Automated Decision-Making Technologies: Risks and Rewards

Imagine

A commodity exchange *where decisions are the commodity* traded.

Traders bet on risks or rewards going up or down in value.

Analysts watch for decision events which could impact entire investment portfolios or industries.

Politicians and Policymakers assess consumer sentiment prior to taking a position on issues.

Parents conduct their own background checks on childcare providers.

Patients look for alternative treatments or specific surgeons.

How Yo-ai weighs the benefits of using AI against the risks of using AI

Yo-ai's Registered Stakeholders evaluate the personal risks to them specifically – not risks to the economy, society, esoteric values, or, to the organizations which control the AI.

Yo-ai's Registered Stakeholders also evaluate the personal rewards to them specifically – not theoretical benefits to the economy, society, esoteric values, or, to the organizations which control the AI.

Often, when an organization's use of AI poses a risk to individual consumers, the stated benefit may be "to provide better service, reduce costs, increase profits, etc." In general, one could argue that consumers benefit when costs are lower, a business is more profitable, and can provide better service.

Registered Stakeholders want to know, "Are my costs lower? Are profits reinvested into improving the company's financial stability and services? Has the service they provide to me improved? Do I need to become an investor or shareholder to realize the rewards?"

Yo-ai conducts risk assessments and impact assessments that may answer some of these questions and help individual consumers make informed decisions about engaging with AI-enabled organizations.

Yo-ai asks first, "What are the risks to these specific Registered Stakeholders?"

Then Yo-ai searchs for possible corresponding benefits or rewards which may offset the risks.

Many organizations publish information about risks associated with AI.

Yo-ai uses the MIT AI Risk Repository as the standard taxonomy for classifying AI risks.

Currently, thirteen (13) of these risks are evaluated on the Yo-ai Platform.

Risks

1 *Discrimination & toxicity*

<u>1.1 Unfair discrimination and misrepresentation</u>

Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.

<u>1.3 Unequal performance across groups</u>

Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.

2 *Privacy & security*

<u>2.1 Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information</u>

AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or cause loss of confidential intellectual property.

<u>2.2 AI system security vulnerabilities and attacks</u>

Vulnerabilities that can be exploited in AI systems, software development toolchains, and hardware that results in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.

3 *Misinformation*

<u>3.1 False or misleading information</u>

AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional, or material harms.

4 *Malicious actors & misuse*

<u>4.1 Disinformation, surveillance, and influence at scale</u>

Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim of manipulating political processes, public opinion, and behavior.

<u>4.3 Fraud, scams, and targeted manipulation</u>

Using AI systems to gain a personal advantage over others through cheating, fraud, scams, blackmail, or targeted manipulation of beliefs or behavior. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.

5 *Human-computer interaction*

5.2 Loss of human agency and autonomy

Delegating by humans of key decisions to AI systems, or AI systems that make decisions that diminish human control and autonomy. Both can potentially lead to humans feeling disempowered, losing the ability to shape a fulfilling life trajectory, or becoming cognitively enfeebled.

6 *Socioeconomic & environmental harms*

6.1 Power centralization and unfair distribution of benefits

AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality.

6.2 Increased inequality and decline in employment quality

Social and economic inequalities caused by widespread use of AI, such as by automating jobs, reducing the quality of employment or producing exploitative dependencies between workers and their employers.

6.4 Competitive dynamics

Competition by AI developers or state-like actors in an AI "race" by rapidly developing, deploying, and applying AI systems to maximize strategic or economic advantage, increasing the risk they release unsafe and error-prone systems.

6.5 Governance failure

Inadequate regulatory frameworks and oversight mechanisms that fail to keep pace with AI development, leading to ineffective governance and the inability to manage AI risks appropriately.

7 *AI system safety, failures & limitations*

7.4 Lack of transparency or interpretability

Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability to identify and correct errors.

The MIT AI Risk Taxonomy was developed by searching documents on AI Risks.
"Due to our interest in broad, structured classifications of risks from AI, we excluded documents which focused only on risks from AI that are present in a single location or sector or that discussed risks specific to particular risk categories (e.g., content solely focused on different types of unfair decision-making) or very specific AI tools (e.g., content solely focused on risks from DALL-E).
…
We excluded anything which discussed sources of risk at a high level of abstraction (e.g., the sources of sociotechnical risk in AIs) or risk-assessment processes (e.g, how organizations can assess risks from AI) rather than focusing on classifying AI risks more specifically."

Yo-ai concentrates on unfair decision-making and how organizations assess AI risk.

Yo-ai enriches these risks by pairing them with corresponding rewards.

Rewards

Rewards are typically theoretical or aspirational. Although there is no taxonomy of AI Rewards, rewards are discovered from marketing claims and offers, legitimate business purposes for processing personal information, and formal rewards programs such as “loyalty programs”. Currently, there is insufficient quantifiable data on benefits to consumers receiving these rewards. The Yo-ai Platform is designed to help fill this information gap.

Many benefits attributed to the use of AI are described as ‘benefitting society’ or other broad-brush, macro-economic terms, that most individuals would find irrelevant to them personally. In fact, most benefits derived from AI are only to the entity who controls the AI, with little regard given to individuals who are adversely impacted.

Yo-ai conducts its assessments based on harms suffered by individual stakeholders, and the likelihood of rewards or benefits from using AI to individual stakeholders. One of the approaches used by Yo-ai is to showcase or construct positive models that support a common objective – in this case, distributing rewards to incentivize adoption and acceptance of new technologies like AI.

* * * * * * * * * The Yo-ai Rewards Program * * * * * * * * *

* Individual consumers can earn Reward Points for sharing information about harms that impacted them from specific organizations they subscribe to as Registered Stakeholders on the Yo-ai Platform.
* Monetary damages awarded by regulators, courts, or class action settlements to one Registered Stakeholder may be distributed equally among all Registered Stakeholders.
* Individual consumers can acquire additional skills, tools, and powers (privileges) as they increase their effective use of the Yo-ai Platform.
* Registered Stakeholders can become certified as Responsible AI practitioners to compensate for harm suffered from being displaced by employers’ AI initiatives.

* *

Risk Assessments on Automated Decision-making Technologies (ADMT)

Consumers who are positively or negatively impacted by ADMT can conduct their own risk assessments and submit them to other Stakeholders, including regulators, to offset false or misleading certifications and claims made on behalf of organizations. Yo-ai’s team of AI Agents takes care of the discovery, collection, and analysis of ADMT risks and rewards, but individual, Responsible Humans make the final determination about personal risk to themselves.

Using the Yo-ai Rewards Program as an example, let’s discover how Yo-ai’s AI Agents make autonomous, automated decisions about organizations: Is it explainable? Which components and vendors are included in the decision-making process? Is the data used as input, accurate and relevant? How was the model or algorithm trained, and what level of consistency can we expect from the output?

Yo-ai helps all Registered Stakeholders become Responsible Humans by controlling their own AI systems, learning how they work, and how they can cause harm in unexpected ways.

Yo-ai helps Organizations listen to Registered Stakeholders, who are not anonymous but do protect their own privacy as members of a larger group.

Finally, Yo-ai helps all stakeholders, registered or not, by giving consumers a voice with specificity and confidentiality, which includes employees of the organizations being assessed.