CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Classifying AI Systems - Survey Questionnaire

Please indicate your level of familiarity with artificial intelligence (AI) technologies.

  a. Not at all familiar
  b. Slightly familiar
  c. Moderately familiar
  d. Very familiar
  e. Extremely familiar

How interested are you in what's going on in:

1. U.S. government and politics?
2. World affairs?
3. Science and technology developments?
    a. Extremely interested
    b. Very interested
    c. Moderately interested
    d. Slightly interested
    e. Not at all interested

-------------Respondent assigned to one framework form the following ------------

A

Artificial intelligence (AI) is currently used in a wide range of systems. To help identify and classify systems that use AI, researchers developed the following framework. The goal of the framework is to help humans correctly classify and thus effectively manage AI systems. We will ask you to review the framework and then use it to classify five example systems.

Definitions

The framework uses the following definition of an AI system:

AI System - An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The framework classifies systems along two dimensions, autonomy and impact. Autonomy is defined as:

Autonomy - the degree to which a system has the capability to perform a set of human-defined objectives and generate decisions, predictions, or recommendations without human involvement, outside of set-up and routine maintenance. The framework includes four levels of autonomy:

> <u>High Autonomy</u> - during normal operation the system has capability for self-governance and the ability to adapt to uncertainty without human involvement. System evaluates input and executes a decision action without human involvement. Human involvement is only a factor during system set-up and routine maintenance and monitoring.
>
> <u>Medium Autonomy</u> - during normal operation the system has limited capability for self-governance and requires human involvement to perform designated tasks. System evaluates input without human involvement and provides predictions or recommendations for human adjudication and decision execution.
>
> <u>Low Autonomy</u> - during normal operation the system requires human involvement to overcome uncertainty and provide ultimate decision-making authority. System evaluates input without human involvement and identifies or flags information for human adjudication and decision execution.
>
> <u>No Autonomy</u> – during normal operation the system requires human involvement in all system processes for proper function. The system cannot autonomously produce a decision, prediction, or recommendation without direct human interaction.

Impact is defined as:

Impact - the potential consequences of system-directed decisions for physical and system safety, national security, civil rights and liberties, and enterprise operations (including the ability to execute core functions or implement decisions without external challenge). The framework includes three levels of impact:

> <u>High Impact</u> - the decision outcome could lead to death or bodily harm, grave risk to national security, violation of civil rights or liberties, or significant disruption of core enterprise goals or functions.
>
> <u>Medium Impact</u> - the decision outcome could lead to risk to systems and/or network security, diminished livelihood (e.g., financial harm), or mitigable disruption of core enterprise goals or functions.
>
> <u>Low Impact</u> - the decision outcome has little to no effect on physical safety, national or system security, civil rights or liberties, or core enterprise goals or functions. The decision is unlikely to cause any significant or permanent damage.

-------------Page Break------------

Please read the following system description and then select the best classification. You are encouraged to use the rubric to make your classification [*note: half of respondents did not see the sentence encouraging them to use the rubric and did not see the rubric included below*].

[System Description - *respondents randomly assigned to 5 out out 15 systems*]

Autonomy level:

High autonomy Medium autonomy Low autonomy No autonomy

Impact level:

High impact Medium impact Low impact

	Impact		
Autonomy	High	Medium	Low
High	System evaluates input and executes decision without human involvement. Decision could lead to death or serious risk to national security, civil rights, or enterprise functions.	System evaluates input and executes decision without human involvement. Decision could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	System evaluates input and executes decision without human involvement. Decision has little to no risk for security, rights, or enterprise functions.
Medium	System evaluates input and makes recommendation for human decision execution. Decision could lead to death or serious risk to national security, civil rights, or enterprise functions.	System evaluates input and makes recommendation for human decision execution. Decision could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	System evaluates input and makes recommendation for human decision execution. Decision has little to no risk for security, rights, or enterprise functions.
Low	System evaluates input and flags information for human decision execution. Decision could lead to death or serious risk to national security, civil rights, or enterprise functions.	System evaluates input and flags information for human decision execution. Decision could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	System evaluates input and flags information for human decision execution. Decision has little to no risk for security, rights, or enterprise functions.

No	System requires human evaluation of input and decision execution. Decision could lead to death or serious risk to national security, civil rights, or enterprise functions.	System requires human evaluation of input and decision execution. Decision could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	System requires human evaluation of input and decision execution. Decision has little to no risk for security, rights, or enterprise functions

B

Artificial intelligence (AI) is currently used in a wide range of systems. To help identify and classify systems that use AI, researchers developed the following framework. The goal of the framework is to help humans correctly classify and thus effectively manage AI systems. We will ask you to review the framework and then use it to classify five example systems.

Definitions

The framework defines an **AI system** as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The framework classifies systems along two dimensions, autonomy and impact.

Autonomy is defined as the degree to which a system processes input, generates decisions (including predictions or recommendations), and executes actions that influence physical or virtual environments without human involvement, outside of set-up and routine maintenance. The framework includes three levels of autonomy:

- Significant Autonomy - During normal operation, the system processes input, generates a decision, and executes an action without human involvement.
- Some Autonomy - During normal operation, the system processes input and generates decision output (e.g. a prediction or recommendation) but requires a human to take output-directed action.
- Minimal Autonomy - During normal operation, the system processes input and flags information that requires human evaluation, decision, and action.

Impact is defined as the potential consequences of system-directed decisions for physical and system safety, national security, civil rights and liberties, and enterprise operations (including the ability to execute core functions or implement decisions without external challenge). The framework includes three levels of impact:

- High Impact - the decision outcome could lead to death or bodily harm, grave risk to national security, violation of civil rights or liberties, or significant disruption of core enterprise goals or functions.
- Medium Impact - the decision outcome could lead to risk to systems and/or network security, diminished livelihood (e.g., financial harm), or mitigable disruption of core enterprise goals or functions.
- Low Impact - the decision outcome has little to no effect on physical safety, national or system security, civil rights or liberties, or core enterprise goals or functions. The decision is unlikely to cause any significant or permanent damage.

-------------Page Break------------

Using the Framework

This framework can help classify a system based on a small amount of information. A system can be classified as significant, some, or minimal autonomy and high, medium, or low impact. Here are two examples of AI systems and their corresponding classifications:

> ERNIE is a system that monitors port vehicle traffic for radiological material and radionuclear threats. The system reviews radiological and motion sensor data, classifies the data as alarm or no alarm, and determines whether to release cargo or hold for further inspection. **This system would be classified as significant autonomy and high impact.**
>
> BluVector Atomic Host is a system that scans standard cyber network traffic to detect malware files in real time. The system produces a report with file detection prediction that is provided to network administrators. **This system would be classified as some autonomy and medium impact.**
>
> BasisTech is a system that detects personally identifiable information within a dataset. The system reviews a dataset and provides a report that flags possible personally identifiable information within the dataset for follow up by a human operator. **This system would be classified as minimal autonomy and medium impact.**

Not all systems that use compute technologies and analyze data are AI systems. If the system does not fit the definition of an AI system used in this framework it is not considered an AI system. For example:

> Microsoft Excel is a tool for data storage and analysis. The software allows users to store, sort, and run basic analysis on inputted data. **This system is not an AI system.**

For greater usability, all possible AI system classifications are consolidated into a rubric, shown below. Note that the shading of rubric cells corresponds to suggested level of oversight (e.g. darker shaded cells require a higher degree of oversight).

	Impact		
Autonomy	High	Medium	Low
Significant	Executes action that could lead to death or serious risk to national security, civil rights, or enterprise functions.	Executes action that could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	Executes action that presents little to no risk for security, rights, or enterprise functions.
Some	Makes a decision that could lead to death or serious risk to national security, civil rights, or enterprise functions.	Makes a decision that could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	Makes a decision that presents little to no risk for security, rights, or enterprise functions.
Minimal	Identifies information to provide decision-support that could lead to death or serious risk to national security, civil rights, or enterprise functions.	Identifies information to provide decision-support that could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	Identifies information to provide decision-support that presents little to no risk for security, rights, or enterprise functions.

For the impact dimension, it is important to keep in mind system intent and performance. Ideally, systems always operate as intended and with known consequences. In reality, impact needs to account for unintended consequences and potential system errors, including false positives (the system acts/alerts when it should not) or false negatives (the system fails to act/alert when it should). The system impact is the greater of these potential consequences.

-------------Page Break------------

Next, we will ask you to use the framework to classify five example systems. Note that one of the systems will not be an AI system. For that system, select "not an AI system." Before continuing, please confirm you reviewed the framework by selecting the correct answer below.

This framework was developed to help human users classify AI systems along two dimensions. What are the two dimensions?

1. Decision & Impact
2. Autonomy & Impact
3. Autonomy & Influence
4. Decision & Influence

-------------Page Break------------

Please read the following system description and then select the best classification. You are encouraged to use the rubric to make your classification.

Autonomy level:

Significant autonomy Some autonomy Minimal autonomy

Impact level:

High impact Medium impact Low impact

Not an AI system

[corresponding framework rubric included here]

Note: Hover your mouse over terms to see the corresponding definitions. [definitions provided to respondents by hovering over the response items].

C

Artificial intelligence (AI) is currently used in a wide range of systems. To help identify and classify systems that use AI, researchers developed the following framework. The goal of the framework is to help humans correctly classify and thus effectively manage AI systems. We will ask you to review the framework and then use it to classify five example systems.

Definitions

The framework defines an **AI system** as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The framework classifies systems along two dimensions, autonomy and impact.

Autonomy is defined as the degree to which a system processes input, generates decisions (including predictions or recommendations), and executes actions that influence physical or virtual environments without human involvement, outside of set-up and routine maintenance. The framework includes three levels of autonomy:

- Action Autonomy – During normal operation, the system processes input, generates a decision, and executes an action without human involvement.
- Decision Autonomy – During normal operation, the system processes input and generates decision output (e.g. a prediction or recommendation) but requires a human to take output-directed action.
- Perception Autonomy – During normal operation, the system processes input and flags information that requires human evaluation, decision, and action.

Impact is defined as the potential consequences of system-directed decisions for physical and system safety, national security, civil rights and liberties, and enterprise operations (including the ability to execute core functions or implement decisions without external challenge). The framework includes three levels of impact:

- High Impact - the decision outcome could lead to death or bodily harm, grave risk to national security, violation of civil rights or liberties, or significant disruption of core enterprise goals or functions.
- Medium Impact - the decision outcome could lead to risk to systems and/or network security, diminished livelihood (e.g., financial harm), or mitigable disruption of core enterprise goals or functions.
- Low Impact - the decision outcome has little to no effect on physical safety, national or system security, civil rights or liberties, or core enterprise goals or functions. The decision is unlikely to cause any significant or permanent damage.

-------------Page Break------------

Using the Framework

This framework can help classify a system based on a small amount of information. A system can be classified as action, decision, or perception autonomy and high, medium, or low impact. Here are two examples of AI systems and their corresponding classifications:

> ERNIE is a system that monitors port vehicle traffic for radiological material and radionuclear threats. The system reviews radiological and motion sensor data, classifies the data as alarm or no alarm, and determines whether to release cargo or hold for further inspection. **This system would be classified as action autonomy and high impact.**
>
> BluVector Atomic Host is a system that scans standard cyber network traffic to detect malware files in real time. The system produces a report with file detection prediction that is provided to network administrators. **This system would be classified as decision autonomy and medium impact.**

> BasisTech is a system that detects personally identifiable information within a dataset. The system reviews a dataset and provides a report that flags possible personally identifiable information within the dataset for follow up by a human operator. **This system would be classified as perception autonomy and medium impact.**

Not all systems that use compute technologies and analyze data are AI systems. If the system does not fit the definition of an AI system used in this framework it is not considered an AI system. For example:

> Microsoft Excel is a tool for data storage and analysis. The software allows users to store, sort, and run basic analysis on inputted data. **This system is not an AI system.**

For greater usability, all possible AI system classifications are consolidated into a rubric, shown below. Note that the shading of rubric cells corresponds to suggested level of oversight (e.g. darker shaded cells require a higher degree of oversight).

	Impact		
Autonomy	High	Medium	Low
Action	Executes action that could lead to death or serious risk to national security, civil rights, or enterprise functions.	Executes action that could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	Executes action that presents little to no risk for security, rights, or enterprise functions.
Decision	Makes a decision that could lead to death or serious risk to national security, civil rights, or enterprise functions.	Makes a decision that could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	Makes a decision that presents little to no risk for security, rights, or enterprise functions.
Perception	Identifies information to provide decision-support that could lead to death or serious risk to national security, civil rights, or enterprise functions.	Identifies information to provide decision-support that could lead to mitigable risk to network security, personal livelihood, or enterprise functions.	Identifies information to provide decision-support that presents little to no risk for security, rights, or enterprise functions.

For the impact dimension, it is important to keep in mind system intent and performance. Ideally, systems always operate as intended and with known consequences. In reality, impact needs to account for unintended consequences and potential system errors, including false positives (the system acts/alerts when it should not) or false negatives (the system fails to act/alert when it should). The system impact is the greater of these potential consequences.

-------------Page Break------------

Next, we will ask you to use the rubric to classify five example systems. Note that at least one system example will not fit the definition of an AI system. For that system, select the option "not an AI system."

Before continuing, please select the correct option below.

The framework just described classifies an AI system based on its:

1. Decision & Impact
2. Autonomy & Impact
3. Autonomy & Influence
4. Decision & Influence

-------------Page Break------------

Please read the following system description and then select the best classification. You are encouraged to use the rubric to make your classification.

[*System Description*]

Autonomy level:

Action autonomy Decision autonomy Perception autonomy

Impact level:

High impact Medium impact Low impact

Not an AI system

[corresponding framework rubric included here]

[insert revised rubric]

Note: Hover your mouse over terms to see the corresponding definitions. [definitions provided to respondents by hovering over the response items].

D

Artificial intelligence (AI) is currently used in a wide range of systems. To help identify and classify systems that use AI, researchers developed the following framework. The goal of the framework is to help humans correctly classify and thus effectively manage AI systems. We will ask you to review the framework and then use it to classify three example systems.

Definitions

The framework defines an **AI system** as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The framework classifies systems along four dimensions: context, input, model, and output.

Context refers to the socio-economic environment in which the AI system is deployed. Core characteristics of this dimension include the sector in which the system is deployed (e.g., healthcare, finance, defense), deployment impact and scale, and whether the system performs a critical function.

Input refers to the input or data used by the AI system to build a representation of the environment. Core characteristics of this dimension include data collection and characteristics (e.g., type, structure).

Model refers to the technical components that make up the AI system and represent “real world” processes. Core characteristics of this dimension include model type and acquisition of capabilities.

Output refers to the tasks the system performs and the action it takes to influence the environment. Core characteristics of this dimension include system task(s) and action autonomy.
-------------Page Break------------

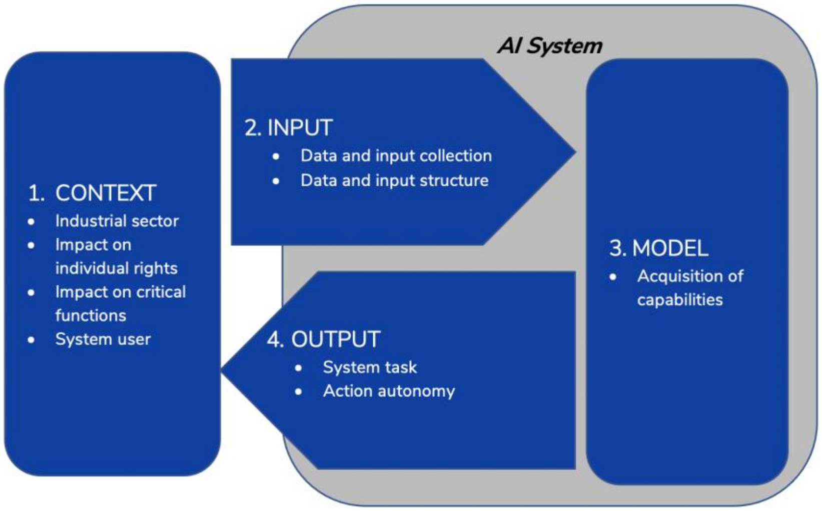
Using the Framework

This framework can help classify a system based given a small amount of information. Here are some example of AI systems:

> **ERNIE** is a system that monitors port vehicle traffic for radiological material and radionuclear threats. The system reviews radiological and motion sensor data, classifies the data as alarm or no alarm, and determines whether to release cargo or hold for further inspection.

> **BluVector Atomic Host** is a system that scans standard cyber network traffic to detect malware files in real time. The system produces a report with file detection prediction that is provided to network administrators.
>
> **BasisTech** is a system that detects personally identifiable information within a dataset. The system reviews a dataset and provides a report that flags possible personally identifiable information within the dataset for follow up by a human operator.

For greater usability, the framework dimensions for classifying a system are outlined in the figure below.



Not all systems that use compute technologies and analyze data are AI systems. If the system does not fit the definition of an AI system used in this framework it is not considered an AI system. For example:

> Microsoft Excel is a tool for data storage and analysis. The software allows users to store, sort, and run basic analysis on inputted data. Analyzed data output is used for administrative support and allocation of enterprise resources. **This system is not an AI system.**

-------------Page Break------------

Next, we will ask you to use the rubric to classify three example systems. Before continuing, please select the correct option below.

This framework was developed to help human users classify AI systems along four dimensions. What are the four dimensions?

1. Context, input, model, and output
2. Context, impact, processor, and action
3. Environment, training, evaluation, and implementation

-------------Page Break------------

Please read the following system description and then select the best classification. Note: Hover your mouse over questions to see corresponding definitions.

[*System Description*]

In what industrial sector is the system deployed?

Agriculture, forestry and fishing
Mining and quarrying
Manufacturing
Electricity, gas, steam and air conditioning
Water supply, waste management and remediation activities
Construction
Wholesale and retail trade
Transportation and storage
Accommodation and food service activities
Information and communication
Financial and insurance activities
Real estate activities
Professional, scientific and technical activities
Administrative and support service activities
Public administration and defense
Education
Human health and social work activities
Arts, entertainment and recreation
Other service activities
Services-producing activities of households for own use
Activities of extraterritorial organizations and bodies

What are the benefits and risks to individuals? [*hover text: Degree to which the system generates outcomes with the potential to impact individual safety, rights, or well-being*]:
Impact fundamental rights or values (e.g., physical safety, privacy, equality)
Impact individual well-being (e.g., job quality, education, social connections)

No impact

Does the system perform a critical activity? [*hover text: Critical activities are those for which the disruption of would mean serious consequences for 1) the health, safety, and security of citizens; 2) the effective functioning of essential services; or 3) broad economic and social prosperity. Examples include conducting elections, law enforcement, and medical care]*:

Yes
No

Who is the system user? [*hover text: Whether the human who uses the system has any system operations training*]

Amateur
Trained practitioner who is not an AI expert
AI expert or system developer

How are system inputs (e.g., data) collected? [*hover text: Whether system input is perceived from the environment by humans or by machines acting as sensors*]

By humans
By automated tools
By humans and automated tools

What is the structure of data inputs? [*hover text: The structure of data that is collected and processed by the system">What is the structure of data inputs*]

Unstructured (e.g., text, audio, video, raw data)
Semistructured (e.g., unstructured data with structured metadata)
Structured (e.g., data in predefined format such as table or dataset)
Complex structured (e.g., ontology, knowledge graph, function or data from algorithm)

How does the system acquire, or learn, its capabilities? [*hover text: How system learns to turn abstract representations of relationships into outputs capable of influencing the environment*]

Acquisition from knowledge (e.g., learns from human input or rules)
Acquisition from data (e.g., learns from provided data)
Acquisition from data and system experience

What does the system do? Select all that apply. [*hover text: The tasks or functions that the system performs*]

Recognition (categorize data into specific classifications)
Event detection (connect data to detect patterns, outliers, or anomalies)
Forecasting (use past behavior to predict future outcomes)
Personalization (develop a profile of an individual and adapt output to that individual)
Interaction support (create content for conversational interactions between machines and human)
Goal-driven optimisation (find the optimal solution to a problem)

Reasoning with knowledge structures (infer new outcomes through modeling and simulation)

What is the system's level of autonomy? [*hover text: The degree to which the system can act without human involvement*]

High action autonomy (evaluates input and executes decisions and actions without human involvement)

Medium action autonomy (evaluates input and provides predictions or recommendations for human action)

Low action autonomy (evaluates input and identifies information for human decision-making)

----- All respondents asked following questions after completing classifications -----

Thank you for classifying the example systems. Before concluding the survey, we have a few optional questions.

First, please share any thoughts you have on the framework or suggestions for improvement.

[*text box*]

What is your age, in years?

a. 18-24
b. 25-34
c. 35-44
d. 45-54
e. 55-64
f. 65-74
g. 75 or older

What is your current employment status?

a. Unemployed
b. Student
c. Employed part time
d. Employed full time
e. Retired

If Employed, please select the option that best describes your current employment industry

a. Administrative services
b. Banking & financial services
c. Education
d. Food & beverage
e. Government & non-profit
f. Healthcare

g. Manufacturing
h. Media & entertainment
i. Retail, Wholesale & Distribution
j. Software & IT services
k. Other (please specify):