1	Human Rights	A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2	Well-being	A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3	Data Agency	A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4	Effectiveness	A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5	Transparency	The basis of a particular A/IS decision should always be discoverable.
6	Accountability	A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7	Awareness of Misuse	A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8	Competence	A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

EU: Ethics guidelines for trustworthy AI

	Principle	Description (excerpts of longer descriptions)
1	Respect for human autonomy	…AI systems should be designed to augment, complement and empower human cognitive, social and cultural skills. … The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. …
2	Prevention of harm	AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. …
3	Fairness	The development, deployment and use of AI systems must be fair … free from unfair bias, discrimination and stigmatization. If unfair biases can be avoided, AI systems could even increase societal fairness. …
4	Explicability	...processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. … The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.

…EU: Ethics guidelines for trustworthy AI - Requirements



Microsoft: Guidelines for Human-AI Interaction

1/2

1	Initially	Make clear what the system can do	Help the user understand what the AI system is capable of doing
2		Make clear how well the system can do what it can do	Help the user understand how often the AI system may make mistakes
3	During interaction	Time services based on context	Time when to act or interrupt based on the user's current task and environment
4		Show contextually relevant information	Display information relevant to the user's current task and environment
5		Match relevant social norms	Ensure the experience is delivered in a way that users would expect, given their social and cultural context
6		Mitigate social biases	Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases
7	When wrong	Support efficient invocation	Make it easy to invoke or request the AI system's services when needed
8		Support efficient dismissal	Make it easy to dismiss or ignore undesired AI system's services
9		Support efficient correction	Make it easy to edit, refine, or recover when the AI system is wrong

Microsoft: Guidelines for Human-AI Interaction

2/2

10	...When wrong	Scope services when in doubt	Engage in disambiguation or gracefully degrade the AI system's services, when uncertain about a user's goals
11		Make clear why the system did what it did	Enable the user to access an explanation of why the AI system behaved as it did
12	Over time	Remember recent interactions	Maintain short-term memory and allow the user to make efficient references to that memory
13		Learn from user behavior	Personalize the user's experience by learning from their actions over time
14		Update and adapt cautiously	Limit disruptive changes when updating and adapting the AI system's behaviors
15		Encourage granular feedback	Enable the user to provide feedback indicating their preferences during regular interaction with the AI system
16		Convey the consequences of user actions	Immediately update or convey how user actions will impact future behaviors of the AI system
17		Provide global controls	Allow the user to globally customize what the AI system monitors and how it behaves
18		Notify users about changes	Inform the user when the AI system adds or updates its capabilities