November 2021

Classifying AI Systems

CSET Data Brief

AUTHOR
Catherine Aiken

Executive Summary

Artificial intelligence (AI) governance is a pressing policy issue. AI systems help us complete a wide range of tasks, from driving to the store or vacuuming, to diagnosing illnesses and providing disaster relief. AI systems are rapidly being adopted to assist, or independently complete, many more tasks. The result is a deserved focus on the safe development and deployment of AI systems. Governments are putting forth AI ethics principles, compiling AI inventories, and mandating AI risk assessments. But efforts to ensure AI systems are safe and effective require a standardized approach to classifying the varied types of AI systems in use.

Classifying AI systems involves identifying a set of observable system characteristics and assigning individual systems a predefined label for each characteristic. For example, for every AI system, we can observe its autonomy and assign it a predefined autonomy level that informs our understanding of what kind of system it is and how it works. Combining the identified system characteristics, a framework defines an AI system along those characteristics, referred to as framework dimensions. Using such a framework, system developers, governing bodies, and users can classify systems in a uniform way and use those classifications to inform consequential decisions about AI technologies, while effectively monitoring risk and bias and managing system inventories.

To that end, CSET partnered with the Organization for Economic Cooperation and Development (OECD) AI Policy Observatory and U.S. Department of Homeland Security (DHS) Office of Strategy, Policy, and Plans to develop several frameworks for classifying AI systems.[*] Each framework was built following the same process:

---

[*] CSET did not receive any funding for this research from the U.S. Department of Homeland Security or any other government entity. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DHS, and do not constitute a DHS endorsement of the rubric tested or evaluated.

1. Identify policy-relevant characteristics of AI systems (e.g., autonomy, impact, data collection method) to be the framework dimensions;

2. Define a set of levels for each dimension (e.g., low, medium, high) into which an AI system can be assigned.

To test the usability of four resulting frameworks, CSET fielded two rounds of a survey experiment where more than 360 respondents completed over 1,800 unique AI system classifications using the frameworks. We tested the ability of individuals to assign consistent and accurate classifications across frameworks to provide insight into how the public and policymakers could use the framework and what frameworks may be most effective.

Based on the survey experiment, we find:

- **Certain frameworks produced more consistent and accurate classifications.** Higher-performing frameworks more than doubled the percentage of consistent and accurate classifications, compared to the lowest-performing framework.

- **Including a summary rubric of framework dimensions improves classification.** We found a significant decrease in consistent and accurate classifications when users were not provided with a rubric when making a system classification, compared to instances when a rubric was provided.

- **Users were better at classifying an AI system's impact level than autonomy level.** Users consistently assigned the accurate system impact but struggled to consistently assign the accurate level of system autonomy, across all frameworks. Users were also better at classifying system deployment context than technical system characteristics.

- **Users were better at classifying an AI system's autonomy level when the framework provided more descriptive levels.** We found that consistent and accurate classifications were higher when system autonomy levels

were labeled as “action,” “decision,” and “perception” as opposed to “high,” “medium,” and “low.”

- **Classification depends on sufficient accessible information about the system.** We found classifications were more varied when system descriptions did not include a specific use case, suggesting that the provided descriptions shaped how well users could classify the systems. More broadly, this process showed that classifying technical characteristics requires more information than is typically available about an AI system.

Table of Contents

Executive Summary 2
Introduction 6
- Classifying AI Systems: The Example of AlphaGo Zero 7
- Existing Approaches to Classifying AI Systems 9

Classification Framework Development 10
Methodology 13
Evaluating Framework Performance 15
- Frameworks A, B, and C 16
- Framework D 20
- Comparing Frameworks 25

Discussion 29
Next Steps 31
Author 33
Acknowledgements 33
Appendices 34
- Appendix 1. Survey Distribution 34
- Appendix 2. Survey Questionnaire 36
- Appendix 3. AI System Descriptions 36
- Appendix 4. Additional Analysis 40

Endnotes 46

Introduction

Policymakers and the public are concerned about the rapid adoption and deployment of AI systems. In April 2021, the European Commission proposed a first-of-its-kind AI legal framework that relies on a "risk-based approach" to regulating AI systems.[1] Meanwhile, U.S. government agencies are compiling department-wide AI system inventories and drafting processes for reducing bias in AI design and deployment.[2] These examples represent a few efforts within a wider discussion around AI governance.[3]

These efforts, and the conversation around AI governance, show the need for a tool to guide how we identify and characterize AI systems and for a uniform approach to classifying diverse systems. Such a tool is important for efficient governance, consistent evaluation, and informed user interactions with AI systems. While some may argue this requires only an accepted definition of AI, existing definitions vary widely and a static "universally applicable definition" may be unattainable and even undesirable.[4] A focus on AI classification, as opposed to definition, allows us to know what we are looking at and talking about, providing a common foundation for assessing AI systems, without constraining AI to a single defining attribute.

A classification framework is a tool that identifies important, policy-relevant characteristics of AI systems and provides a structured way to distill down complex technologies and standardize information gathering.[5] Frameworks ensure organized and accessible paths for governments and policymakers to create and target regulations. Frameworks that clearly characterize AI systems can foster public understanding of AI, which is important when developing AI policies, as a lack of clarity "enables technology pessimists to warn AI will conquer humans, suppress individual freedom, and destroy personal privacy."[6] Beyond public awareness, enhanced understanding of AI technologies through classification can ensure technologies being developed and deployed in the private and public sector are appropriately operated, managed, and regulated. A lack of clarity here could lead to accidents and leave underdeveloped or ad hoc ethical

principles and legal frameworks to fill the void.[7] Overall, a framework for classifying AI systems can help policymakers "focus on the measurement and assessment of AI systems, both in terms of technical attributes and societal impact" on the path to safe and effective AI development and deployment.[8]

Understanding classification as a critical first step, this brief works toward a tested framework for classifying AI systems. First, it provides an example to illustrate the benefits of classifying AI systems and reviews some existing approaches to AI classification. Next, it outlines the process by which we developed our frameworks in collaboration with OECD and DHS and our methodology for testing them through a survey experiment. Then, it compares user classification consistency and accuracy across the frameworks. It concludes with lessons learned and paths for continued research.

Classifying AI Systems: The Example of AlphaGo Zero

Does the AI system act autonomously or does it provide a recommendation to a human who then acts? Who is impacted by an AI system? What sector is the AI system deployed in? How does the AI system collect data? These are questions that are relevant for AI governance and policymaking. They are questions that point to observable system attributes that we can characterize along predefined levels—our classification dimensions. Resulting classifications help us answer questions about an AI system, compare it to other AI systems, and assess the landscape of AI systems. Consider the example of the AI system AlphaGo Zero.

AlphaGo Zero is a system that plays the board game Go in a virtual environment, often beating professional human players. AlphaGo Zero uses both human-based inputs, including the rules of Go, and machine-based inputs, primarily data learned through repeated play against itself, to play the game. It abstracts data into a model of moves through reinforcement learning and then uses the model to make its next move based on the current state of play. The system is deployed in a narrow setting for research and recreational purposes and has no immediate impact beyond the outcome of the game.[9]

Using an AI system classification framework described below (Framework A, see also Appendix 2) and only this general information about AlphaGo Zero, we can classify it as a high autonomy and low impact AI system. We can now clearly communicate two policy-relevant characteristics of AlphaGo Zero—it can take an action (make a move in the game) autonomously, but that action has minimal impact on individuals, aside from the competing player, or on society. We can also consider how AlphaGo Zero relates to other systems classified using the same framework. For example, we can say it is similar, along these dimensions, to a virtual website navigation assistant, also classified as high autonomy and low impact. Meanwhile, it is very different from a missile defense AI system, classified as high autonomy and high impact.

In addition to clear, consistent, and comparable system information, we can now use AlphaGo Zero's classification to characterize its risk profile and articulate how its risk is different from systems with other classifications. Specifically, that it has a lower risk profile than systems with higher impact, especially high autonomy and high impact, but a slightly higher risk profile than a system with low autonomy and low impact.

Using the classification and risk profile, one can identify management and regulatory needs, or the absence of such needs. Specifically, one could now flag AlphaGo Zero as a system that does not require a high degree of governance and focus efforts on regulating systems with different risk profiles. More broadly, this shows how classification enables the development of policies around appropriate subsets of systems, rather than attempting to govern all AI systems through a single approach.

In sum, a classification framework allows us to use publicly available information about AlphaGo Zero to classify it along two policy-relevant dimensions, compare it to other AI systems along the same dimensions, assess its risk profile, and determine the required level of regulation. Applying the same framework to a wider sample of systems allows for analysis of the distribution of AI systems along these dimensions and the ability to subset AI systems for more targeted management and regulation.

Existing Approaches to Classifying AI Systems

There are a variety of existing frameworks for classifying AI systems. Some frameworks classify AI systems by their technical characteristics, based on the openness of their algorithms and data or according to their AI approach (e.g., supervised machine learning) and problem domain (e.g., reasoning).[10] Technical frameworks may work well for academics, VC firms, and AI experts, but they are less useful for non-technical and policy-making audiences. Additionally, the rapid evolution of AI technologies makes complex frameworks more likely to lose their relevance, particularly in a policy setting. Rather, what is needed is a user-friendly framework to provide a balanced assessment of an AI system's policy-relevant characteristics. Retaining broad usability is a continued challenge and a motivation for the development of the classification frameworks examined here.

Other existing frameworks focus on the degree of system automation to characterize an AI system—for example, using an "automation spectrum" or identifying levels of function-specific automation.[11] Often the focus on automation is at the expense of other relevant factors, like application, risk, and impact. Public sector actors in particular highlight the need for a focus on assessing risk and impact. The European Commission's proposed "Regulation on a European Approach for Artificial Intelligence" claims to provide a "proportionate regulatory system centered on a well-defined risk-based regulatory approach."[12] The Canadian government has already implemented its "Algorithmic Impact Assessment Tool," which "determines the impact level of an automated decision-system" using a questionnaire and provides guidance based on the calculated impact level.[13] Germany's Data Ethics Commission outlined a "risk criticality pyramid" for "algorithmic systems" to link system risk levels to regulatory measures.[14] Meanwhile, U.S. government guidance on regulating AI emphasizes that "the kind of AI adopted and the way it works in decision-making may present new demands on existing risk frameworks."[15] Across these efforts, there is emphasis on the need to understand how AI systems interact with humans and how the actions humans take using information provided from an AI system

impact human rights, wellness, and freedoms. Building on existing frameworks while prioritizing usability, an AI classification framework should address system interactions with humans and system impacts, and be informed by technical system characteristics.

Classification Framework Development

The classification frameworks evaluated here were motivated by expressed needs from policymakers and developed in the course of discussions with two groups in the policy community. Specifically:

1. The U.S. Department of Homeland Security's Office of Strategy, Policy, and Plans. DHS sought a viable AI system classification framework to support the department's AI Strategy and Implementation Plan.[16] Frameworks A, B, and C are versions of the frameworks discussed within the context of these plans.

2. The OECD Network of Experts on AI (ONE.AI) working group on the Classification of AI. This working group was tasked with developing a framework to classify AI systems. Framework D is a version of the framework developed by this group, which is expected to launch in early 2022.[17]

Expressed policy concerns centered around the need to efficiently identify, monitor, and characterize AI systems in development or use, and determine which AI systems are higher risk and warrant greater policy attention. CSET helped address these questions as a ONE.AI contributor over more than a year of collaborative discussion and formulation and pursued several avenues for collecting feedback on the frameworks and testing their usability.

It should be noted that CSET's discussions with these groups took place independently, but did inform one another. DHS officials were involved in the OECD working group and the proposed OECD framework draws on dimensions developed in our discussions with DHS. Additionally, as noted below, both use the OECD definition of an AI system. We highlight this to show that the need for AI system classification is not specific to a single agency, or even country, and the proposed solution of a classification framework is based on input from a wide range of stakeholders. It is also worth noting that these discussions speak to ongoing AI governance efforts by the U.S. National Institute of Standards and Technology, the European Union, and other governing bodies.

The need for AI system classification is not specific to a single agency, or even country, and the proposed solution of a classification framework is based on input from a wide range of stakeholders.

OECD Definition of AI System

- A machine-based system that is capable of influencing the environment by making recommendations, predictions or decisions for a given set of objectives, which uses machine and human-based inputs to perceive the environment, abstract perceptions into models, and formulate options for outcomes.[18]

Using the OECD definition of an AI system as a baseline, each framework identifies policy-relevant characteristics of an AI system and includes them as core dimensions, as displayed in Table 1. Frameworks A, B, and C each have two core dimensions: autonomy and impact. Framework D has four core dimensions—context, input, AI model, and output—each with one or more sub-categories, for a total of nine dimensions.[19]

Table 1. AI System Classification Framework Dimensions

Framework	Core Dimensions	Dimension Levels
A	Autonomy and Impact	Autonomy: high, medium, low, none Impact: high, medium, low
B	Autonomy and Impact	Autonomy: significant, some, minimal Impact: high, medium, low
C	Autonomy and Impact	Autonomy: action, decision, perception Impact: high, medium, low
D	Context, Input, Model, Output	Context: • Sector: International Standard Industrial Classification • Impact: fundamental values, well-being, no impact • Critical activity: yes, no • System user: AI expert or system developer, trained practitioner who is not an AI expert, amateur Input: • Data collection: by humans, by automated sensing devices, by humans and automated tools • Data structure: unstructured, semistructured, structured, complex structured Model: • Acquisition of capabilities: from knowledge, from data, from data and system experience Output: • Task: recognition, event detection, forecasting, personalization, interaction support, goal-driven optimization, reasoning with knowledge structures • Autonomy: high, medium, low, none

Source: CSET Classifying AI Systems Survey.

The variation in number and type of core dimensions largely stemmed from two factors, the first being differing perspectives on the appropriate balance between the number of dimensions to include and framework usability. In theory, there are many characteristics of an AI system that are relevant for policymaking, but including every characteristic as a framework dimension would limit its usability by 1) increasing complexity and length, 2) requiring technical, often proprietary, information about systems, and 3) reducing comparability across systems. Thus, we test frameworks with different numbers of dimensions. Second, we wanted some variation within frameworks with the same number of dimensions to see if factors beyond the number of dimensions influence overall classification performance.

Our discussions with policymakers and working group experts highlighted several policy-relevant characteristics of systems—notably, the system's impact. Ideally, this dimension would capture the impact of the system's operation and output on individuals and communities while also accounting for the reversibility (or not) of system-guided decisions or actions. Given this emphasis on impact, we tested frameworks that combine these considerations into a single impact dimension (e.g., high, medium, low impact) and a framework that breaks down these considerations into multiple dimensions (e.g., impact on individual rights, impact on critical activities). Autonomy also came up as a key system dimension for policymaking. Discussions highlighted the need to understand the degree of human involvement in the system operation, specifically the degree to which the system can make a decision or carry out an action independent of human involvement, and how humans use system output. Like impact, we approached this dimension in two different ways. We capture these considerations in a single autonomy dimension in Frameworks A, B, and C. Then in Framework D, we include multiple dimensions that speak to system-human interactions (e.g., autonomy, task, end user).

Methodology

We conducted two rounds of a survey experiment to test the usability of the resulting frameworks and compare user classifications.[20] First, we identified several AI systems deployed in

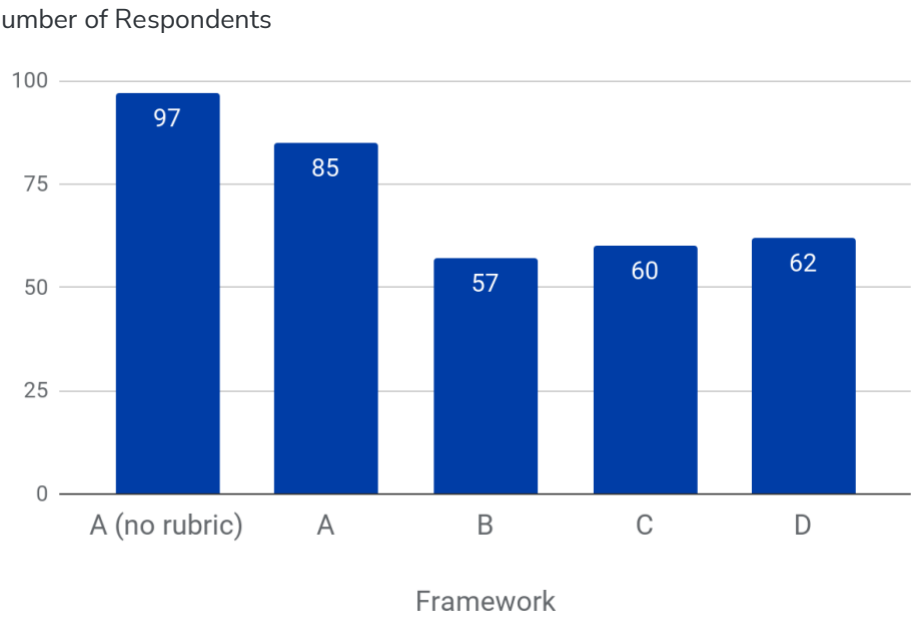
the real world in one of three ways: 1) collected in an internal DHS call and shared with CSET for purposes of this study; 2) identified by CSET researchers as a deployed AI system; or 3) identified by the ONE.AI working group as a deployed AI system. We crafted a short description for each system using general, publicly available information about the system.[21] While some survey respondents may have been aware of a system prior to taking the survey, it is not likely that they knew much information beyond what we provided in the descriptions and therefore prior knowledge was unlikely to cause them to be better at classifying those systems. Survey respondents were randomly assigned to use one of the frameworks to classify three to five of the following systems:[22]

- Aegis missile defense (Aegis)
- AlphaGo Zero (AGZ)
- Caster drug interaction predictor (Caster)
- C-CORE iceberg detection (C-CORE)
- Credit scoring system (SCORE)
- Facial image quality evaluator (FIQ)
- Search and rescue detection (SRD)

Survey respondents were recruited through Amazon's Mechanical Turk. Respondents were required to be located in the United States, have a high school degree or more education, and be assigned Amazon Mechanical Turk Masters Qualification.[23] We also included questions in the survey to capture respondent familiarity with AI technologies and political interest.

In total, 361 respondents provided 1,831 system classifications using the four frameworks. Figure 1 displays the number of respondents assigned to each framework. Note that we tested two variations of Framework A. One included a rubric which summarized the dimensions while the other did not include the rubric (see Table 3 for an example rubric).[24]

Figure 1. Respondent Assignment to Classification Frameworks



Note: Two variations of Framework A were split between 182 respondents. Frameworks B, C, and D were split between 179 respondents.
Source: CSET Classifying AI Systems Survey.

Overall, we think it is informative to assess framework usability and classification performance among public users because one goal of this framework is to establish a transparent process for classifying AI systems that is effective for policymakers but understandable to the public. We also sent the survey to a small sample of U.S. government officials, but with only 11 responses, we do not report findings here (see some discussion in Appendix 4.3). A version of the survey was also made available online as a part of the public consultation stage for the OECD Framework for Classifying AI Systems, but responses are still being analyzed. Finally, a version of the survey remains open and is accessible through the Classifying AI Systems interactive tool.[25]

Evaluating Framework Performance

To assess how well users classified AI systems using the different frameworks, we compare classification consistency and accuracy.[26] By classification consistency we mean agreement, in terms of users assigning a system dimension the same level, among 65 percent or more users. By classification accuracy, we mean 65

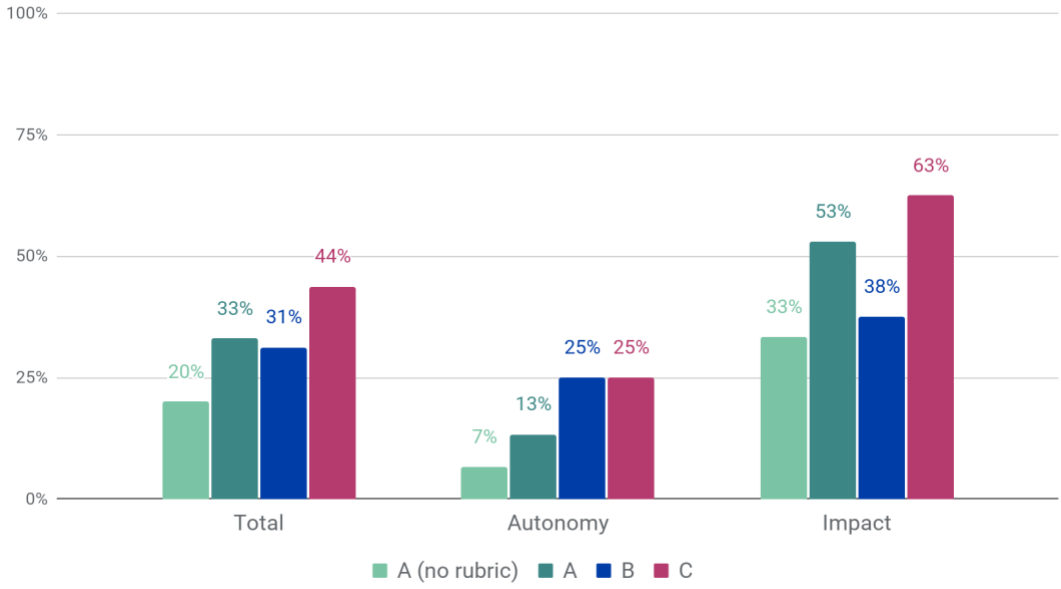
percent or more users assigning the classification that matches an expert classification.[27] We primarily assess classification consistency and accuracy at the system dimension level, meaning we evaluate classifications for each framework dimension for each example system (e.g., AlphaGo Zero's level of autonomy). Overall, we found classification consistency and accuracy varied by classification framework used and the system and dimension being classified, which indicate potential framework features associated with improved usability and understandability.

Frameworks A, B, and C

Given the different number of dimensions in Frameworks A, B, and C compared to Framework D, we first review the former, two-dimension frameworks. As shown in Figure 2, Framework C led to a higher rate of consistent and accurate classifications, with 44 percent of system dimensions assigned a consistent, accurate classification.

Figure 2. Classification Performance for Frameworks A, B, and C

Percentage of Accurate and Consistent Dimension Classifications



Note: Total system dimensions varied by framework due to the number of example systems included in the two rounds. Framework A included 15 systems (15 autonomy and 15 impact = 30 total dimensions) while Frameworks B and C included 8 systems (8 autonomy and 8 impact = 16 total dimensions).
Source: CSET Classifying AI Systems Survey.

Looking at the autonomy and impact classifications separately, we see accuracy and consistency was higher across the board for system impact. For example, using Framework C, five out of eight example systems were assigned a consistent impact level but only two out of eight were assigned a consistent autonomy level. Table 2 shows the dimensions with accurate, consistent classification for Frameworks A, B, and C across the seven example AI systems.

Table 2. Accurate, Consistent Classifications by System Dimension

	Framework							
	A (no rubric)		A		B		C	
System	Autonomy	Impact	Autonomy	Impact	Autonomy	Impact	Autonomy	Impact
Aegis	✓	✓	✓	✓	✓	✓	✓	✓
C-CORE			✓					
SCORE	–	–	–	–				
FIQ								✓
SRD				✓		✓		✓
Caster							✓	
AGZ	–	–	–	–	✓	✓		✓
ARTCC		✓		✓				✓

Note: Eight additional systems were included when testing Framework A, which are not included in this table.[28] For Framework A (no rubric), three additional system impact classifications were accurate and consistent. For Framework A, five additional system impact classifications were accurate and consistent.
Source: CSET Classifying AI Systems Survey.

The primary difference between Frameworks A, B, and C was the labels for the autonomy dimension levels. Framework A included four autonomy levels: high, medium, low, and no autonomy. Framework B included three autonomy levels: significant, some, and minimal. Framework C used three autonomy levels: action, decision, and perception.[29] Framework C, which had the overall highest rate of consistent and accurate classifications, used descriptive labels for the levels of autonomy, intended to highlight what the AI system can complete without human intervention. Another difference was that Framework A included a longer definition of autonomy and longer autonomy level descriptions.

Frameworks B and C shortened the definitions for clarity. For the full text for each framework and their accompanying rubrics, see Appendix 2 or the Classifying AI Systems Interactive.

Framework C Autonomy Definition and Levels

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.

These differences led to slight variation in consistent classification of a system's autonomy level, but primarily between Framework A, which had lower consistency, compared to B and C. Notably, the increase in consistent autonomy classification going from Framework A to B and C was starker when we decreased the threshold for consistent classifications to more than 50 percent of respondents assigning the same level of autonomy, as seen in Appendix 4.4. This suggests that the shortened definitions and descriptions and removal of the "no autonomy" option, produced more consistent classifications, even though users generally struggled with classifying a system's autonomy level.

We also see a notable difference in classification performance when a rubric that summarizes the dimensions was provided for

the user.[30] This variation was only tested in the first round of the survey experiment when we evaluated Framework A. In that round, we found classification consistency and accuracy was higher when the users could consult the rubric, compared to instances when the user was not provided the rubric. We decided to include a corresponding rubric in the next round of the survey when we evaluated Frameworks B and C, as shown in Table 3.

Table 3. AI System Classification Rubric for Framework C

	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.

Source: CSET Classifying AI Systems Survey.

Framework D

Framework D included four core dimensions—context, input, model, and output—which are broken down into subcategories, resulting in nine dimensions that were classified for each system. These dimensions represent a subset of dimensions included in the current OECD classification framework and were chosen because they were the most developed definitions and labels at the time we fielded this survey.[31] These include autonomy, within the output dimension, and impact, within the context dimension like Frameworks A, B, and C but also include sector of deployment, criticality, end user, data collection, data structure, acquisition of capabilities, and tasks.

Framework D Dimensions

Context

- **Sector** - The industrial sector in which the system is deployed.
- **Impact** - Whether the system generates outcomes that present a risk or benefit to fundamental human rights, to individual well-being, or has no impact.
- **Criticality** - Whether the system performs a critical activity, where critical activities are those for which the interruption of would mean serious consequences for the health, safety, and security of citizens or the effective functioning of services essential to the economy, society, or government.
- **End user** - Whether the intended user of the system in current deployment context is an AI expert or system developer, trained practitioner who is not an AI expert, or an amateur.

Input

- **Data collection** - Whether data providing the system input is perceived from the environment by humans or by machines acting as sensors.

- **Data structure** - Whether the structure of the data providing the system input is unstructured, semistructured, structured, or complex structured.

Model

- **Acquisition of capabilities** - Whether the system's models acquire learning capabilities from expert knowledge, data, or system experience.

Output

- **Task** - The tasks that the system performs, to include recognition, event detection, forecasting, personalization, interaction support, goal-driven optimization, or reasoning with knowledge structures.
- **Autonomy** - The degree to which the system can evaluate input, make decisions, and act without human involvement.

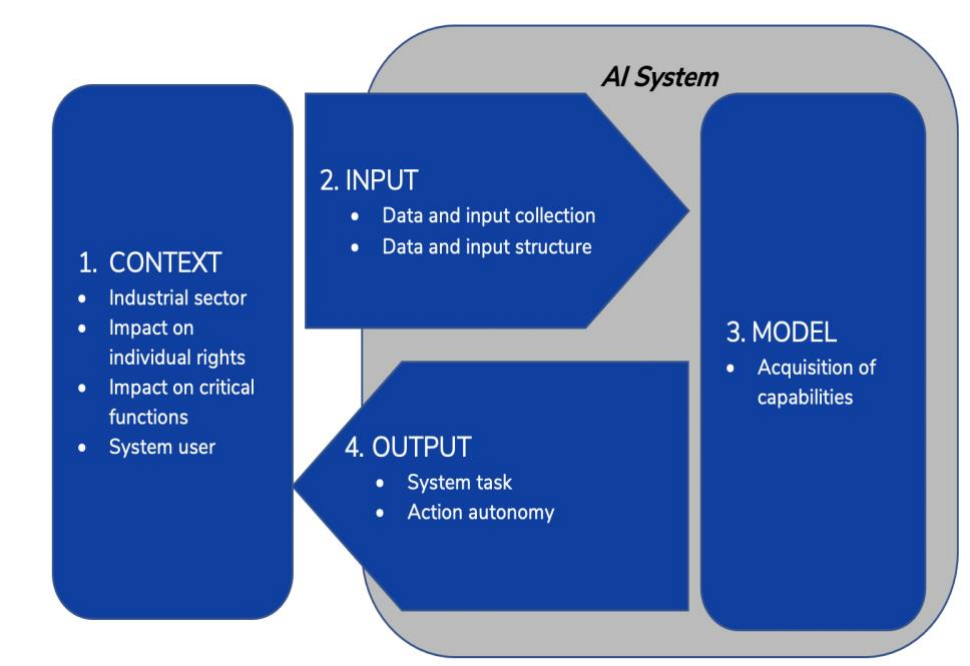


Table 4 displays the Framework D dimensions and seven example AI systems, with a check indicating those assigned an accurate, consistent classification. We see more context dimensions with consistent classifications but fewer input or model dimensions with consistent classifications. Overall, 51 percent of dimensions were assigned a consistent classification.[32]

Table 4. Accurate, Consistent Classifications by System Dimension

	Context				Input		Model	Output	
System	Sector	Impact	Critical	User	Collection	Structure	Acquisition	Tasks	Autonomy
Aegis	✓	✓	✓		✓			✓	
C-CORE			✓	✓	✓			✓	
SCORE	✓	✓		✓		✓		✓	
FIQ				✓				✓	
SRD		✓	✓	✓		✓		✓	
Caster	✓	✓*	✓	✓					✓*
AGZ	✓	✓	✓		✓		✓	✓	

Note: * indicates classifications that were consistent, but not accurate. For details of the seven example systems classified by users assigned to Framework D, see Appendix 3. Source: CSET Classifying AI Systems Survey.

In addition to variation by framework dimension, we see variation in classification performance by AI system. The two systems with the lowest classification accuracy and consistency using Framework D were FIQ and C-CORE. These systems have similar attributes—both are deployed in the professional, scientific sector and have image-based scanning recognition capabilities that pose minimal risk to individual rights. Lower classification performance for these systems could be due to unexpected ambiguity in the provided descriptions, but may represent a more general lack of understanding about these types of systems and their applications.

System Context

As shown in Table 4, users did relatively well with the context dimension, which consists of the industrial sector in which the system is deployed, the system's impact on individuals and critical activities, and the system's end user.

For the industrial sector of deployment, a large majority of users assigned the same, accurate classification to most systems. Three systems—FIQ, SRD, and C-CORE—had only a minority of users

assign the same, accurate sector classification. This was likely due to a lack of clarity around the system's deployment context in the provided description in these cases. Meanwhile, the systems that had more consistent and accurate classification are deployed in industries easily recognizable to most users (e.g., the SCORE credit-scoring system being deployed in the finance sector).

For the system's impact on individuals, most systems were assigned a consistent and accurate classification. The high-impact systems—Aegis and SRD—had overwhelmingly accurate responses. As both systems have clear impacts on national security and human safety, the awareness of risk seems appropriate. But two systems, FIQ and C-CORE, had a range of classifications and were far below the threshold for consistency.

For criticality, most systems were assigned accurate, consistent classifications. AlphaGo Zero was the only system that was accurately classified as *not* performing a critical activity, given the recreational function of the system. While not reaching the threshold for consistent classification, more users considered FIQ and SCORE as performing critical activities, though they were deemed to be non-critical by experts. It appears users were reluctant to cast any of the example AI systems as performing a non-critical activity. This suggests the public may assume that the use of AI implies a level of criticality, even when it is being deployed for non-critical activities.

For end user, most systems were assigned accurate, consistent classifications. Two systems failed to meet this threshold—Aegis and AlphaGo Zero. For Aegis, respondents were split whether the primary system user was an AI expert or a non-expert trained practitioner, and for AlphaGo Zero they were split whether the end user was an amateur (i.e., player of the game Go) or AI expert.

System Input

Users did not have much success classifying the input dimensions, which consisted of data collection method and data structure. Respondents had slightly more success classifying a system's mode of data collection. Three systems—Aegis, C-CORE, and

AlphaGo Zero—were assigned the same, accurate data collection classification by a majority of users. But more systems had inconsistent classifications. The data structure dimension also had low success, with only two systems consistently assigned the accurate classification. User classifications of data structure for three of the example systems was no better than random.

System Model

For the model dimension, which asked users to classify how the system's model acquired its learning capabilities, only one system had an accurate, consistent classification—AlphaGo Zero. Most systems had a small majority of users (50–60 percent) who assigned the accurate classification, but did not quite meet the threshold of consistency.

System Output

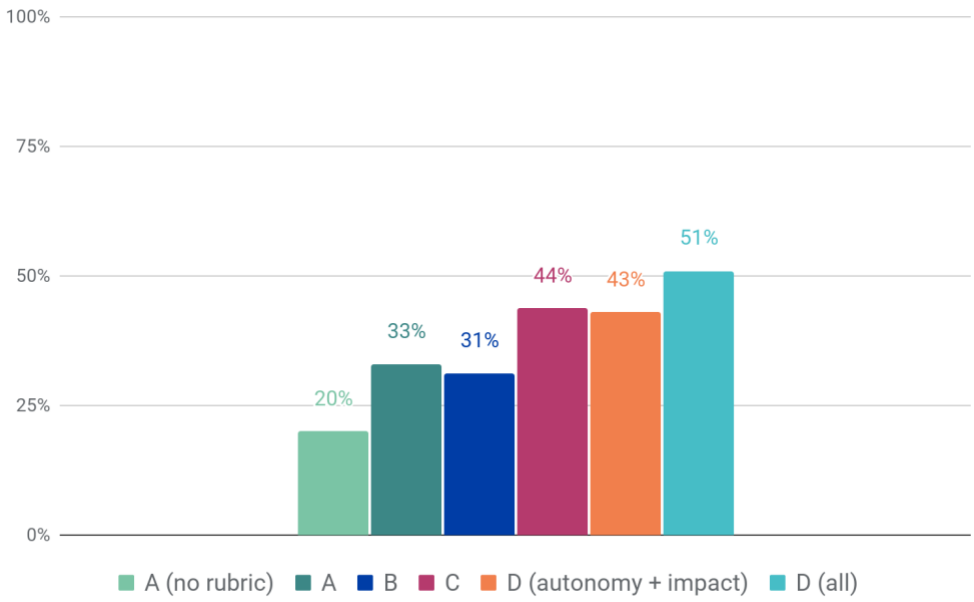
Users had mixed success classifying the output dimension, composed of system task and autonomy. In fact, system tasks had the most accurate, consistent classifications of any Framework D dimension.[33] All but one system had at least one task accurately identified by a majority of respondents.[34] Event detection, recognition, and forecasting were identified accurately for each of the systems performing these tasks, suggesting they were easier to understand. When they were performed by the system, interaction support and goal-driven optimization tasks were less often selected by users.

In terms of system autonomy, users consistently assigned a classification for only one system—Caster. Interestingly, the classification a majority of users chose (76 percent chose medium autonomy) did not match the expert classification (low autonomy). Users did not assign any system a consistent classification that matched the expert classification. This reflects the classification performance for Frameworks A, B, and C—where at best, only a quarter of systems had accurate, consistent autonomy classifications.

Comparing Frameworks

Figure 3 displays the percentage of consistent classifications for each framework. Given the difference between these frameworks, these comparisons are only suggestive. While Framework D had the highest rate of consistent classification, with 51 percent of system dimensions assigned a consistent classification, that includes seven dimensions only classified in that framework. When we restrict our comparison to only the autonomy and impact dimensions, meaning only dimensions classified in all frameworks, we see almost identical consistency between Frameworks C and D, 44 and 43 percent respectively. Notably, when we compare consistency across all frameworks for only the five AI systems classified using every framework, Framework C performs best, with 60 percent of dimensions classified consistently (see Figure C in Appendix 4.5).

Figure 3. Classification Consistency by Framework, All Dimensions

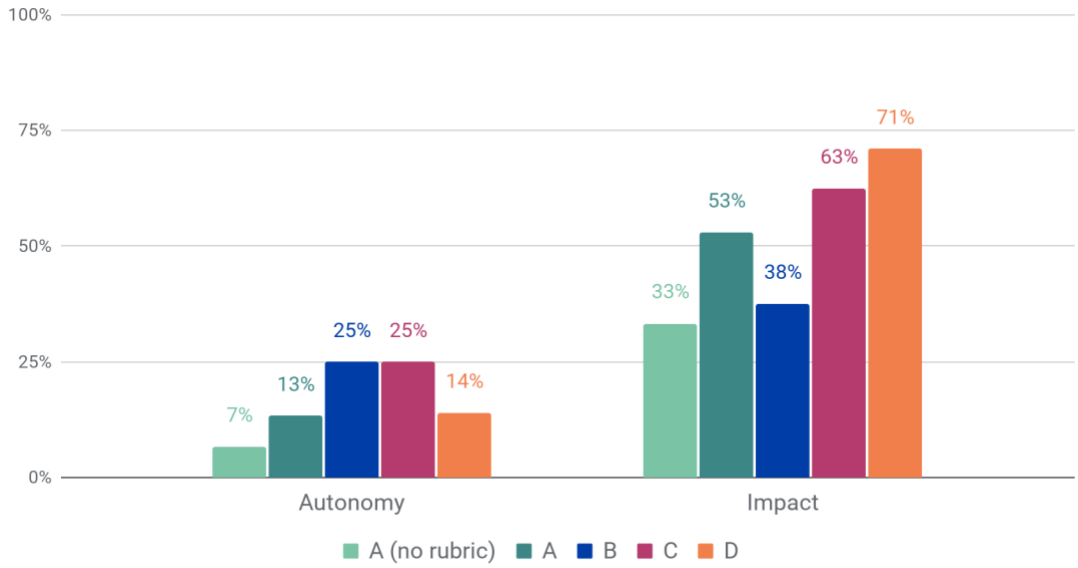


Note: Total system dimensions varied by framework due to the number of example systems included. Framework A included 15 systems (15 autonomy and 15 impact = 30 total dimensions) while Frameworks B and C included eight systems (eight autonomy and eight impact = 16 total dimensions) and Framework D included seven systems (nine dimensions * seven systems = 63 total dimensions). Source: CSET Classifying AI Systems Survey.

Across frameworks, classification consistency is higher for the impact dimension compared to the autonomy dimension, as seen in Figure 4. In fact, for Frameworks C and D the higher rate of consistent classification for all dimensions, compared to Frameworks A and B, appears to be primarily driven by more consistent impact classifications. Autonomy classifications were relatively inconsistent throughout.[35]

Figure 4. Classification Consistency by Framework, Autonomy and Impact

Percentage of Consistent Dimension Classifications



Note: Total system dimensions varied by framework due to the number of example systems included. Framework A included 15 systems (15 autonomy and 15 impact for 30 total dimensions) while Frameworks B and C included eight systems (eight autonomy and eight impact for 16 total dimensions) and Framework D included seven systems (seven autonomy and seven impact for 14 total dimensions). Source: CSET Classifying AI Systems Survey.

Figures 3 and 4 report classification consistency only, but accuracy maps on to classification consistency with two exceptions. For Framework D, 48 percent of system dimensions were accurately classified, as opposed to 51 percent consistently classified. This is due to two system dimensions (Caster impact and Caster autonomy) having a consistent user classification that did not match the expert classification. It should be noted that we consider accuracy to be the weaker measure of framework performance, as

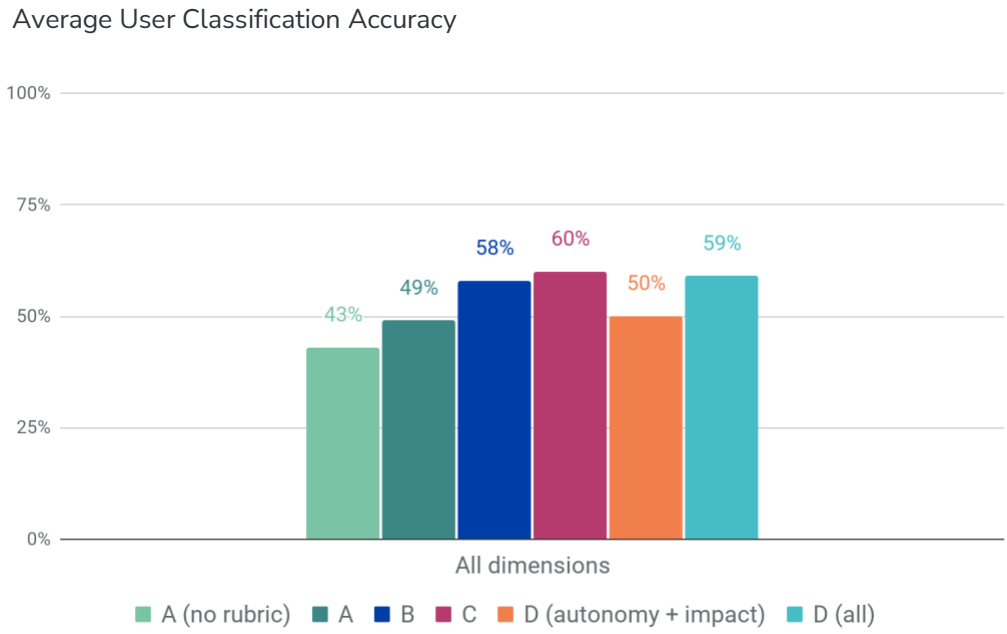
expert classifications contained some disagreement and aggregating expert classifications into a single classification required making decisions without complete knowledge of the AI system in question.

User Classification Accuracy

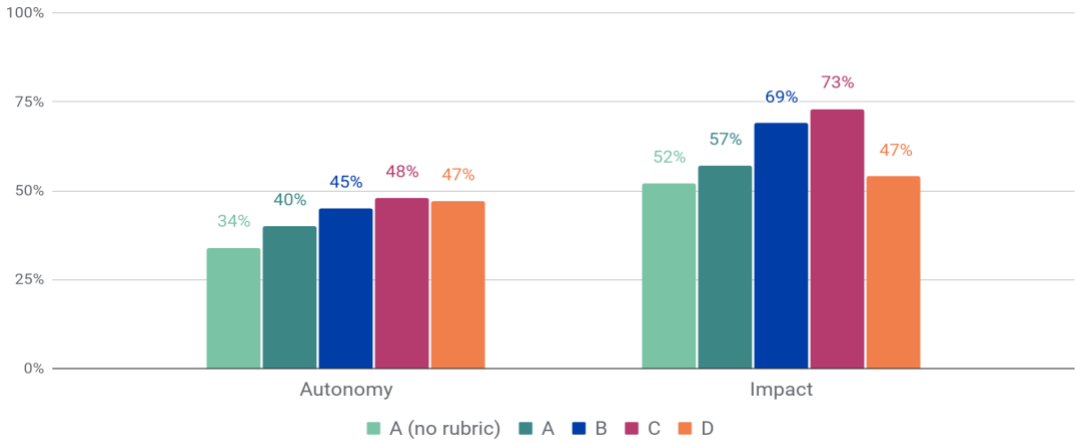
As an alternative way to compare performance across all frameworks, we evaluate accuracy at the individual user level. As a reminder, an accurate classification means a user assigned a classification that matched the expert classification. We calculated an average accuracy score for each user by dividing the number of system dimensions they accurately classified by the total number of system dimensions they classified.[36]

On average, across all frameworks, users accurately classified just over half, 51 percent, of the dimensions they classified. Users were more accurate when 1) classifying impact level and 2) provided with the framework rubric to reference when making their classifications. Figure 5 displays the average user classification accuracy for each framework.

Figure 5. Rate of Accurate User Classifications by Framework



Average User Classification Accuracy



Source: CSET Classifying AI Systems Survey.

In general, we see less variation in average user classification accuracy than we found in consistent system dimension classifications. Yet Framework C does maintain a slight lead, with users correctly classifying an average of 73 percent of the system impact dimensions and 48 percent of the system autonomy dimensions they classified, for a combined average 60 percent of system dimensions classified correctly. Framework B had similar average user accuracy rates, but given the lower rate of accurate and consistent dimension classification, it seems these accurate user classifications were more widely distributed across systems and dimensions than in Framework C. Meanwhile, using Framework A without the assistance of a rubric, a lower average of 43 percent of system dimensions were accurately classified.[37]

Discussion

While classification performance varied by framework tested, some specific framework features led to more consistent and accurate classifications. Two frameworks performed well in our evaluation when accompanied by a rubric that summarized the framework for user reference:

- Framework C, relative to Frameworks A and B: A two-dimension framework with concise, descriptive autonomy levels and clear impact level distinctions, and;
- Framework D: A four-dimension framework that requires a more comprehensive understanding of a system and its operating environment.

Additionally, we note several factors that seemed particularly relevant for human classification of AI systems and likely led to some variation in classifications.

First, the amount and type of information provided about the system is critical in the classification process. Our written descriptions of each system had, as one can imagine, an impact on the classification of that system. For example, clear information about the system deployment context was important. In one example, the description for FIQ—a system used for evaluating the quality of digital face images—implied but did not specify the scientific and technical activities sector, so users classified the system as deployed in a range of sectors, from information and communication to public administration, likely stemming from a lack of information regarding the specific use case.

Second, users struggled classifying the more technical characteristics of systems. Autonomy classifications were more inconsistent and inaccurate compared to system impact classifications. Specific to Framework D, respondents struggled to connect examples of data inputs in the system descriptions to a defined data structure or collection method. Similarly, classifications of the model dimension—how the system acquires its learning capabilities—were largely inconsistent and inaccurate. Beyond users having a lack of technical expertise, information on

these aspects of AI systems is often not readily available to the public, which may limit the ability to include them in a framework despite the utility of knowing these features of an AI system for governance and policymaking.

Third, our results suggest that public discourse around AI and current events may influence classifications. For example, some system descriptions left more room for interpretation for some dimensions, such as system impact. Systems that appeared to allow for wider interpretations were FIQ, C-CORE, and Caster. We can reasonably assume many respondents had some exposure to or familiarity with conversations around facial recognition, satellite image scanning, and the use of AI in medicine. Perhaps media coverage of these technologies and public discourse around ethical considerations shaped the decisions users made regarding system impact. For example, 64 percent of respondents classified FIQ as posing a fundamental risk to human rights. While the description mentioned that the images were evaluated for quality and were used by developers of facial recognition technologies only as a means to refine the system, respondents may have been concerned about privacy and other risks associated with these technologies.

Next Steps

A standardized framework that offers specific and consistent information about an AI system is important for viable AI governance. An effective, usable framework would enable efficient risk assessment, organization-level AI management, and public awareness of AI systems and their potential impact. But a comprehensive framework is only part of the task. We found that with minimal system information individuals with limited technical knowledge can, with moderate success, classify some characteristics of AI systems that matter for policymaking. But ensuring users have access to relevant system information, including technical system characteristics, is key to the process of classifying AI systems.

An effective, usable framework would enable efficient risk assessment, organization-level AI management, and public awareness of AI systems and their potential impact.

There are already several paths of continued research in the works. First, an expanded version of Framework D, published by OECD, underwent a period of public consultation in June 2021, which included an option to complete a similar survey to classify the same set of systems. This yielded more than 160 additional system classifications. Those classifications and the accompanying in-depth feedback provided by those users are currently being analyzed and incorporated by the ONE.AI working group, and the framework is slated to be launched in early 2022. Depending on what we learn from that process, we may continue to survey expert populations to assess the utility of a revised framework.

Second, CSET is exploring an alternate approach to identifying AI systems in use and classifying them. Using the Partnership on AI's database of real-life incidents that involve AI systems, we identified systems involved in "AI Incidents."[38] From the news articles covering these incidents, human annotators extracted the system involved and information about the system needed to classify it along the dimensions in Framework D. To date, this approach has led to more than 100 additional identified and

classified AI systems. While analysis of this approach is preliminary, it could lead to an automated process for extracting system information from news articles and other documents that would allow for standard system classification and potentially to clustering AI systems based on those classifications.[39]

Finally, CSET continues to support the implementation of these frameworks, or improved versions of them, in actual policy contexts. This includes continued participation in discussions with the ONE.AI working group, continued support for testing these frameworks, as well as complementary frameworks for assessing AI risk, bias, or trustworthiness. Given the importance of classification for AI governance, CSET is eager to continue developing an agreed-upon and usable classification framework for AI systems.

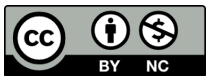
Author

Catherine Aiken is the Director of Data Science and Research at CSET.

Acknowledgements

The author is grateful to Eish Sumra for invaluable research assistance. A special thanks to Nicholas Reese for his thoughtful input and support over the course of this research. Thank you to the OECD.AI working group on the classification of AI for sharing their insights and feedback; and to the OECD Secretariat Directorate for Science, Technology, and Innovation, specifically Karine Perset, Louise Hatem, and Luis Aranda for organizing these important discussions and expanding this research.

The author also thanks Dewey Murdick and Igor Mikolic-Torreira for many lengthy discussions on these questions and their support for this work. Thanks to Matthew Daniels, Margarita Konaev, Emma Westerman, and Osonde Osoba for their thorough reviews and helpful comments and suggestions. Thanks to Lynne Weil, Tessa Baker, and Owen Daniels for several rounds of review and feedback on this brief. Melissa Deng and Alex Friedland provided editorial support.



Document Identifier: doi: 10.51593/20200025

Appendices

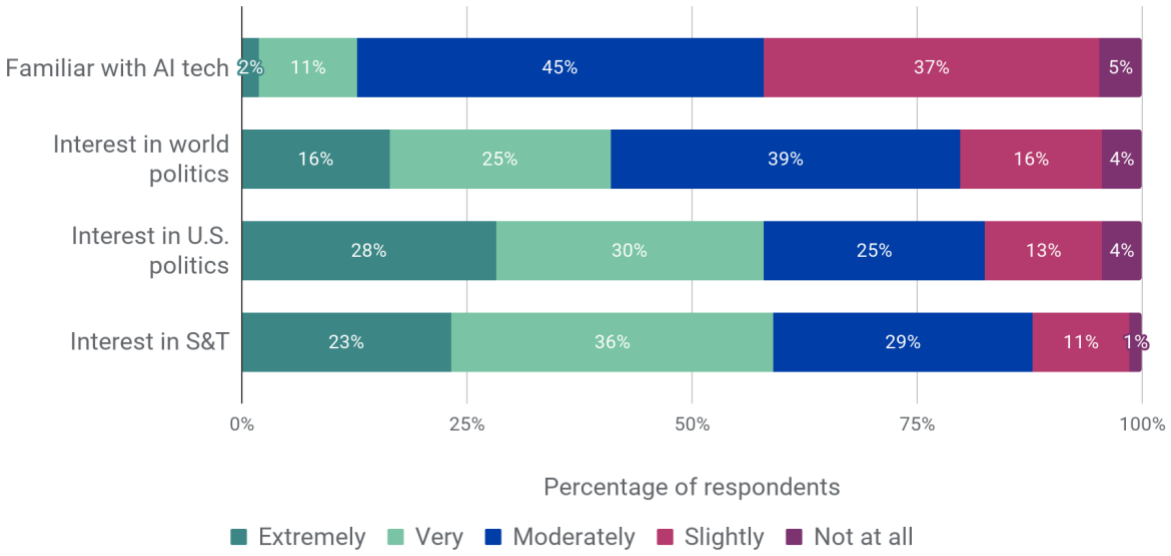
Appendix 1. Survey Distribution

In wave one (September 2020), 182 respondents completed the survey using Framework A, with or without the rubric, resulting in 910 system classifications. In wave two (December 2020), 179 respondents classified systems using Frameworks B, C, and D, resulting in 921 system classifications. Average time to complete the survey, including all example system classifications, was just under 11 minutes. We included an attention check question in the wave two survey and 94 percent of respondents selected the accurate answer. This, combined with the fact that a large majority of respondents provided detailed feedback on the task suggests respondents were paying close attention while completing the survey.

The survey asked respondents to provide optional background information such as employment status and age. 71 percent of respondents reported being employed full-time and 12 percent employed part-time. 11 percent reported being unemployed and 6 percent retired. Respondents were employed in several sectors, with 15 percent reporting working in IT and software. In terms of age, respondents skewed young, with 63 percent between 18 to 45 years old and 37 percent older than 45.

To estimate respondents' knowledge of AI and political interest, we asked respondents to indicate their familiarity with AI technologies and their interest in U.S. politics, world affairs, and science and technology (S&T) developments. As seen in Figure A, respondents were not very familiar with AI technologies—a mere 13 percent indicated high familiarity. 5 percent reported no familiarity at all.

Figure A. Respondent AI Familiarity and Political Interest



Source: CSET Classifying AI Systems Survey.

Respondents reported higher levels of political interest. A majority of respondents were extremely or very interested in U.S. politics and S&T developments, while a majority were very or moderately interested in global affairs.

Examining whether self-reported familiarity with AI or political interest related to individual classification performance, we found that there was no significant difference in the mean individual classification accuracy based on an individual's self-reported AI familiarity. In fact, the only difference was a slightly lower mean accuracy score among individuals who reported being *extremely familiar* with AI technologies, but it may be due to the small number of respondents in this group (n=7). Average individual accurate scores were nearly identical for all levels of respondent interest in U.S. politics, world affairs, and S&T.

In future work, it will be interesting to continue to compare classification performance among participants with more relevant experience and education to performance among the MTurk participants to see if accuracy and consistency do relate to experience and education, despite these null findings based on

self-reported familiarity and interest—see additional analysis below for some preliminary findings.

Appendix 2. Survey Questionnaire

The complete survey instrument, including the framework and instruction for each experimental condition, is available here and on the Classifying AI Systems interactive.

Appendix 3. AI System Descriptions

Respondents were randomly assigned to use one of the frameworks to classify three to five AI systems from the following example systems. Note that some system names were altered or created for purposes of the survey and may not reflect actual system names. Systems included in both rounds (Frameworks A, B, C, and D) were:

- **Aegis**: Aegis Ballistic Missile Defense is a system that uses radar to detect incoming intermediate-range ballistic missiles from marine vessels. If detected, the system transmits target-detection information to interceptors and fires them at incoming missiles. The system makes a decision based on the information it perceives and transfers the decision to military warships, enabling them to automatically disarm missiles.

- **Caster**: Caster is a system that reviews inputted molecular information of drugs for medical research purposes. The system is trained to recognize possible drug interactions and then models organic chemical reactions to predict drug-to-drug interactions, including potential harmful interactions. Scientists evaluate the model's output to determine the potential for adverse drug reactions.

- **C-CORE**: C-CORE is a system that scans satellite imagery over ocean areas to locate marine environmental structures (e.g., icebergs). The system processes satellite imagery to identify structures or objects. The system determines object type, position, and size of identified structures and automatically enters that information into a marine safety

database. The dataset is used by officials to evaluate marine safety, including transportation and shipping routes.

- **Facial Image Quality**: Facial Image Quality is a tool for determining the quality of a digital face image. The system reviews a digital face image and produces a face image quality score. The score is used by developers of facial recognition technologies to help determine the reliability of a face image collected through facial recognition technology.

- **Search and Rescue Detection**: Search and Rescue Detection is an aircraft-based system that scans full motion video to augment human detection of marine vessels in distress and persons in the water. The system provides a display that identifies possible vessels or persons in the water for the aircraft crew. The system uses human and sensory inputs to flag the presence of objects or persons that fit defined criteria for being in distress. The identified information is shared in real-time with human operators to determine the need for emergency response.

- **Air Route Traffic Control Center** (included as a non-AI system in A, B, and C): An ARTCC is a system set up to oversee and manage air traffic within a controlled zone. Radar technology provides human operators with flight path and aircraft information for aircraft flying in the zone's airspace. The human operators monitor the radar and flight information in order to assist aircraft pilots.

Two systems were added only in our round two distribution (Frameworks B, C, and D):

- **AlphaGo Zero**: AlphaGo Zero is a system that plays the board game Go better than professional human players. The board game's environment is virtual and player positions are constrained by the rules of the game. AlphaGo Zero uses both human-based inputs, including the rules of Go, and machine-based inputs, primarily data learned through repeated play against itself. It abstracts data into a model of

actions, or moves, through reinforcement learning. It then uses the model to make its next move based on the current state of play.

- **SCORE credit scoring**: SCORE is a system that makes credit score recommendations to help gauge a loan applicant's credit-worthiness. It does so by using human-based inputs (e.g., a set of rules) and data inputs (e.g., loan payments histories) to assess whether applicants are repaying loans on a regular basis. It automatically aggregates inputs and uses a scoring algorithm to formulate a credit score and recommendation of options for providing or denying a loan. The resulting score and recommendation are reviewed by a human who decides whether to grant or deny a loan.

Eight systems were included in our round one distribution (for a total of 15 systems tested using Framework A) but removed for subsequent distributions:

- **EMMA**: EMMA is a system that provides virtual assistance with website navigation. The system understands a wide range of questions and provides immediate answers in written or spoken format. The system can also help users search websites and direct them to the most relevant pages.

- **ERNIE**: 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. The system notifies officers in the case of hold for further inspection.

- **Checkpoint Property Screening**: The Checkpoint Property Screening System checks luggage for non-explosive threat items such as knives or guns. The system reads a 3D image of screened luggage and classifies items into the following categories: weapons, sharps, blunts, or benign. If the system

predicts the presence of a threat item it alerts an officer to search the bag.

- **Blue Prism**: Blue Prism is a system that reviews and audits National Flood Insurance Program claims applications. The system reviews claims and provides a recommendation of approve, deny, or flag for review. Given preset rules and constraints, the system helps determine which claims fall outside preset qualifications. The system includes a front-facing platform for human claims reviewers to assess claimant information and review system-generated recommendations.

- **Cyber Network Vulnerability Assessment**: AWARE is a system that evaluates cyber inventory and assets to identify vulnerabilities. It produces a report with visualizations and recommendations of the vulnerabilities that are most likely to be exploited in a cyber incident. Reports are used by administrators to help decide which vulnerabilities to prioritize.

- **Griffeye Brain**: Griffeye Brain is a system that scans media seized by law enforcement for the presence of child exploitation images or videos. The system classifies images and videos to alert officials of potential child exploitation material and bookmarks files that require investigation.

- **Program Health Assessment**: Program Health Assessment (PHA) is a system that reviews comma-separated values files (CSVs) and application programming interfaces (APIs) to aggregate raw data and flag possible metrics for analysts. The collected data is used by analysts to inform evaluations of investment risk, program success, and product delivery.

- **Commerce Control List** (included as a non-AI system): The Commerce Control List (CCL) lists goods and technology items regulated by the U.S. government. Regulated items are considered dual-use, meaning they are designed for commercial purposes but have possible military

applications. Officials use the CCL to determine if technical goods and data can be exported outside the United States.

Appendix 4. Additional Analysis

4.1 User Feedback

In the survey, we asked respondents to provide feedback on the framework and the task of classifying the example systems using the assigned framework. Analyzing 180 free-text responses using the AI assistant "Elicit", we tagged 66 (37 percent) responses that provided some constructive criticism of the framework or the task, and 64 (36 percent) that provided positive feedback (e.g., "it was well done").[40] 13 respondents (7 percent) provided a response but with no feedback (e.g., "fine as is") while 36 (20 percent) typed a response to indicate they had no feedback (e.g., "none" or "NA").

Of the feedback that offered a criticism of the framework or task, the most common criticism was that the definitions were unclear. The next most common critique was that more examples were needed. Other constructive feedback included clarifying the impact of the framework, clarifying the instructions, enhancing the survey experience, clarifying the rubric, or making the task easier.

4.2 Removing Non-AI Systems

In round one, we found low accuracy and consistency rates among our three "no autonomy" systems (intended to be non-AI systems). We checked to see if removing those classifications increases performance. Removing these systems from the analysis, we see a minor increase in consistent classifications and accurate classifications. In general, our other findings hold with slight (but not necessarily distinguishable) increases in accuracy.

4.3 Expert Respondents

Over the two waves of the survey, we were able to recruit a small sample of 11 U.S. government personnel to classify systems using Frameworks A, B, and C. These users performed somewhat better than our MTurk sample in terms of accuracy. They accurately classified 58 percent of their system dimension classifications,

compared to 52 percent of MTurk respondents. Meaning, on average, these participants accurately classified 5.8 (out of 10) dimensions.

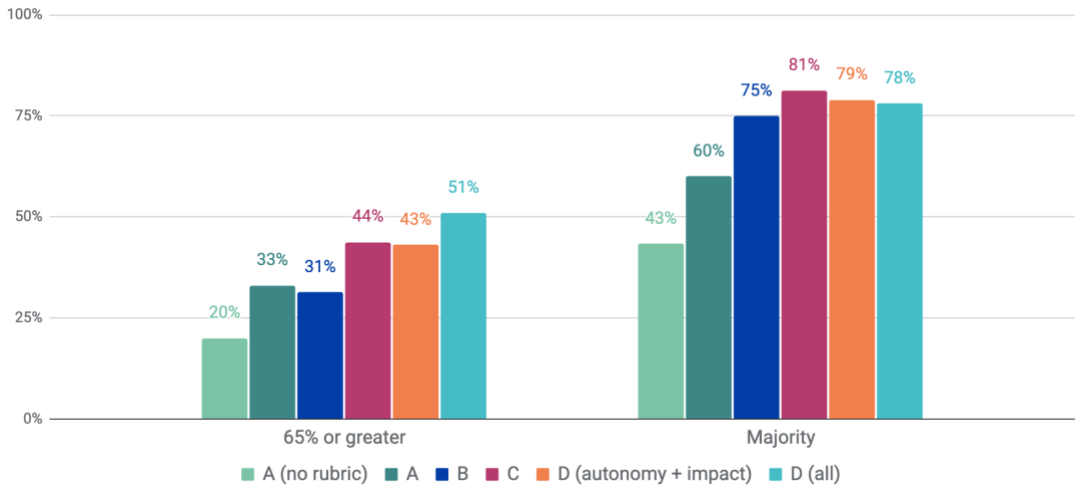
It is more difficult to assess consistency among this sample due to the small number of classifications (since each system was classified by four to six respondents and each framework was used by three to five respondents). Taking all of the classifications from this sample (70 total dimension classifications), expert respondents consistently classified 42 percent of dimensions. While not directly comparable, this is close to the proportion of system dimensions consistently classified by MTurk respondents using Framework C. Similar to MTurk respondents, expert respondents in this sample more consistently classified system impact (57 percent of impact classifications) than system autonomy (29 percent of autonomy classifications). Again, these numbers do align with the proportions of consistent classifications by MTurk respondents using the highest-performing Framework C, suggesting experts may be more equipped to use the frameworks to derive consistent classifications, but with the caveat that this is a small sample and aggregates classifications using Frameworks A, B, and C, so it does not allow for comparison across these frameworks as we do in the main analysis.

4.4 Reducing the Threshold for Classification Consistency

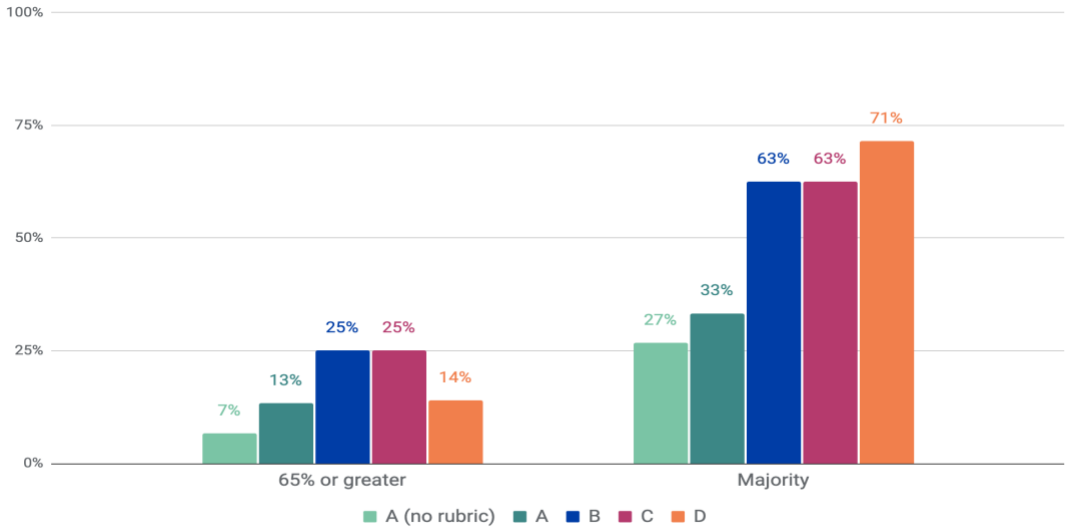
The threshold for consistency used in our main analysis is 65 percent or more users assigning the same classification. This threshold was chosen as a more conservative estimate of majority consistency, to account for a wider margin of error given the sample size for each framework. But we did analyze classification consistency as more than 50 percent of respondents assigning the same classification and report the differences in overall and dimension-specific classification consistency in Figure B.

Figure B. Dimensions with Consistent Classifications by Consistency Thresholds

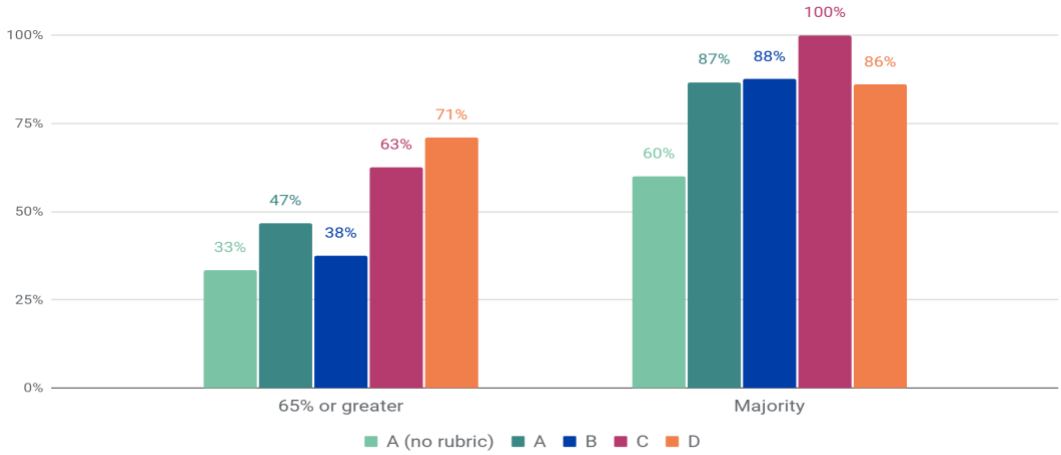
Percentage of Consistent Dimension Classifications, All Dimensions



Percentage of Consistent Dimension Classifications, Autonomy Dimension



Percentage of Consistent Dimension Classifications, Impact Dimension



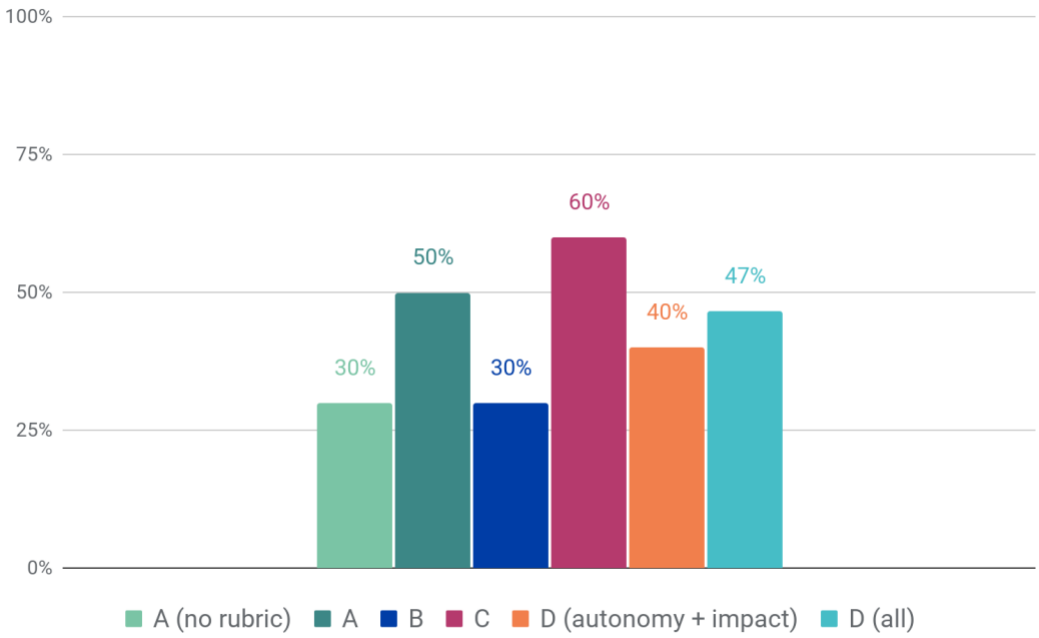
Note: Total system dimensions varied by framework; 30 dimensions for Framework A (15 autonomy and 15 impact), 16 dimensions for Frameworks B and C (8 autonomy and 8 impact), and 14 dimensions for Framework D (7 autonomy and 7 impact). Source: CSET Classifying AI Systems Survey.

As expected, there is a marked increase in consistency across frameworks at the lower threshold, driven largely by an increase in the proportion of autonomy dimensions with consistent classifications in Frameworks B, C, and D—with the percentage of consistent autonomy dimensions more than doubling for Frameworks B and C and tripling for Framework D.

4.5 Comparing Only Systems Classified by Every Framework

Figure C compares classification consistency across all frameworks, as reported in Figure 3 in the main text, but for only the five AI systems classified using every framework.

Figure C. Classification Consistency for Systems Classified for All Frameworks



Note: 10 total system dimensions for Frameworks A, B, C, and D (autonomy + impact) and 45 for D (all). Source: CSET Classifying AI Systems Survey.

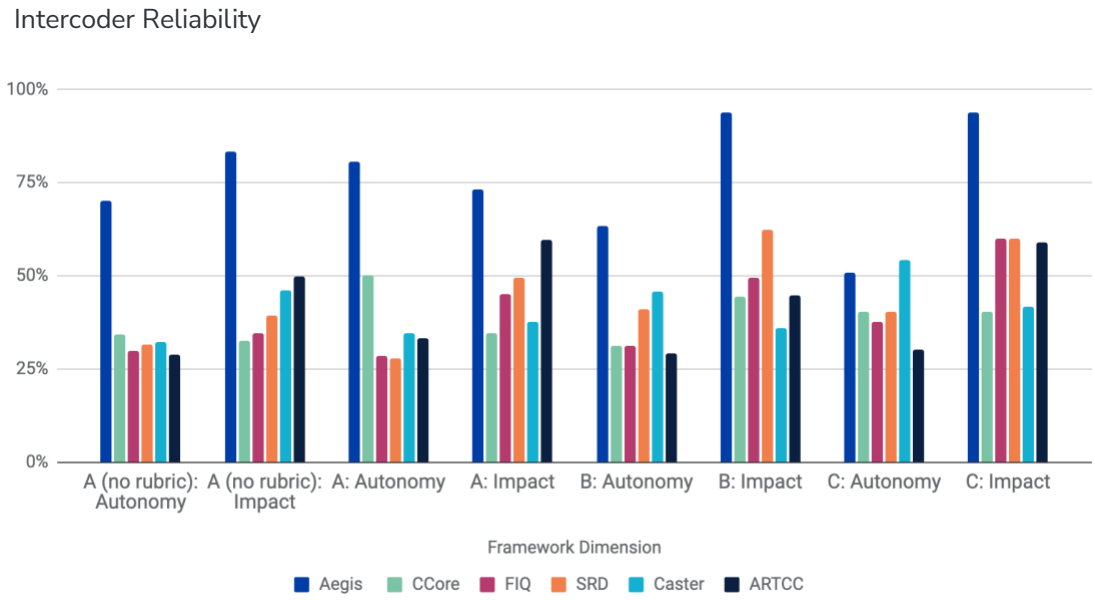
4.6 Calculating Intercoder Reliability

Intercoder reliability provides another assessment of framework quality, so it is provided here for each dimension classification for Frameworks A, B, and C. Because this task involved more than two respondents, intercoder agreement is not zero or one. Rather, agreement is computed by looking at all pairwise agreements among respondents over all respondent pairs for each classified dimension. In other words, for each system dimension (i.e., Aegis autonomy), take the number of respondents that provided a classification for that dimension to calculate the number of respondent pairs that "coded" that dimension, and then calculate

the number of agreements among those pairs. Dividing the number of pair agreements by the number of user pairs for that system dimension results in a value between zero and one, presented here as a percentage.

Intercoder reliability ranged from 94 percent (Aegis impact using Frameworks B and C) to 28 percent (SRD and FIQ autonomy using Framework A with rubric). Figure D shows the intercoder reliability for each framework dimension for each system. While some trends emerge, like Aegis having consistently higher intercoder reliability, there is generally variation across frameworks and systems.

Figure D. Intercoder Reliability for Frameworks A, B, and C, by System Dimension



Source: CSET Classifying AI Systems Survey.

Endnotes

[1] European Commission, "Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence," April 21, 2021, https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682. See also the Commission's earlier *White Paper on Artificial Intelligence – A European Approach to excellence and trust* (Brussels: European Commission, February 19, 2020), https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[2] Exec. Order No. 13960, 85 FR 78939 (2020); Brian Stanton and Theodore Jensen, *Trust and Artificial Intelligence* (Washington, DC: National Institute of Standards and Technology, 2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf; For a list of existing AI ethical principles and frameworks, check out the Awesome AI Guidelines GitHub repository, accessed at https://github.com/EthicalML/awesome-artificial-intelligence-guidelines/.

[3] AI governance is meant here as a broad term, to refer to processes for developing global norms, policies, and institutions to best ensure the responsible and beneficial development and use of AI. For more discussion on the concept of AI governance, see research by the Centre for the Governance of AI at https://governance.ai/research.

[4] In previous research, CSET argued for an evolving definition grounded in three criteria: the regularly updated judgments of experts, active research methods and tasks, and examples of AI-relevant research and products. See Dewey Murdick, James Dunham, and Jennifer Melot, "AI Definitions Affect Policymaking" (Center for Security and Emerging Technology, June 2020), http://doi.org/10.51593/20200004. Others highlight divergent definitions among practitioners and policymakers and propose criteria for a "policy-facing definition" that encompasses deployed and future AI applications, is accessible for non-expert audiences, and allows for policy implementation. See P. M. Krafft et al., "Defining AI in Policy versus Practice," in *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, February 2020, https://doi.org/10.1145/3375627.3375835.

[5] A framework goes beyond a definition to allow for system characterization along multiple dimensions. A framework is more structured and condensed compared to a primer or attempt to explain or describe AI or AI systems in prose. For a helpful primer on AI targeted to a policymaking audience, see Greg Allen, *Understanding AI Technology* (Washington, DC: Department of Defense, 2020), https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf.

[6] Darrell M. West, "What is artificial intelligence?" *Brookings*, October 4, 2018, https://www.brookings.edu/research/what-is-artificial-intelligence/.

[7] Zachary Arnold and Helen Toner, "AI Accidents: An Emerging Threat" (Center for Security and Emerging Technology, July 2021), https://doi.org/10.51593/20200072; James E. Baker, "Ethics and Artificial Intelligence: A Policymaker's Introduction" (Center for Security and Emerging Technology, April 2021), https://doi.org/10.51593/20190022.

[8] "AI and Efficiency," *OpenAI Blog*, May 5, 2020, https://openai.com/blog/ai-and-efficiency/.

[9] For more information on AlphaGo Zero see David Silver and Demis Hassabis, "AlphaGo Zero: Starting from scratch," *DeepMind Blog*, October 18, 2017, https://deepmind.com/blog/article/alphago-zero-starting-scratch.

[10] Open Data Institute's 2018 report classifies systems along five categories: 1. Proprietary data model: open algorithm and closed data; 2. Closed model: closed algorithm and closed data; 3. Proprietary algorithm model: closed algorithm and open data; 4. Open model: open algorithm and open data; 5. Shared model: data and algorithms are shared only by select parties. The other referenced framework comes from Francesco Corea, "AI Knowledge Map: how to classify AI technologies," *Medium*, August 29, 2018, https://francesco-ai.medium.com/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020. In this AI Knowledge Map, there are three AI paradigms—symbolic, statistical, and subsymbolic—that encompass eight approaches: logic-based, knowledge-based, probabilistic methods, supervised machine learning, unsupervised machine learning, reinforcement learning, embodied intelligence, and search and optimisation. There are five AI problem domains: perception, reasoning, knowledge, planning, and communication.

[11] AI is not the same thing as autonomy, and autonomy and automation are not the same thing. Different AI systems have varying degrees of autonomy, but a system can be wholly autonomous without being AI (e.g., a landmine). The author thanks Matthew Daniels for pointing out the need to clarify this distinction. See also Lt. Col. Aaron Celaya and Sriraj Aiyer, "A Foundation of Automation for Future Artificial Intelligence Strategy," *Journal of the Homeland Defense & Security Information Analysis* 7, no. 1 (June 15, 2020), https://www.hdiac.org/journal-article/a-foundation-of-automation-for-future-artificial-intelligence-strategy/; R. Parasuraman et al., "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, no. 3 (2000): 286–297, https://ieeexplore.ieee.org/document/844354.

[12] European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence

(ARTIFICIAL INTELLIGENCE ACT) and Amending Certain Union Legislative Acts," April 21, 2021, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence.

[13] For an overview of Canada's Algorithmic Impact Assessment Tool, see "Algorithmic Impact Assessment Tool," Government of Canada, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[14] Germany's Data Ethics Commission Risk Criticality Pyramid (2020) informed by an earlier report (2019) that proposes a risk classification matrix (see section 3). Another example is the AI Impact Assessment developed by Open Loop, an initiative launched in 2021 to connect global policymakers and technology companies to help develop evidence-based policies around AI and other emerging technologies, see Norberto Nuno Gomes Andrade and Verena Kontschieder, "AI Impact Assessment: A Policy Prototyping Experiment" (Open Loop, 2021), https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf.

[15] Russell T. Vought, *Guidance for Regulation of Artificial Intelligence Applications*, M-21-06 (Washington, DC: Office of Management and Budget, November 17, 2020), https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.

[16] The DHS AI Strategy was released on December 3, 2020 and can be accessed at https://www.dhs.gov/sites/default/files/publications/dhs_ai_strategy.pdf.

[17] For more information on the OECD Network of Experts on AI and the research of the working group on the Classification of AI, see "Public consultation on the OECD Framework for Classifying AI Systems," OECD.AI Policy Observatory, May 31, 2021, https://oecd.ai/wonk/classification.

[18] This definition includes the note that, "AI systems are designed to operate with varying levels of autonomy." Organisation for Economic Co-operation and Development (OECD), *Artificial Intelligence in Society* (OECD Publishing, June 11, 2019), https://doi.org/10.1787/eedfee77-en; OECD Directorate for Science, Technology and Innovation, *OECD Framework for the Classification of AI Systems – Public Consultation on Preliminary Findings* (OECD Publishing, May 20, 2021), https://aipo-api.buddyweb.fr/app/uploads/2021/06/Report-for-consultation_OECD.AI_Classification.pdf.

[19] See Appendix 2 for full description of each framework.

[20] See Appendix 2 for the full survey questionnaire or to take the anonymous survey.

[21] The example system descriptions were reviewed by the small sample of experts who provided a baseline, expert classification for each system. They were also reviewed by other CSET analysts and OECD working group members. The goal was to ensure the descriptions provided accurate representations of the system and included adequate information for classification, without providing the exact wording for choosing the correct classification. In a couple of cases, the descriptions were written or reviewed by system developers.

[22] The example systems included in the two rounds of the survey varied slightly. The first round (Framework A) included eight additional systems and did not include SCORE or AGZ. For the second round (Frameworks B, C, and D), we removed eight systems but added SCORE or AGZ for a total of seven AI systems in that round. See Appendix 3 for a full list of systems included in both rounds and the full system descriptions provided in the survey.

[23] See Appendix 1 for more information on survey respondents and survey distribution.

[24] See Appendix 2. In the first round, the experimental treatment was the inclusion of a framework rubric. All respondents were asked to review and use the same framework, but only half were provided with the rubric to guide them while they were classifying the systems. The goal at that stage was to understand whether users using the same framework were better able to classify a system with the assistance of the rubric. We found that classifications were more consistent and accurate when the rubric was available to users, so in the second round all respondents were provided with a corresponding framework rubric, but the framework they were asked to review and use varied.

[25] The survey is anonymous and no identifying information is collected.

[26] An alternate way of assessing the quality of the frameworks, in terms of their ability to guide human users to consistent and accurate classifications, is by calculating intercoder reliability. See Appendix 4 for a brief discussion of our analysis of intercoder reliability, which reinforces our reported findings.

[27] 65 percent was chosen as the threshold to confidently assume, accounting for margin of error, that more than half of users assigned the same classification. To determine expert classification, each system was classified using the framework by at least one expert. Usually three to five experts, including CSET researchers and members of ONE.AI, classified each system. All expert classifications were consulted to develop one, consensus expert classification for each example system.

[28] The systems listed in Table 2 were classified by respondents for each framework, with the exception of SCORE and AGZ, which were not included in the survey testing Framework A. Air route traffic control center (ARTCC) was an example of a non-AI system included for testing Frameworks A, B, and C but not D. ARTCC was included in the survey for testing Frameworks A, B, and C with the goal of assessing whether users could use the framework to differentiate between AI and non-AI systems. This example was not included in Framework D. See Appendix 3 for more details about the example systems included.

[29] Instead of including a "no autonomy" level as in Framework A, Frameworks B and C instead offered a "not an AI system option" that a user could select instead of assigning the example system an autonomy and impact level classification. Every respondent had one example system that was not an AI system and were informed in the instructions that one of the five systems they received would not be an AI system.

[30] Again, this difference is starker when we look at the more than 50 percent consistency threshold, instead of 65 percent or more, suggesting the rubric is a useful tool for classification, even though it may not produce 100 percent consistency. See Appendix 4.

[31] Additional framework dimensions were being considered by the ONE.AI working group at the time this survey was designed and distributed and have since been included in the OECD classification framework, which can be accessed at https://oecd.ai/wonk/classification.

[32] 48 percent of dimensions were assigned the accurate classification. As noted in Table 3, two dimension classifications were consistent but did not match the expert classification. Specifically for the Caster system, 76 percent of users assigned impact and autonomy levels that did not match the expert classification.

[33] The survey allowed respondents to pick as many task classifications as they saw fit, which may have contributed to the higher classification performance.

[34] Caster did not quite meet the threshold, but 60 percent of users selected an accurate system task.

[35] At this threshold, the systems with consistent autonomy classifications tended to be those with the highest level of autonomy (high/significant/action).

[36] For example, a user who classified five systems using a two dimension framework classified a total of ten system dimensions. If that user accurately classified two system autonomy and four impact dimensions, she had an

average correct score of six. We did not assess classification consistency at the individual user level.

[37] The difference in proportion of accurate classifications between no rubric and rubric conditions for Framework A was statistically significant (at the 95 percent confidence level for a one-tailed test). Framework A is the only framework for which we varied inclusion of the rubric to support user classification. The reported percentages are based on 970 total dimension classifications completed without the rubric and 850 total dimension classifications completed with the rubric.

[38] The Artificial Intelligence Incident Database can be accessed at https://incidentdatabase.ai/. In addition to the research described here, CSET partnered with PAI to develop and publish a taxonomy of AI incidents, tracking features such as severity of harm and location of incidents. To learn more about the CSET taxonomy and work on AI incidents see Artificial Intelligence Incident Database, "CSET Taxonomy," Partnership on AI, https://incidentdatabase.ai/taxonomy/cset and Arnold and Toner, "AI Accidents."

[39] Since the current corpus of documents is news articles from PAI's AI Incidents database, the identified AI systems are not representative of all systems, rather it is exclusively systems that were involved in an accident or incident covered by the media. While this skews the sample of classified systems, it offers a first step, proof-of-concept to assess whether system information required to classify a system can be extracted, whether by humans or machines, from public text.

[40] Thanks to Jungwon Byun for exploring this task using Elicit and providing these preliminary results. Elicit is a research assistant developed by the applied research lab Ought, which makes use of the GPT-3 language transformer developed by OpenAI. See "Elicit: The AI Research Assistant," https://elicit.org; and Jungwon Byun and Andreas Stuhlmüller, "Automating reasoning about the future at Ought," Ought, November 9, 2020, https://ought.org/updates/2020-11-09-forecasting.