

US 20230360388A1

(19) **United States**

(12) **Patent Application Publication**  
**SINGH**

(10) **Pub. No.: US 2023/0360388 A1**

(43) **Pub. Date: Nov. 9, 2023**

(54) **TRAINING A GENERATIVE ARTIFICIAL INTELLIGENCE / MACHINE LEARNING MODEL TO RECOGNIZE APPLICATIONS, SCREENS, AND USER INTERFACE ELEMENTS USING COMPUTER VISION**

(52) **U.S. Cl.**  
CPC ..... *G06V 10/82* (2022.01); *G06V 10/774* (2022.01); *G06V 40/20* (2022.01); *G06V 2201/02* (2022.01)

(71) Applicant: **UiPath, Inc.**, New York, NY (US)

(57) **ABSTRACT**

(72) Inventor: **Prabhdeep SINGH**, Bellevue, WA (US)

(73) Assignee: **UiPath, Inc.**, New York, NY (US)

(21) Appl. No.: **18/355,877**

(22) Filed: **Jul. 20, 2023**

**Related U.S. Application Data**

(63) Continuation-in-part of application No. 17/070,108, filed on Oct. 14, 2020.

**Publication Classification**

(51) **Int. Cl.**  
*G06V 40/20* (2006.01)  
*G06V 10/82* (2006.01)  
*G06V 10/774* (2006.01)

Techniques for training a generative artificial intelligence (AI) / machine learning (ML) model to recognize applications, screens, and UI elements using computer vision (CV) and to recognize user interactions with the applications, screens, and UI elements are disclosed. Optical character recognition (OCR) may also be used to assist in training the generative AI/ML model. Training of the generative AI/ML model may be performed without other system inputs such as system-level information (e.g., key presses, mouse clicks, locations, operating system operations, etc.) or application-level information (e.g., information from an application programming interface (API) from a software application executing on a computing system), or the training of the generative AI/ML model may be supplemented by other information, such as browser history, heat maps, file information, currently running applications and locations, system level and/or application-level information, etc.

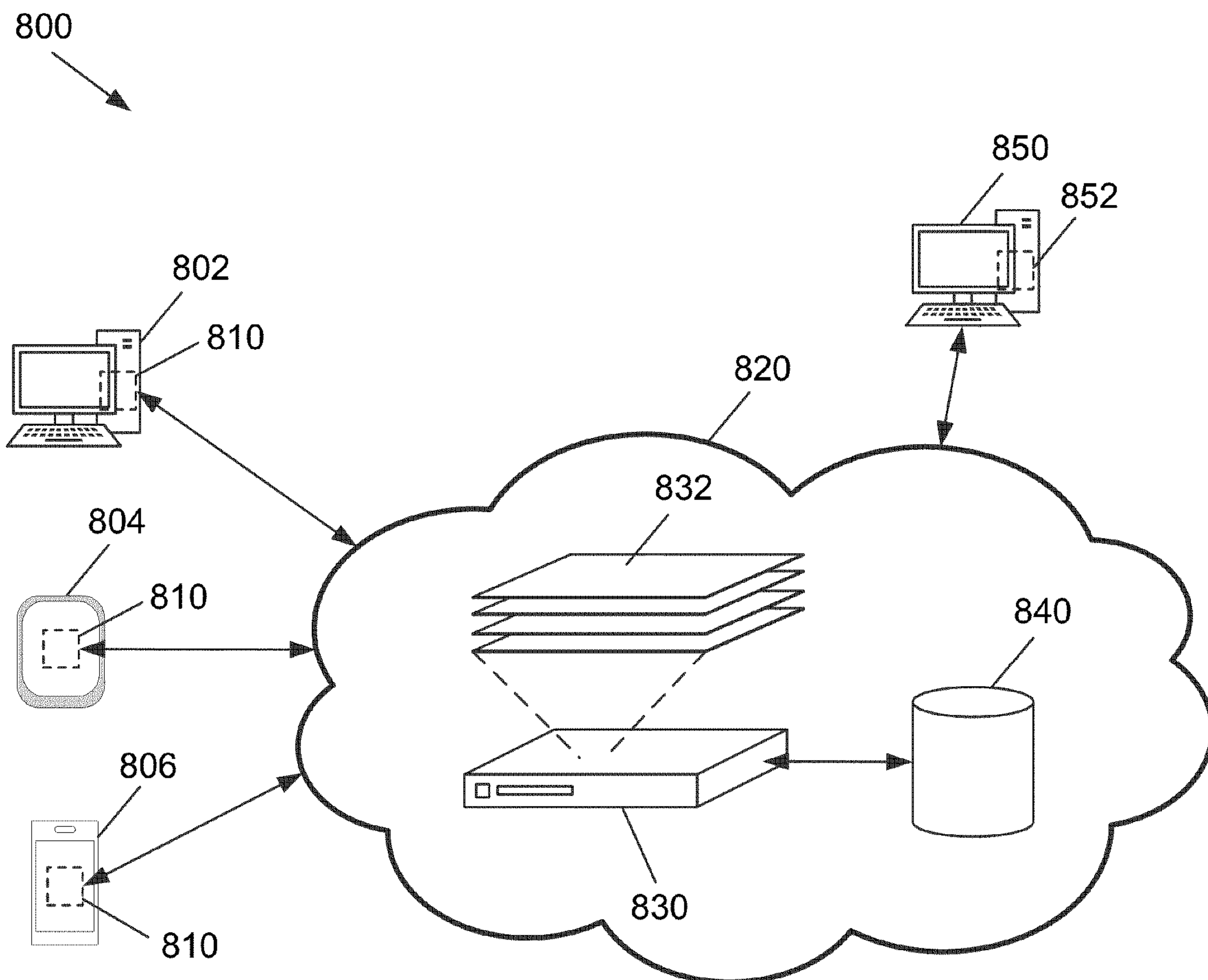


FIG. 1

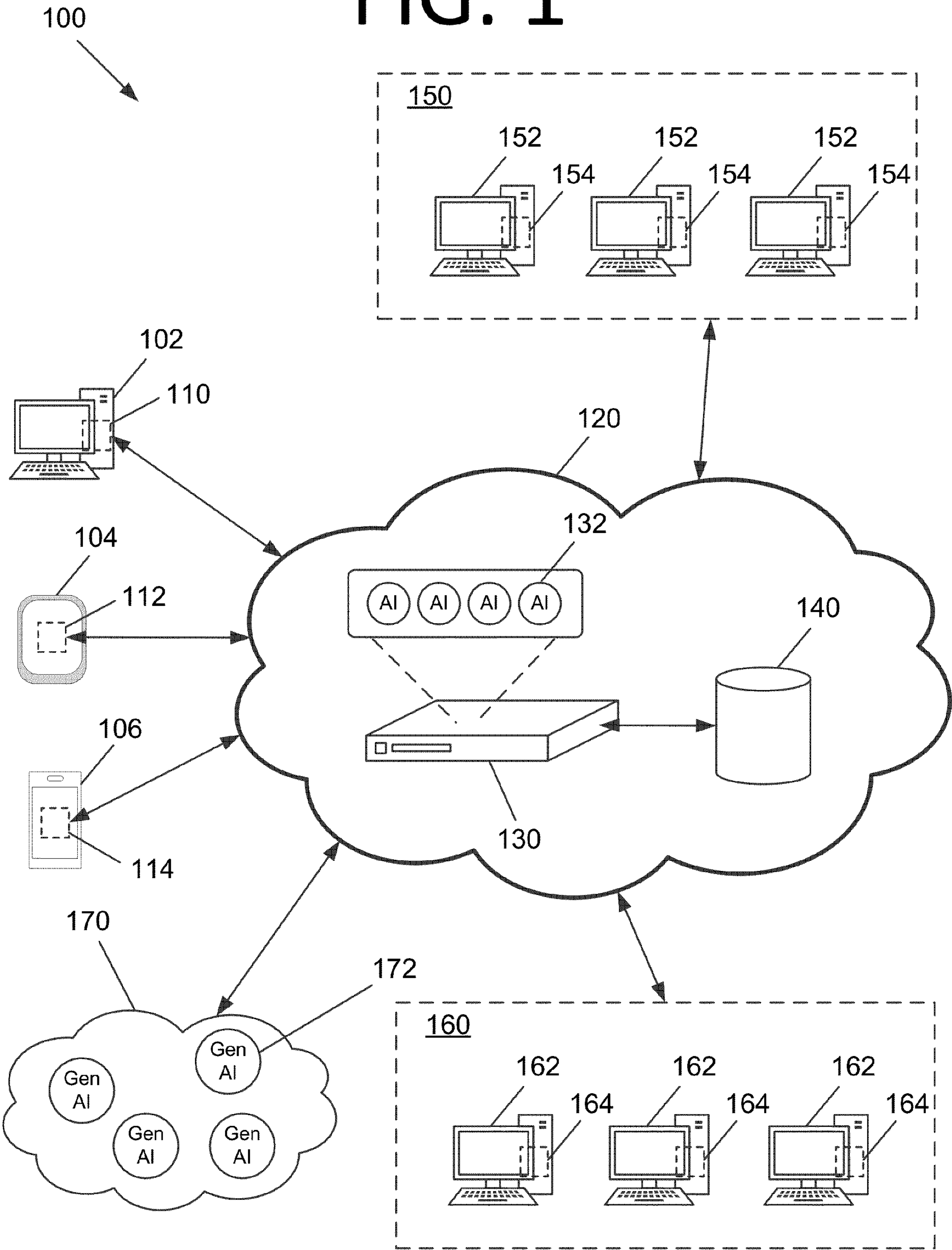
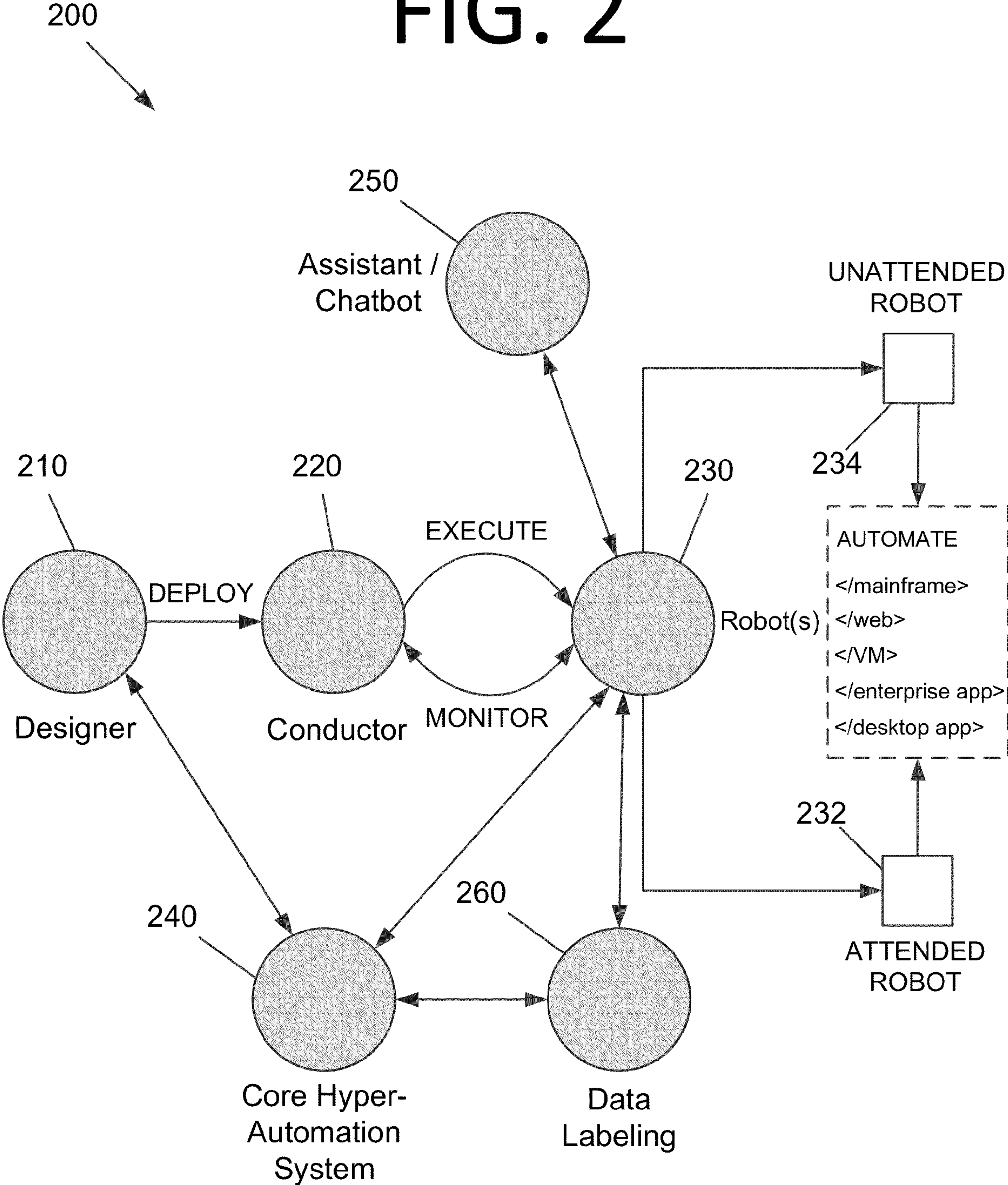




FIG. 2



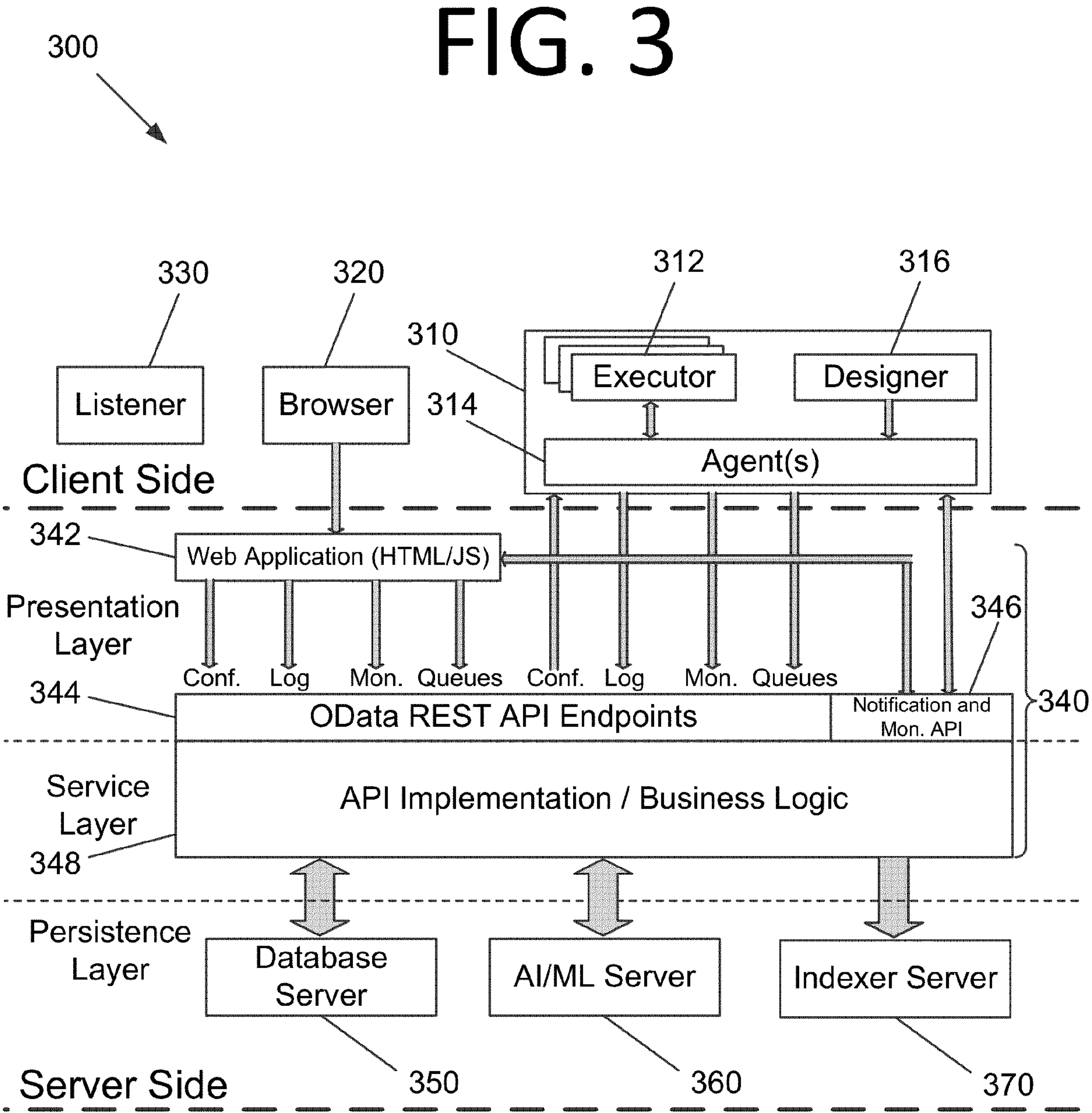
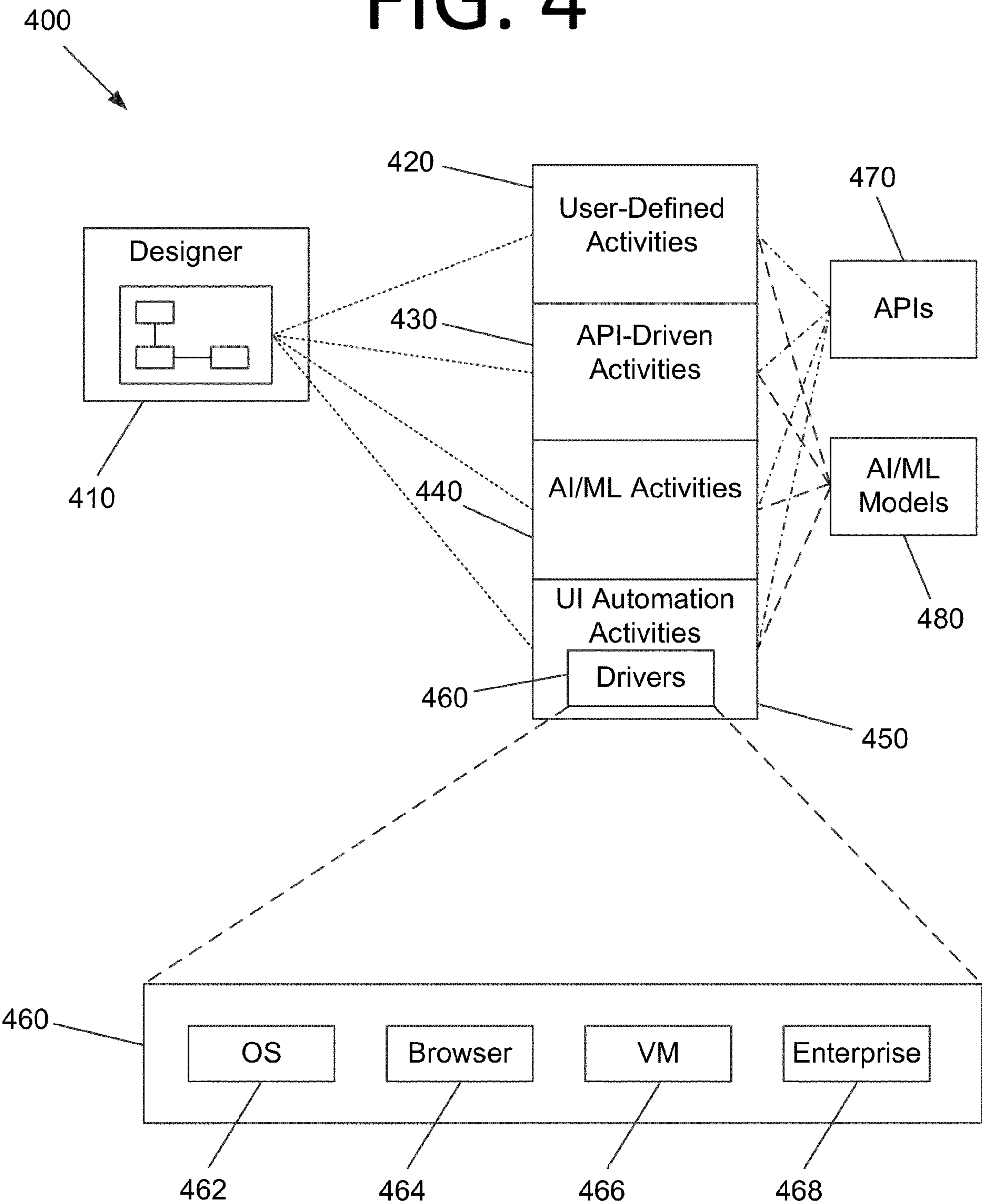
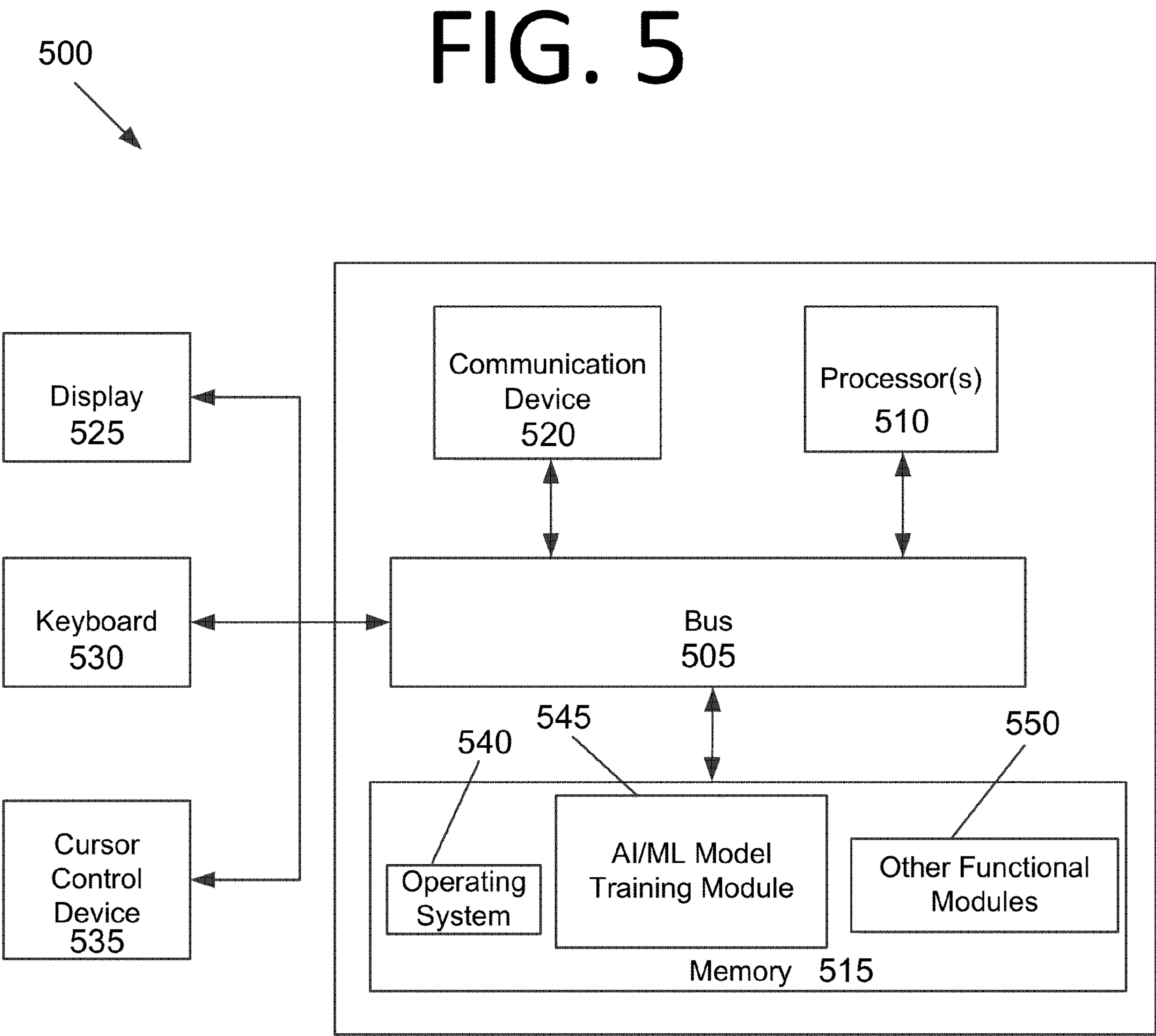


FIG. 4







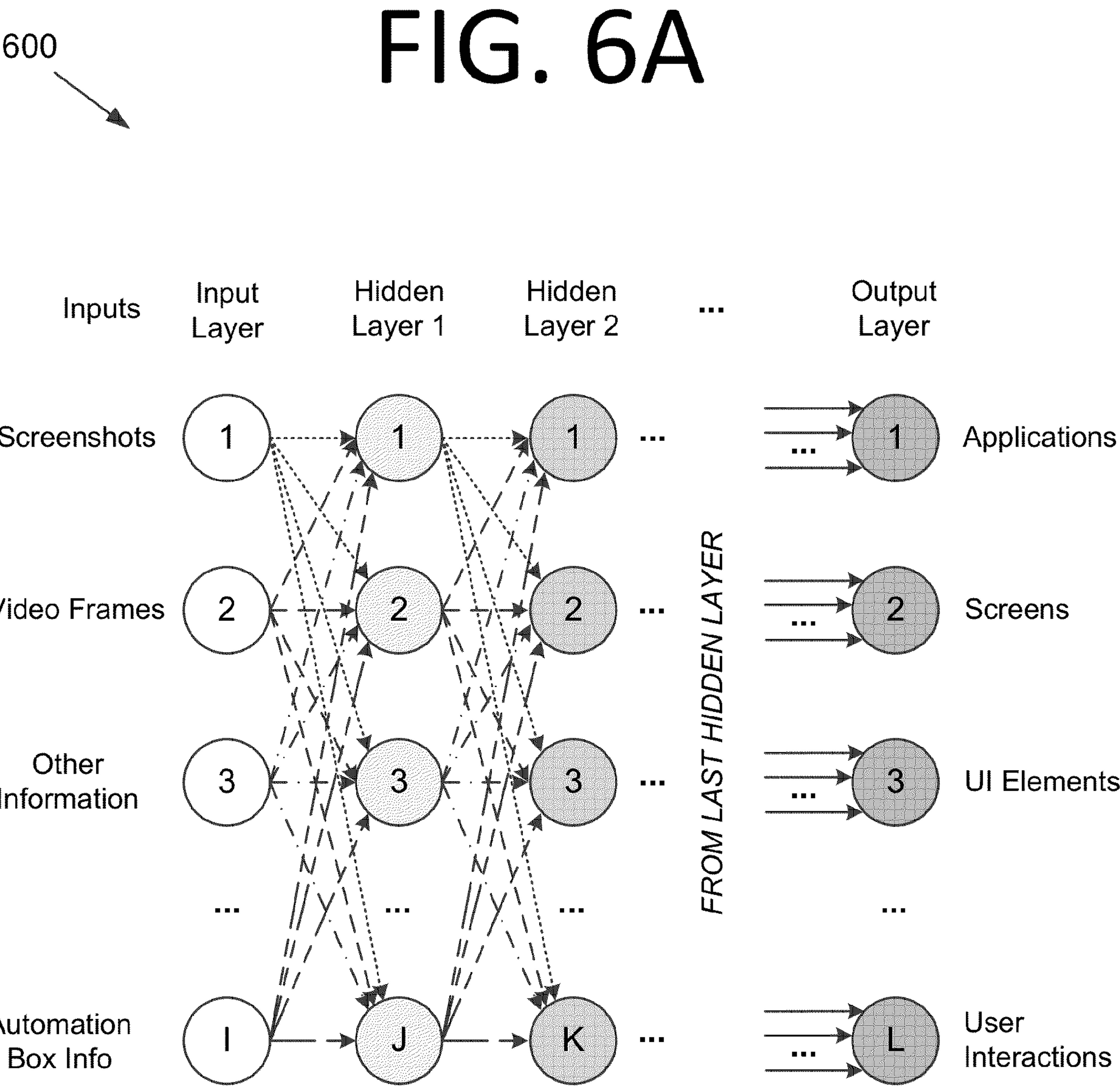
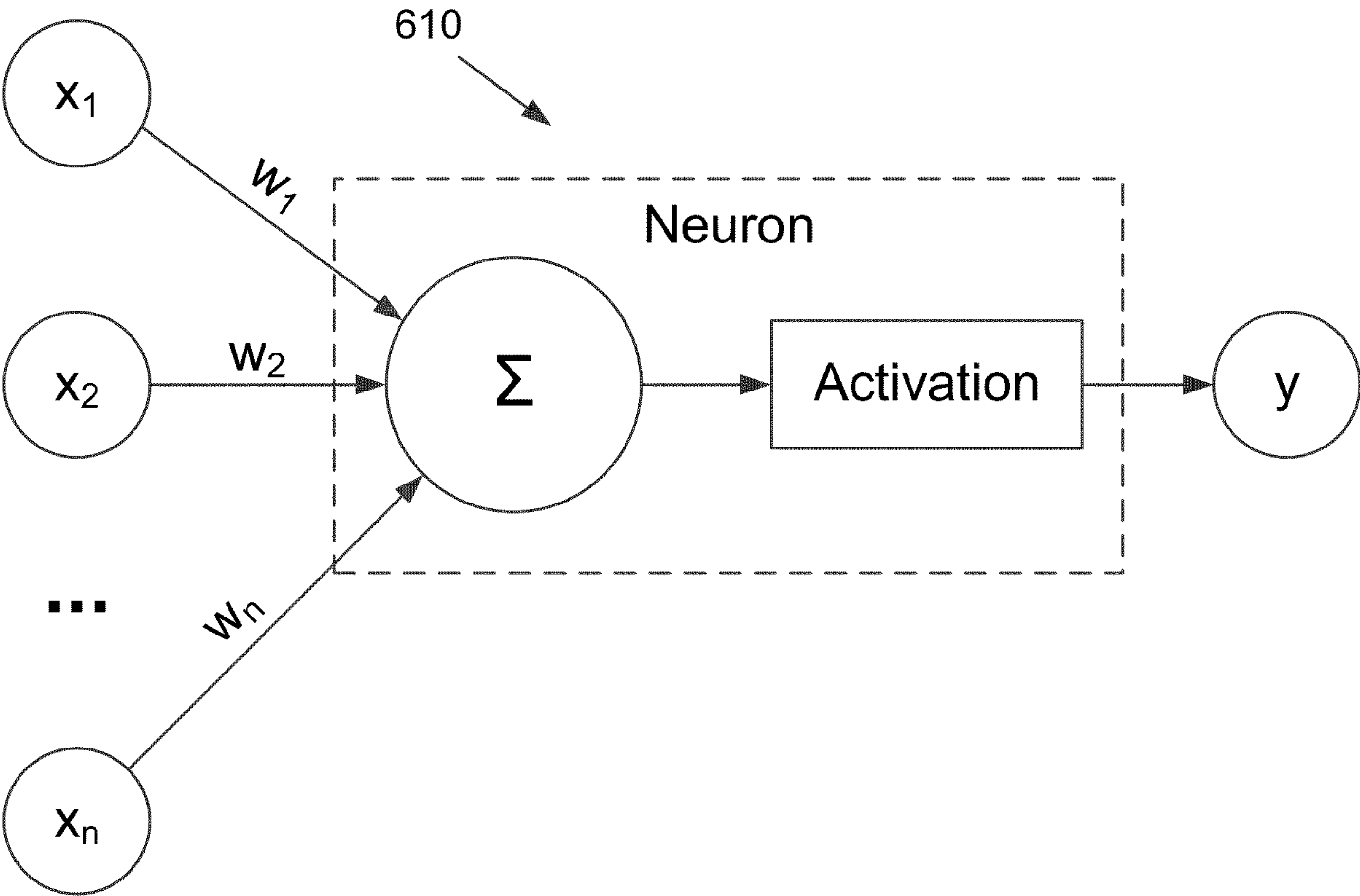


FIG. 6B





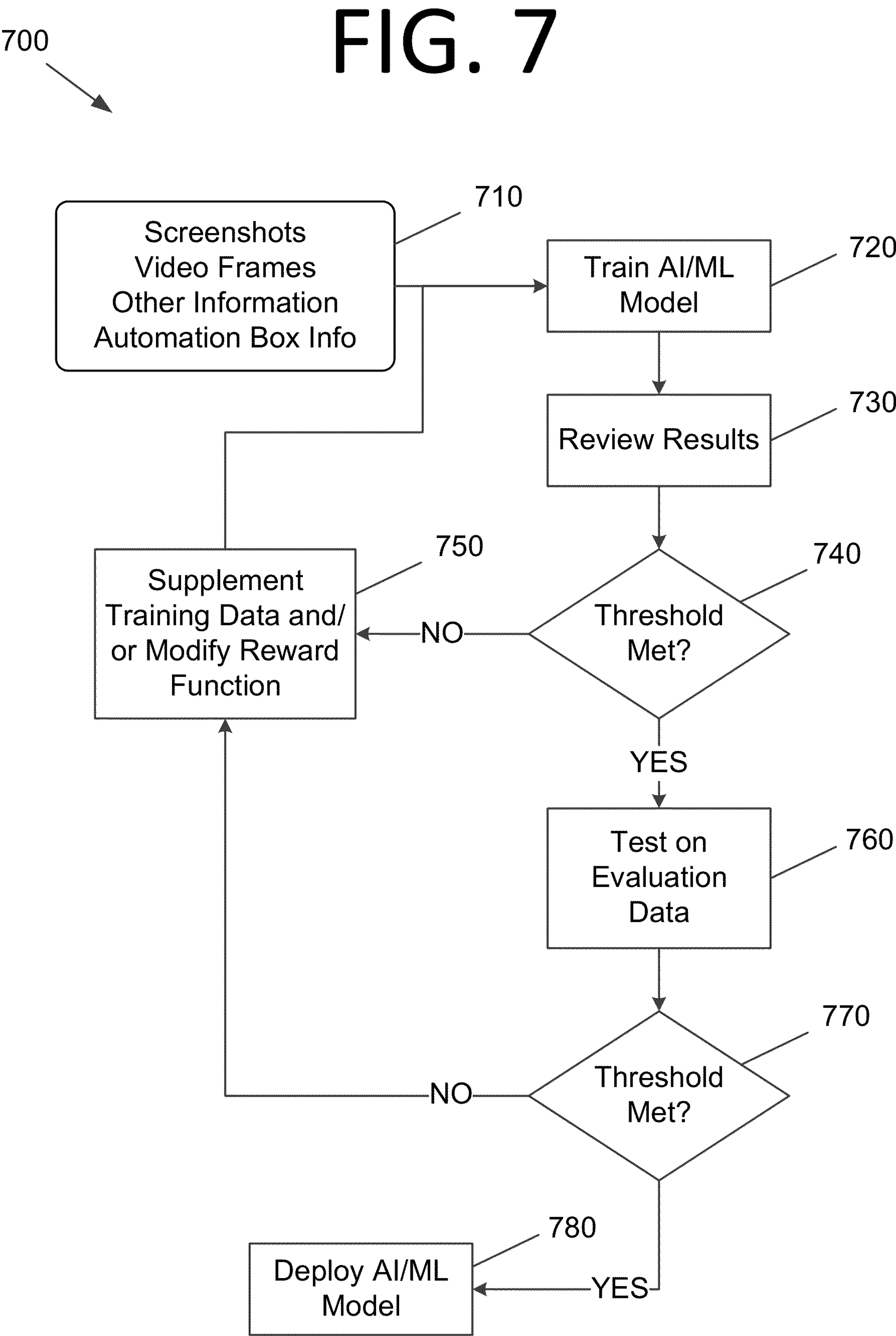
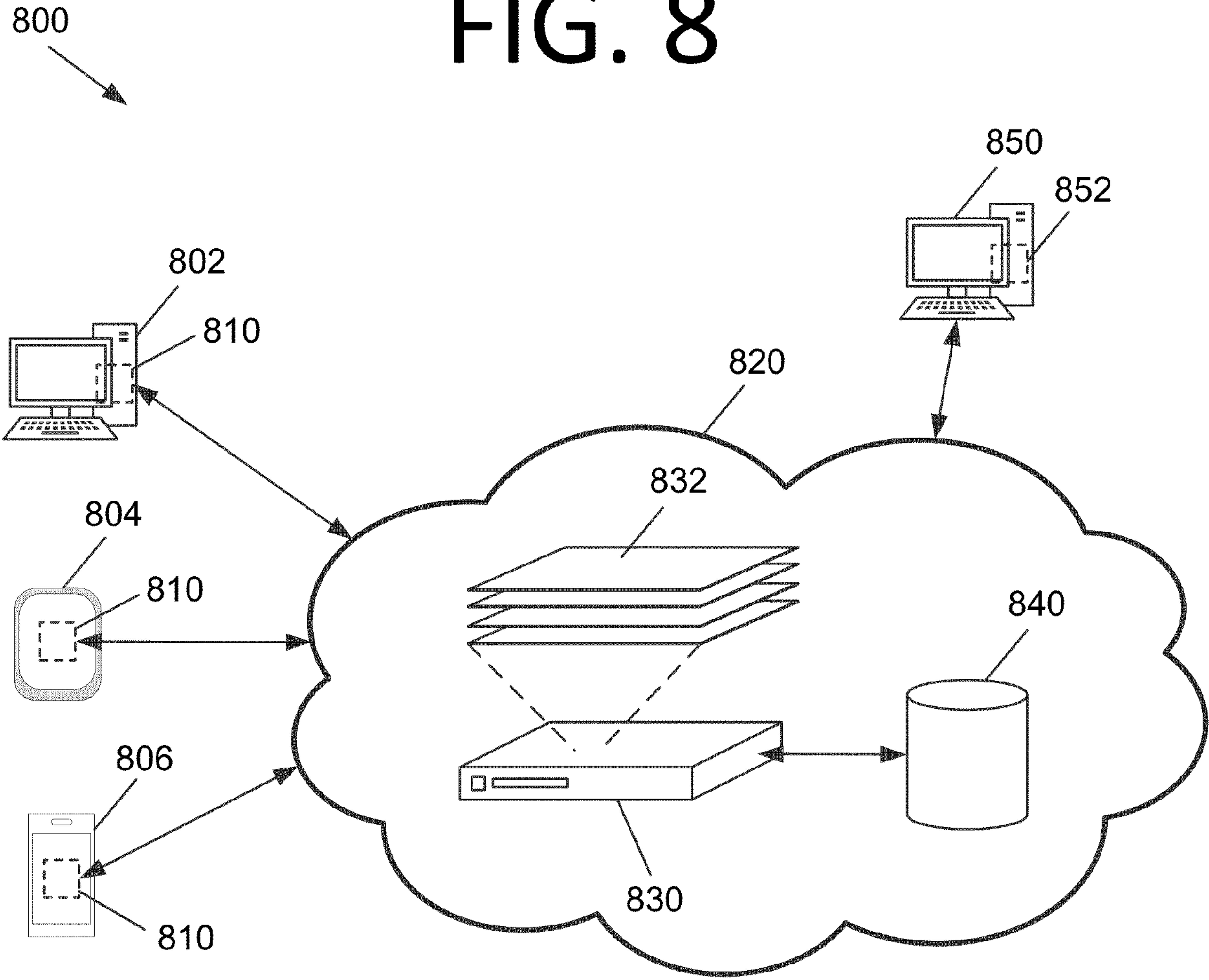
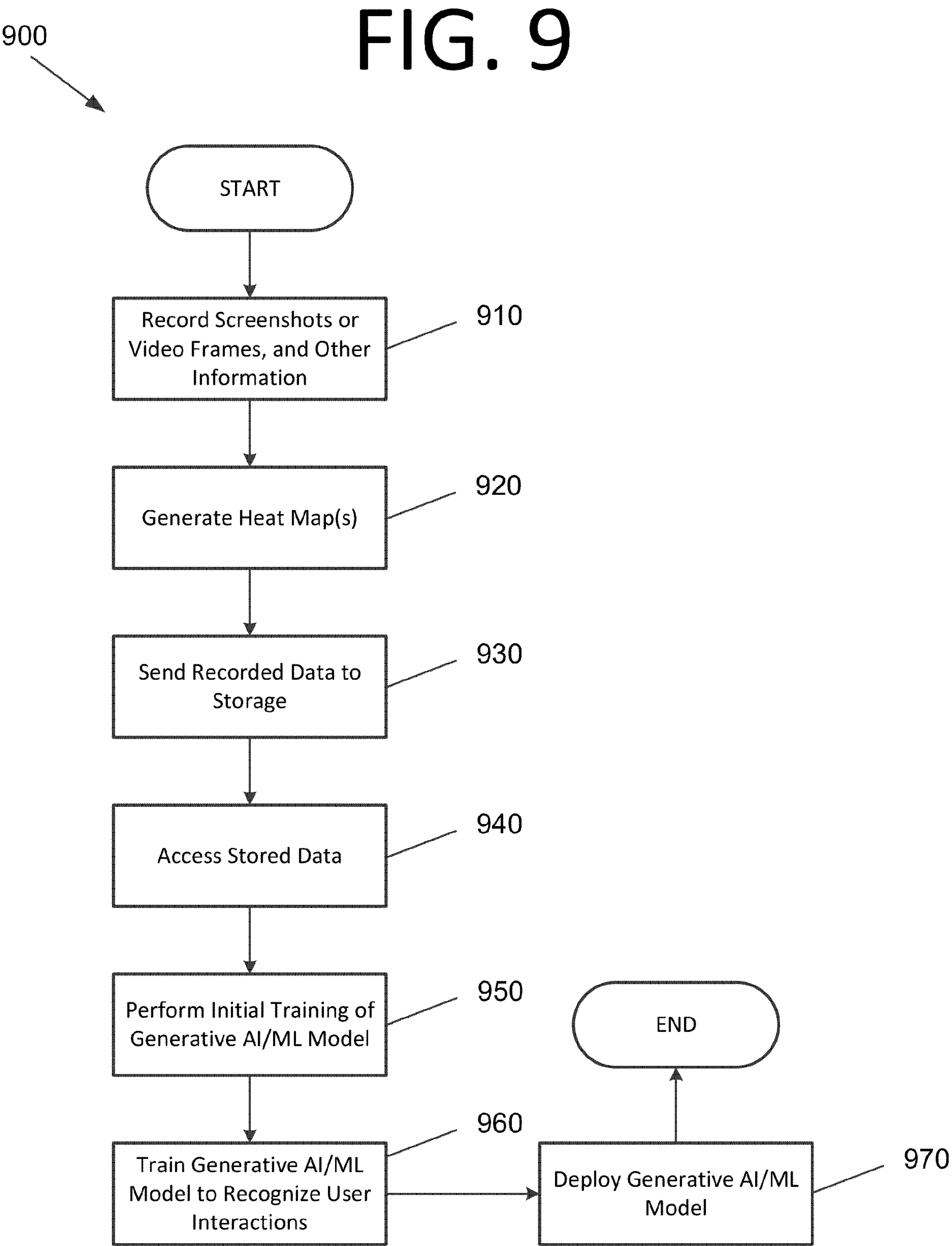
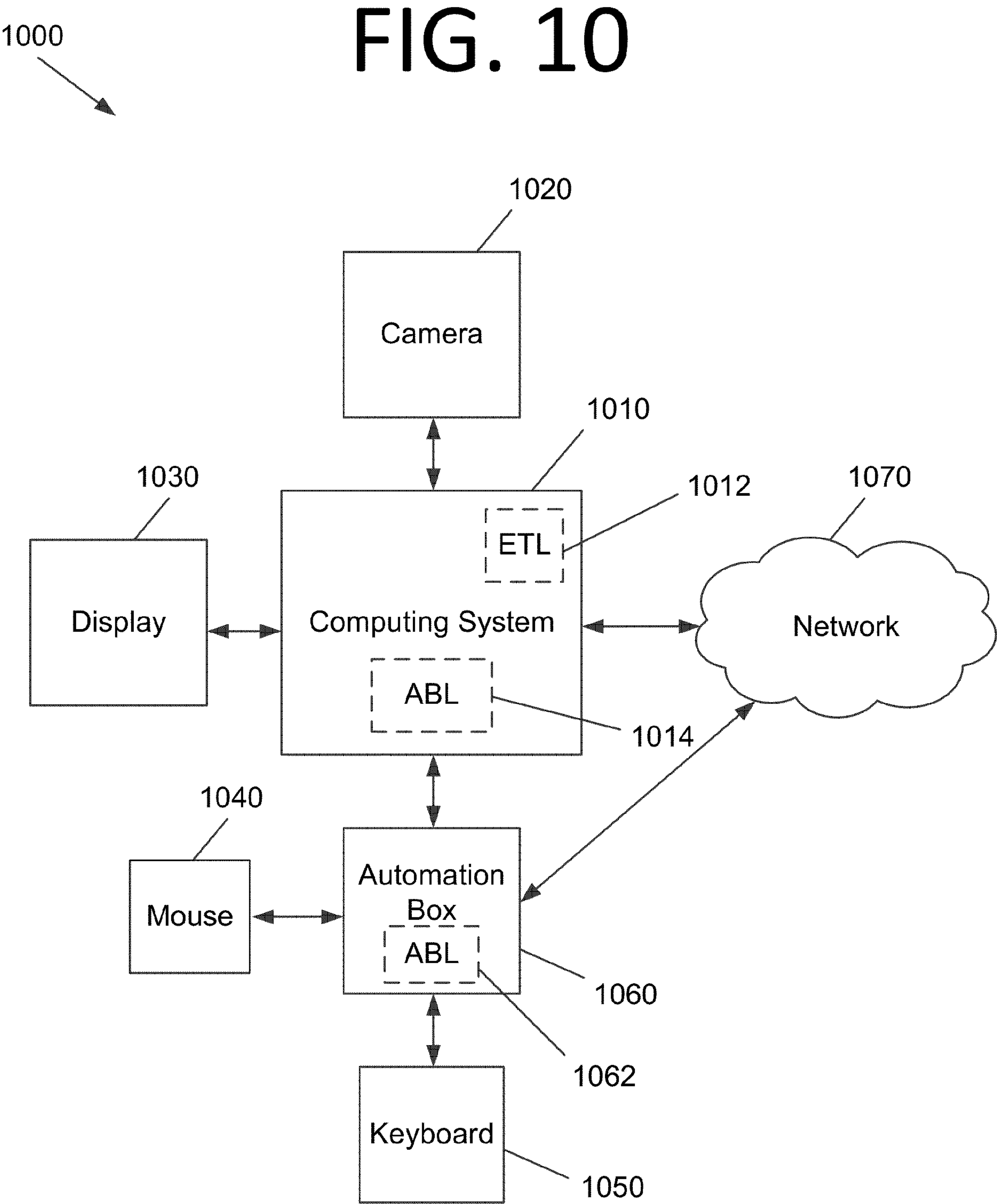


FIG. 8









**TRAINING A GENERATIVE ARTIFICIAL  
INTELLIGENCE / MACHINE LEARNING  
MODEL TO RECOGNIZE APPLICATIONS,  
SCREENS, AND USER INTERFACE  
ELEMENTS USING COMPUTER VISION**

**CROSS REFERENCE TO RELATED  
APPLICATION**

**[0001]** This application is a continuation-in-part (CIP) of, and claims the benefit of, U.S. Nonprovisional Pat. Application No. 17/070,108 filed Oct. 14, 2020. The subject matter of this earlier filed application is hereby incorporated by reference in its entirety.

**FIELD**

**[0002]** The present invention generally relates to user interface (UI) automation, and more specifically, to training a generative artificial intelligence (AI) / machine learning (ML) model to recognize applications, screens, and UI elements using computer vision (CV) and to recognize user interactions with the applications, screens, and UI elements.

**BACKGROUND**

**[0003]** In order to perform UI automation, RPA technologies may utilize driver and/or application-level interactions to click buttons, enter text, and perform other interactions with a UI. However, key presses, mouse clicks, and other kernel hook information may not be available at the system level in some embodiments, or when building a new UI automation platform. Extensive driver level and application level functionality is typically required to implement such a UI automation platform. Accordingly, alternative techniques for providing UI automation may be beneficial.

**SUMMARY**

**[0004]** Certain embodiments of the present invention may provide solutions to the problems and needs in the art that have not yet been fully identified, appreciated, or solved by current UI automation technologies. For example, some embodiments of the present invention pertain to training a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements.

**[0005]** In an embodiment, a system includes one or more user computing systems including respective recorder processes and a server configured to train a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements. The respective recorder processes are configured to record screenshots or video frames of a display associated with the respective user computing system and other information. The respective recorder processes are also configured to send the recorded screenshots or video frames, and the other information, to storage accessible by the server. The server is configured to initially train the generative AI/ML model to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information. After the generative AI/ML model can recognize the applications, screens, and UI elements in the recorded screen-

shots or video frames with a confidence, the server is also configured to train the generative AI/ML model to recognize individual user interactions with the UI elements.

**[0006]** In another embodiment, a non-transitory computer-readable medium stores a computer program configured to train a generative AI/ML model to recognize applications, screens, and UI elements using CV and/or to recognize user interactions with the applications, screens, and UI elements. The computer program is configured to cause at least one processor to access recorded screenshots or video frames of displays associated with one or more computing systems and access other information associated with the one or more computing systems. The computer program is also configured to cause at least one processor to initially train the generative AI/ML model to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information. The initial training of the generative AI/ML model is performed without a priori knowledge of the applications, screens, and UI elements in the screenshots or video frames.

**[0007]** In yet another embodiment, a computer-implemented method for training a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements includes accessing recorded screenshots or video frames of displays associated with one or more computing systems and accessing other information associated with the one or more computing systems. The computer-implemented method also includes initially training the generative AI/ML model to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information. After the generative AI/ML model can recognize the applications, screens, and UI elements in the recorded screenshots or video frames with a confidence, the computer-implemented method further includes training the generative AI/ML model to recognize individual user interactions with the UI elements.

**BRIEF DESCRIPTION OF THE DRAWINGS**

**[0008]** In order that the advantages of certain embodiments of the invention will be readily understood, a more particular description of the invention briefly described above will be rendered by reference to specific embodiments that are illustrated in the appended drawings. While it should be understood that these drawings depict only typical embodiments of the invention and are not therefore to be considered to be limiting of its scope, the invention will be described and explained with additional specificity and detail through the use of the accompanying drawings, in which:

**[0009]** FIG. 1 is an architectural diagram illustrating a hyper-automation system, according to an embodiment of the present invention.

**[0010]** FIG. 2 is an architectural diagram illustrating an RPA system, according to an embodiment of the present invention.

**[0011]** FIG. 3 is an architectural diagram illustrating a deployed RPA system, according to an embodiment of the present invention.



**[0012]** FIG. 4 is an architectural diagram illustrating the relationship between a designer, activities, and drivers, according to an embodiment of the present invention.

**[0013]** FIG. 5 is an architectural diagram illustrating a computing system configured to train an AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements using generative AI, according to an embodiment of the present invention.

**[0014]** FIG. 6A illustrates an example of a neural network that has been trained to recognize user interactions with applications, screens, and UI elements, according to an embodiment of the present invention.

**[0015]** FIG. 6B illustrates an example of a neuron, according to an embodiment of the present invention.

**[0016]** FIG. 7 is a flowchart illustrating a process for training AI/ML model(s), according to an embodiment of the present invention.

**[0017]** FIG. 8 is an architectural diagram illustrating a system configured to train a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements, according to an embodiment of the present invention.

**[0018]** FIG. 9 is a flowchart illustrating a process for training a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements, according to an embodiment of the present invention.

**[0019]** FIG. 10 is an architectural diagram illustrating an automation box and eye movement tracking system, according to an embodiment of the present invention.

**[0020]** Unless otherwise indicated, similar reference characters denote corresponding features consistently throughout the attached drawings.

#### DETAILED DESCRIPTION OF THE EMBODIMENTS

**[0021]** Some embodiments pertain to training a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements. In certain embodiments, optical character recognition (OCR) may also be used to assist in training the generative AI/ML model. In some embodiments, training of the generative AI/ML model may be performed without other system inputs such as system-level information (e.g., key presses, mouse clicks, locations, operating system operations, etc.) or application-level information (e.g., information from an application programming interface (API) from a software application executing on a computing system), such as that provided by the driver of UiPath Studio™. However, in certain embodiments, the training of the generative AI/ML model may be supplemented by other information, such as browser history, file information, currently running applications and locations, system level and/or application-level information, etc. Generative AI may be employed to more rapidly understand context and semantic associations.

**[0022]** Some embodiments begin training the generative AI/ML model by feeding the initial version of the generative AI/ML model labeled screen images from one or more computing systems as a training input. The generative AI/ML model provides predictions as output, such as which appli-

cation(s) and graphical element(s) are recognized as being present in a screen. Identification errors can be highlighted by a human reviewer (e.g., by drawing a box around the misidentified element and including the correct identification), and the generative AI/ML model can be trained until its accuracy is sufficiently high to be deployed to observe applications and graphical elements that are present in screens of a UI.

**[0023]** Rather than training from images alone, in some embodiments, tracking code may also be embedded in the user's computing system. For instance, a snippet of JavaScript® may be embedded in a web browser as a listener to track what components the user interacts with, what text the user enters, what locations/components the user clicks with the mouse, what content the user scrolled past, how long the user stopped on a certain part of the content, etc. Scrolling past content may indicate that the content may have been somewhat close but did not have exactly what the user wanted. Clicks may indicate success.

**[0024]** Listener applications need not be JavaScript®, and may be any suitable type of application and in any desired programming language without deviating from the scope of the invention. This may allow “generalizing” of the listener application such that it can track user interactions with multiple applications or any application that the user is interacting with. Using labeled training data from scratch may be difficult since while it may allow the generative AI/ML model to learn to recognize various controls, it does not contain information regarding which controls are commonly used and how. Using listener applications, a “heat map” could be generated to help bootstrap the generative AI/ML model training process. The heat map may include various information, such as the frequency that a user used applications, the frequency that the user interacted with components of the applications, the locations of the components, the content of the applications/components, etc. In some embodiments, the heat map may be derived from screen analysis, such as detection of typed and/or pasted text, caret tracking, and active element detection for a computing system. Some embodiments recognize where on the screen associated with a computing system the user has been typing or pasting text, potentially including hot keys or other keys that do not cause visible characters to appear, and provide the physical position on the screen based on the current resolution (e.g., in coordinates) of where one or more characters appeared, where the cursor was blinking, or both. The physical position of the typing or pasting activity and/or of the caret may allow determination of which field(s) the user is typing in or focused on and what the application is for process discovery or other applications.

**[0025]** Some embodiments are implemented in a feedback loop process that continuously or periodically compares the current screenshot to the previous screenshot to identify changes. Locations where visual changes occurred on the screen may be identified and optical character recognition (OCR) may be performed on the location where the change occurred. Results of the OCR may then be compared to the content of a keyboard queue (e.g., as determined by key hooking) to determine whether a match exists. The locations where the change occurred may be determined by comparing a box of pixels from the current screenshot to a box of pixels in the same location from a previous screenshot. When a match is found, the text of the location where the



change occurred may be associated with that location and provided as part of the listener information.

**[0026]** Once the heat map has been generated, the generative AI/ML model could be trained on screen images (potentially millions of images) based on the initial heat map information. Graphics processing units (GPUs) may be able to process this information and train the generative AI/ML model relatively quickly. Once graphical elements, windows, applications, etc. can be accurately identified, the generative AI/ML model may be trained to recognize labeled user interactions with the applications in the UI to understand incremental actions taken by the user. In some embodiments, generative AI models may be used to understand these user interactions (e.g., providing better understanding of context, semantic associations, etc.). One or a series of graphical element changes may be indicative of a user clicking a button, entering text, interacting with a menu, closing a window, moving to a different screen of an application, etc. For instance, a menu item that the user clicks may become underlined, a button may be shaded darker while it is pressed and then return to the original shade when the user releases the mouse button, the letter “a” may appear in a text field, an image may change to a different image, a screen may assume a different layout when the user moves to the next screen of an application of a series of screens, etc.

**[0027]** Identification errors can again be highlighted by a human reviewer (e.g., by drawing a box around the misidentified element and including the correct identification). The generative AI/ML model can then be trained until its accuracy is sufficiently high to be deployed to understand granular user interactions with the UI. Such a trained generative AI/ML model may then be used to observe multiple users and look for common sequences of interactions in common applications, for example.

**[0028]** In some embodiments, training of the generative AI/ML model may be supplemented with information from “automation boxes”, which are implemented via hardware or software and observe what information is coming from an input device, such as a mouse or the keyboard. In certain embodiments, a camera may be used to track where the user is looking on the screen. The information from automation boxes and/or cameras may be time stamped and used in conjunction with the graphical elements, applications, and screens detected by the generative AI/ML model to assist in its training and better understand what the user is doing at the time.

**[0029]** Certain embodiments may be employed for robotic process automation (RPA). FIG. 1 is an architectural diagram illustrating a hyper-automation system 100, according to an embodiment of the present invention. “Hyper-automation,” as used herein, refers to automation systems that bring together components of process automation, integration tools, and technologies that amplify the ability to automate work. For instance, RPA may be used at the core of a hyper-automation system in some embodiments, and in certain embodiments, automation capabilities may be expanded with AI/ML, process mining, analytics, and/or other advanced tools. As the hyper-automation system learns processes, trains AI/ML models, and employs analytics, for example, more and more knowledge work may be automated, and computing systems in an organization, e.g., both those used by individuals and those that run autonomously, may all be engaged to be participants in the hyper-

automation process. Hyper-automation systems of some embodiments allow users and organizations to efficiently and effectively discover, understand, and scale automations.

**[0030]** Hyper-automation system 100 includes user computing systems, such as desktop computer 102, tablet 104, and smart phone 106. However, any desired user computing system may be used without deviating from the scope of the invention including, but not limited to, smart watches, laptop computers, servers, Internet-of-Things (IoT) devices, etc. Also, while three user computing systems are shown in FIG. 1, any suitable number of user computing systems may be used without deviating from the scope of the invention. For instance, in some embodiments, dozens, hundreds, thousands, or millions of user computing systems may be used. The user computing systems may be actively used by a user or run automatically without much or any user input.

**[0031]** Each user computing system 102, 104, 106 has respective automation process(es) 110, 112, 114 running thereon. In some embodiments, the automation processes are stored remotely (e.g., on server 130 or in database 140 and accessed via network 120) and loaded by RPA robots to implement the automation. Automations may exist as a script (e.g., XML, XAML, etc.) or be compiled into machine readable code (e.g., as a digital link library).

**[0032]** Automation process(es) 110, 112, 114 may include, but are not limited to, RPA robots, part of an operating system, downloadable application(s) for the respective computing system, any other suitable software and/or hardware, or any combination of these without deviating from the scope of the invention. In some embodiments, one or more of process(es) 110, 112, 114 may be listeners. Listeners may be RPA robots, part of an operating system, a downloadable application for the respective computing system, or any other software and/or hardware without deviating from the scope of the invention. Indeed, in some embodiments, the logic of the listener(s) is implemented partially or completely via physical hardware.

**[0033]** Listeners monitor and record data pertaining to user interactions with respective computing systems and/or operations of unattended computing systems and send the data to a core hyper-automation system 120 via a network (e.g., a local area network (LAN), a mobile communications network, a satellite communications network, the Internet, any combination thereof, etc.). The data may include, but is not limited to, which buttons were clicked, where a mouse was moved, the text that was entered in a field, that one window was minimized and another was opened, the application associated with a window, etc. In certain embodiments, the data from the listeners may be sent periodically as part of a heartbeat message. In some embodiments, the data may be sent to core hyper-automation system 120 once a predetermined amount of data has been collected, after a predetermined time period has elapsed, or both. One or more servers, such as server 130, receive and store data from the listeners in a database, such as database 140.

**[0034]** Automation processes may execute the logic developed in workflows during design time. In the case of RPA, workflows may include a set of steps, defined herein as “activities,” that are executed in a sequence or some other logical flow. Each activity may include an action, such as clicking a button, reading a file, writing to a log panel, etc. In some embodiments, workflows may be nested or embedded.



**[0035]** Long-running workflows for RPA in some embodiments are master projects that support service orchestration, human intervention, and long-running transactions in unattended environments. Human intervention comes into play when certain processes require human inputs to handle exceptions, approvals, or validation before proceeding to the next step in the activity. In this situation, the process execution is suspended, freeing up the RPA robots until the human task completes.

**[0036]** A long-running workflow may support workflow fragmentation via persistence activities and may be combined with invoke process and non-user interaction activities, orchestrating human tasks with RPA robot tasks. In some embodiments, multiple or many computing systems may participate in executing the logic of a long-running workflow. The long-running workflow may run in a session to facilitate speedy execution. In some embodiments, long-running workflows may orchestrate background processes that may contain activities performing API calls and running in the long-running workflow session. These activities may be invoked by an invoke process activity in some embodiments. A process with user interaction activities that runs in a user session may be called by starting a job from a conductor activity (conductor described in more detail later herein). The user may interact through tasks that require forms to be completed in the conductor in some embodiments. Activities may be included that cause the RPA robot to wait for a form task to be completed and then resume the long-running workflow.

**[0037]** One or more of automation process(es) **110**, **112**, **114** is in communication with core hyper-automation system **120**. In some embodiments, core hyper-automation system **120** may run a conductor application on one or more servers, such as server **130**. While one server **130** is shown for illustration purposes, multiple or many servers that are proximate to one another or in a distributed architecture may be employed without deviating from the scope of the invention. For instance, one or more servers may be provided for conductor functionality, AI/ML model serving, authentication, governance, and or any other suitable functionality without deviating from the scope of the invention. In some embodiments, core hyper-automation system **120** may incorporate or be part of a public cloud architecture, a private cloud architecture, a hybrid cloud architecture, etc. In certain embodiments, core hyper-automation system **120** may host multiple software-based servers on one or more computing systems, such as server **130**. In some embodiments, one or more servers of core hyper-automation system **120**, such as server **130**, may be implemented via one or more virtual machines (VMs).

**[0038]** In some embodiments, one or more of automation process(es) **110**, **112**, **114** may call one or more AI/ML models **132** deployed on or accessible by core hyper-automation system **120** and trained to accomplish various tasks. For instance, AI/ML models **132** may include models trained to look for various application versions, perform CV, generate UI descriptors, offer suggestions for next activities or sequences of activities in RPA workflows, etc. AI/ML models may be trained using labeled data that includes, but is not limited to, screenshots of various application screens for various versions with their corresponding UI elements, libraries of UI objects, etc. AI/ML models **132** may be trained to achieve a desired confidence threshold while not being overfit to a given set of training data.

**[0039]** AI/ML models **132** may be trained for any suitable purpose without deviating from the scope of the invention, as will be discussed in more detail later herein. Two or more of AI/ML models **132** may be chained in some embodiments (e.g., in series, in parallel, or a combination thereof) such that they collectively provide collaborative output(s). AI/ML models **132** may perform or assist with CV, OCR, document processing and/or understanding, semantic learning and/or analysis, analytical predictions, process discovery, task mining, testing, automatic RPA workflow generation, sequence extraction, clustering detection, audio-to-text translation, any combination thereof, etc. However, any desired number and/or type(s) of AI/ML models may be used without deviating from the scope of the invention. Using multiple AI/ML models may allow the system to develop a global picture of what is happening on a given computing system, for example. For instance, one AI/ML model could perform OCR, another could detect buttons, another could compare sequences, etc. Patterns may be determined individually by an AI/ML model or collectively by multiple AI/ML models. In certain embodiments, one or more AI/ML models are deployed locally on at least one of computing systems **102**, **104**, **106**.

**[0040]** In some embodiments, multiple AI/ML models **132** may be used. Each AI/ML model **132** is an algorithm (or model) that runs on the data, and the AI/ML model itself may be a deep learning neural network (DLNN) of trained artificial “neurons” that are trained on training data, for example. In some embodiments, AI/ML models **132** may have multiple layers that perform various functions, such as statistical modeling (e.g., hidden Markov models (HMMs)), and utilize deep learning techniques (e.g., long short term memory (LSTM) deep learning, encoding of previous hidden states, etc.) to perform the desired functionality.

**[0041]** Hyper-automation system **100** may provide four main groups of functionality in some embodiments: (1) discovery; (2) building automations; (3) management; and (4) engagement. Automations (e.g., run on a user computing system, a server, etc.) may be run by software robots, such as RPA robots, in some embodiments. For instance, attended robots, unattended robots, and/or test robots may be used. Attended robots work with users to assist them with tasks (e.g., via UiPath Assistant™). Unattended robots work independently of users and may run in the background, potentially without user knowledge. Test robots are unattended robots that run test cases against applications or RPA workflows. Test robots may be run on multiple computing systems in parallel in some embodiments.

**[0042]** The discovery functionality may discover and provide automatic recommendations for different opportunities of automations of business processes. Such functionality may be implemented by one or more servers, such as server **130**. The discovery functionality may include providing an automation hub, process mining, task mining, and/or task capture in some embodiments. The automation hub (e.g., UiPath Automation Hub™) may provide a mechanism for managing automation rollout with visibility and control. Automation ideas may be crowdsourced from employees via a submission form, for example. Feasibility and return on investment (ROI) calculations for automating these ideas may be provided, documentation for future automations may be collected, and collaboration may be provided to get from automation discovery to build-out faster.



**[0043]** Process mining (e.g., via UiPath Automation Cloud™ and/or UiPath AI Center™) refers to the process of gathering and analyzing the data from applications (e.g., enterprise resource planning (ERP) applications, customer relation management (CRM) applications, email applications, call center applications, etc.) to identify what end-to-end processes exist in an organization and how to automate them effectively, as well as indicate what the impact of the automation will be. This data may be gleaned from user computing systems **102**, **104**, **106** by listeners, for example, and processed by servers, such as server **130**. One or more AI/ML models **132** may be employed for this purpose in some embodiments. This information may be exported to the automation hub to speed up implementation and avoid manual information transfer. The goal of process mining may be to increase business value by automating processes within an organization. Some examples of process mining goals include, but are not limited to, increasing profit, improving customer satisfaction, regulatory and/or contractual compliance, improving employee efficiency, etc.

**[0044]** Task mining (e.g., via UiPath Automation Cloud™ and/or UiPath AI Center™) identifies and aggregates workflows (e.g., employee workflows), and then applies AI to expose patterns and variations in day-to-day tasks, scoring such tasks for ease of automation and potential savings (e.g., time and/or cost savings). One or more AI/ML models **132** may be employed to uncover recurring task patterns in the data. Repetitive tasks that are ripe for automation may then be identified. This information may initially be provided by listeners and analyzed on servers of core hyper-automation system **120**, such as server **130**, in some embodiments. The findings from task mining (e.g., Extensible Application Markup Language (XAML) process data) may be exported to process documents or to a designer application such as UiPath Studio™ to create and deploy automations more rapidly.

**[0045]** Task mining in some embodiments may include taking screenshots with user actions (e.g., mouse click locations, keyboard inputs, application windows and graphical elements the user was interacting with, timestamps for the interactions, etc.), collecting statistical data (e.g., execution time, number of actions, text entries, etc.), editing and annotating screenshots, specifying types of actions to be recorded, etc.

**[0046]** Task capture (e.g., via UiPath Automation Cloud™ and/or UiPath AI Center™) automatically documents attended processes as users work or provides a framework for unattended processes. Such documentation may include desired tasks to automate in the form of process definition documents (PDDs), skeletal workflows, capturing actions for each part of a process, recording user actions and automatically generating a comprehensive workflow diagram including the details about each step, Microsoft Word® documents, XAML files, and the like. Build-ready workflows may be exported directly to a designer application in some embodiments, such as UiPath Studio™. Task capture may simplify the requirements gathering process for both subject matter experts explaining a process and CoE members providing production-grade automations.

**[0047]** Building automations may be accomplished via a designer application (e.g., UiPath Studio™, UiPath StudioX™, or UiPath Web™). For instance, RPA developers of an RPA development facility **150** may use RPA designer applications **154** of computing systems **152** to build and test

automations for various applications and environments, such as web, mobile, SAP®, and virtualized desktops. API integration may be provided for various applications, technologies, and platforms. Predefined activities, drag-and-drop modeling, and a workflow recorder, may make automation easier with minimal coding. Document understanding functionality may be provided via drag-and-drop AI skills for data extraction and interpretation that call one or more AI/ML models **132**. Such automations may process virtually any document type and format, including tables, checkboxes, signatures, and handwriting. When data is validated or exceptions are handled, this information may be used to retrain the respective AI/ML models, improving their accuracy over time.

**[0048]** RPA designer application **152** may be designed to call one or more of trained AI/ML models **132** on server **130** and/or generative AI models **172** in a cloud environment via network **120** (e.g., a local area network (LAN), a mobile communications network, a satellite communications network, the Internet, any combination thereof, etc.) to assist with the RPA automation development process. In some embodiments, one or more of the AI/ML models may be packaged with RPA designer application **152** or otherwise stored locally on computing system **150**.

**[0049]** In some embodiments, RPA designer application **152** and one or more of AI/ML models **132** may be configured to use an object repository stored in database **140**. The object repository may include libraries of UI objects that can be used to develop RPA workflows via RPA designer application **152**. The object repository may be used to add UI descriptors to activities in the workflows of RPA designer application **152** for UI automations. In some embodiments, one or more of AI/ML models **132** may generate new UI descriptors and add them to the object repository in database **140**. Once automations are completed in designer application **152**, they may be published on server **130**, pushed out to computing systems **102**, **104**, **106**, etc.

**[0050]** An integration service may allow developers to seamlessly combine user interface (UI) automation with API automation, for example. Automations may be built that require APIs or traverse both API and non-API applications and systems. A repository (e.g., UiPath Object Repository™) or marketplace (e.g., UiPath Marketplace™) for pre-built RPA and AI templates and solutions may be provided to allow developers to automate a wide variety of processes more quickly. Thus, when building automations, hyper-automation system **100** may provide user interfaces, development environments, API integration, pre-built and/or custom-built AI/ML models, development templates, integrated development environments (IDEs), and advanced AI capabilities. Hyper-automation system **100** enables development, deployment, management, configuration, monitoring, debugging, and maintenance of RPA robots in some embodiments, which may provide automations for hyper-automation system **100**.

**[0051]** In some embodiments, components of hyper-automation system **100**, such as designer application(s) and/or an external rules engine, provide support for managing and enforcing governance policies for controlling various functionality provided by hyper-automation system **100**. Governance is the ability for organizations to put policies in place to prevent users from developing automations (e.g., RPA robots) capable of taking actions that may harm the organization, such as violating the E.U. General Data Pro-



tection Regulation (GDPR), the U.S. Health Insurance Portability and Accountability Act (HIPAA), third party application terms of service, etc. Since developers may otherwise create automations that violate privacy laws, terms of service, etc. while performing their automations, some embodiments implement access control and governance restrictions at the robot and/or robot design application level. This may provide an added level of security and compliance into to the automation process development pipeline in some embodiments by preventing developers from taking dependencies on unapproved software libraries that may either introduce security risks or work in a way that violates policies, regulations, privacy laws, and/or privacy policies.

**[0052]** The management functionality may provide management, deployment, and optimization of automations across an organization. The management functionality may include orchestration, test management, AI functionality, and/or insights in some embodiments. Management functionality of hyper-automation system **100** may also act as an integration point with third-party solutions and applications for automation applications and/or RPA robots. The management capabilities of hyper-automation system **100** may include, but are not limited to, facilitating provisioning, deployment, configuration, queuing, monitoring, logging, and interconnectivity of RPA robots, among other things.

**[0053]** A conductor application, such as UiPath Orchestrator™ (which may be provided as part of the UiPath Automation Cloud™ in some embodiments, or on premises, in VMs, in a private or public cloud, in a Linux™ VM, or as a cloud native single container suite via UiPath Automation Suite™), provides orchestration capabilities to deploy, monitor, optimize, scale, and ensure security of RPA robot deployments. A test suite (e.g., UiPath Test Suite™) may provide test management to monitor the quality of deployed automations. The test suite may facilitate test planning and execution, meeting of requirements, and defect traceability. The test suite may include comprehensive test reporting.

**[0054]** Analytics software (e.g., UiPath Insights™) may track, measure, and manage the performance of deployed automations. The analytics software may align automation operations with specific key performance indicators (KPIs) and strategic outcomes for an organization. The analytics software may present results in a dashboard format for better understanding by human users.

**[0055]** A data service (e.g., UiPath Data Service™) may be stored in database **140**, for example, and bring data into a single, scalable, secure place with a drag-and-drop storage interface. Some embodiments may provide low-code or no-code data modeling and storage to automations while ensuring seamless access, enterprise-grade security, and scalability of the data. AI functionality may be provided by an AI center (e.g., UiPath AI Center™), which facilitates incorporation of AI/ML models into automations. Pre-built AI/ML models, model templates, and various deployment options may make such functionality accessible even to those who are not data scientists. Deployed automations (e.g., RPA robots) may call AI/ML models from the AI center, such as AI/ML models **132**. Performance of the AI/ML models may be monitored, and be trained and improved using human-validated data, such as that provided by data review center **160**. Human reviewers may provide labeled data to core hyper-automation system **120** via a review application **152** on computing systems **154**. For instance, human reviewers may validate that predictions by AI/ML

models **132** are accurate or provide corrections otherwise. This dynamic input may then be saved as training data for retraining AI/ML models **132**, and may be stored in a database such as database **140**, for example. The AI center may then schedule and execute training jobs to train the new versions of the AI/ML models using the training data. Both positive and negative examples may be stored and used for retraining of AI/ML models **132**.

**[0056]** The engagement functionality engages humans and automations as one team for seamless collaboration on desired processes. Low-code applications may be built (e.g., via UiPath Apps™) to connect browser tabs and legacy software, even that lacking APIs in some embodiments. Applications may be created quickly using a web browser through a rich library of drag-and-drop controls, for instance. An application can be connected to a single automation or multiple automations.

**[0057]** An action center (e.g., UiPath Action Center™) provides a straightforward and efficient mechanism to hand off processes from automations to humans, and vice versa. Humans may provide approvals or escalations, make exceptions, etc. The automation may then perform the automatic functionality of a given workflow.

**[0058]** A local assistant may be provided as a launchpad for users to launch automations (e.g., UiPath Assistant™). This functionality may be provided in a tray provided by an operating system, for example, and may allow users to interact with RPA robots and RPA robot-powered applications on their computing systems. An interface may list automations approved for a given user and allow the user to run them. These may include ready-to-go automations from an automation marketplace, an internal automation store in an automation hub, etc. When automations run, they may run as a local instance in parallel with other processes on the computing system so users can use the computing system while the automation performs its actions. In certain embodiments, the assistant is integrated with the task capture functionality such that users can document their soon-to-be-automated processes from the assistant launchpad.

**[0059]** Chatbots (e.g., UiPath Chatbots™), social messaging applications, an/or voice commands may enable users to run automations. This may simplify access to information, tools, and resources users need in order to interact with customers or perform other activities. Conversations between people may be readily automated, as with other processes. Trigger RPA robots kicked off in this manner may perform operations such as checking an order status, posting data in a CRM, etc., potentially using plain language commands.

**[0060]** End-to-end measurement and government of an automation program at any scale may be provided by hyper-automation system **100** in some embodiments. Per the above, analytics may be employed to understand the performance of automations (e.g., via UiPath Insights™). Data modeling and analytics using any combination of available business metrics and operational insights may be used for various automated processes. Custom-designed and pre-built dashboards allow data to be visualized across desired metrics, new analytical insights to be discovered, performance indicators to be tracked, ROI to be discovered for automations, telemetry monitoring to be performed on user computing systems, errors and anomalies to be detected, and automations to be debugged. An automation management console (e.g., UiPath Automation Ops™) may be provided



to manage automations throughout the automation lifecycle. An organization may govern how automations are built, what users can do with them, and which automations users can access.

**[0061]** Hyper-automation system **100** provides an iterative platform in some embodiments. Processes can be discovered, automations can be built, tested, and deployed, performance may be measured, use of the automations may readily be provided to users, feedback may be obtained, AI/ML models may be trained and retrained, and the process may repeat itself. This facilitates a more robust and effective suite of automations.

**[0062]** In some embodiments, generative AI models are used. Generative AI can generate various types of content, such as text, imagery, audio, and synthetic data. various types of generative AI models may be used, including, but not limited to, large language models (LLMs), generative adversarial networks (GANs), variational autoencoders (VAEs), transformers, etc. These models may be part of AI/ML models **132** hosted on server **130**. For instance, the generative AI models may be trained on a large corpus of textual information to perform semantic understanding, to understand the nature of what is present on a screen from text, and the like. In certain embodiments, generative AI models **172** provided by an existing cloud ML service provider, such as OpenAI®, Google®, Amazon®, Microsoft®, IBM®, Nvidia®, Facebook®, etc., may be employed and trained to provide such functionality. In generative AI embodiments where generative AI model(s) **172** are remotely hosted, server **130** can be configured to integrate with third-party APIs, which allow server **130** to send a request to generative AI model(s) **172** including the requisite input information and receive a response in return (e.g., the semantic matches of fields between application versions and/or screens, a classification of the type of the application on the screen, etc.). Such embodiments may provide a more advanced and sophisticated user experience, as well as provide access to state-of-the-art natural language processing (NLP) and other ML capabilities that these companies offer.

**[0063]** FIG. 2 is an architectural diagram illustrating an RPA system **200**, according to an embodiment of the present invention. In some embodiments, RPA system **200** is part of hyper-automation system **100** of FIG. 1. RPA system **200** includes a designer **210** that allows a developer to design and implement workflows. Designer **210** may provide a solution for application integration, as well as automating third-party applications, administrative Information Technology (IT) tasks, and business IT processes. Designer **210** may facilitate development of an automation project, which is a graphical representation of a business process. Simply put, designer **210** facilitates the development and deployment of workflows and robots. In some embodiments, designer **210** may be an application that runs on a user's desktop, an application that runs remotely in a VM, a web application, etc.

**[0064]** The automation project enables automation of rule-based processes by giving the developer control of the execution order and the relationship between a custom set of steps developed in a workflow, defined herein as "activities." One commercial example of an embodiment of designer **210** is UiPath Studio™. Each activity may include an action, such as clicking a button, reading a file, writing to a log panel, etc. In some embodiments, workflows may be nested or embedded.

**[0065]** Some types of workflows may include, but are not limited to, sequences, flowcharts, FSMs, and/or global exception handlers. Sequences may be particularly suitable for linear processes, enabling flow from one activity to another without cluttering a workflow. Flowcharts may be particularly suitable to more complex business logic, enabling integration of decisions and connection of activities in a more diverse manner through multiple branching logic operators. FSMs may be particularly suitable for large workflows. FSMs may use a finite number of states in their execution, which are triggered by a condition (i.e., transition) or an activity. Global exception handlers may be particularly suitable for determining workflow behavior when encountering an execution error and for debugging processes.

**[0066]** Once a workflow is developed in designer **210**, execution of business processes is orchestrated by conductor **220**, which orchestrates one or more robots **230** that execute the workflows developed in designer **210**. One commercial example of an embodiment of conductor **220** is UiPath Orchestrator™. Conductor **220** facilitates management of the creation, monitoring, and deployment of resources in an environment. Conductor **220** may act as an integration point with third-party solutions and applications. Per the above, in some embodiments, conductor **220** may be part of core hyper-automation system **120** of FIG. 1.

**[0067]** Conductor **220** may manage a fleet of robots **230**, connecting and executing robots **230** from a centralized point. Types of robots **230** that may be managed include, but are not limited to, attended robots **232**, unattended robots **234**, development robots (similar to unattended robots **234**, but used for development and testing purposes), and nonproduction robots (similar to attended robots **232**, but used for development and testing purposes). Attended robots **232** are triggered by user events and operate alongside a human on the same computing system. Attended robots **232** may be used with conductor **220** for a centralized process deployment and logging medium. Attended robots **232** may help the human user accomplish various tasks, and may be triggered by user events. In some embodiments, processes cannot be started from conductor **220** on this type of robot and/or they cannot run under a locked screen. In certain embodiments, attended robots **232** can only be started from a robot tray or from a command prompt. Attended robots **232** should run under human supervision in some embodiments.

**[0068]** Unattended robots **234** run unattended in virtual environments and can automate many processes. Unattended robots **234** may be responsible for remote execution, monitoring, scheduling, and providing support for work queues. Debugging for all robot types may be run in designer **210** in some embodiments. Both attended and unattended robots may automate various systems and applications including, but not limited to, mainframes, web applications, VMs, enterprise applications (e.g., those produced by SAP®, Salesforce®, Oracle®, etc.), and computing system applications (e.g., desktop and laptop applications, mobile device applications, wearable computer applications, etc.).

**[0069]** Conductor **220** may have various capabilities including, but not limited to, provisioning, deployment, versioning, configuration, queueing, monitoring, logging, and/or providing interconnectivity. Provisioning may include creating and maintenance of connections between robots **230** and conductor **220** (e.g., a web application). Deploy-



ment may include assuring the correct delivery of package versions to assigned robots **230** for execution. Configuration may include maintenance and delivery of robot environments and process configurations. Queueing may include providing management of queues and queue items. Monitoring may include keeping track of robot identification data and maintaining user permissions. Logging may include storing and indexing logs to a database (e.g., a structured query language (SQL) database or a “not only” SQL (NoSQL) database) and/or another storage mechanism (e.g., Elasticsearch®, which provides the ability to store and quickly query large datasets). Conductor **220** may provide interconnectivity by acting as the centralized point of communication for third-party solutions and/or applications.

[0070] Robots **230** are execution agents that run workflows built in designer **210**. One commercial example of some embodiments of robot(s) **230** is UiPath Robots™. In some embodiments, robots **230** install the Microsoft Windows® Service Control Manager (SCM)-managed service by default. As a result, such robots **230** can open interactive Windows® sessions under the local system account, and have the rights of a Windows® service.

[0071] In some embodiments, robots **230** can be installed in a user mode. For such robots **230**, this means they have the same rights as the user under which a given robot **230** has been installed. This feature may also be available for High Density (HD) robots, which ensure full utilization of each machine at its maximum potential. In some embodiments, any type of robot **230** may be configured in an HD environment.

[0072] Robots **230** in some embodiments are split into several components, each being dedicated to a particular automation task. The robot components in some embodiments include, but are not limited to, SCM-managed robot services, user mode robot services, executors, agents, and command line. SCM-managed robot services manage and monitor Windows® sessions and act as a proxy between conductor **220** and the execution hosts (i.e., the computing systems on which robots **230** are executed). These services are trusted with and manage the credentials for robots **230**. A console application is launched by the SCM under the local system.

[0073] User mode robot services in some embodiments manage and monitor Windows® sessions and act as a proxy between conductor **220** and the execution hosts. User mode robot services may be trusted with and manage the credentials for robots **230**. A Windows® application may automatically be launched if the SCM-managed robot service is not installed.

[0074] Executors may run given jobs under a Windows® session (i.e., they may execute workflows. Executors may be aware of per-monitor dots per inch (DPI) settings. Agents may be Windows® Presentation Foundation (WPF) applications that display the available jobs in the system tray window. Agents may be a client of the service. Agents may request to start or stop jobs and change settings. The command line is a client of the service. The command line is a console application that can request to start jobs and waits for their output.

[0075] Having components of robots **230** split as explained above helps developers, support users, and computing systems more easily run, identify, and track what each component is executing. Special behaviors may be configured per component this way, such as setting up dif-

ferent firewall rules for the executor and the service. The executor may always be aware of DPI settings per monitor in some embodiments. As a result, workflows may be executed at any DPI, regardless of the configuration of the computing system on which they were created. Projects from designer **210** may also be independent of browser zoom level in some embodiments. For applications that are DPI-unaware or intentionally marked as unaware, DPI may be disabled in some embodiments.

[0076] RPA system **200** in this embodiment is part of a hyper-automation system. Developers may use designer **210** to build and test RPA robots that utilize AI/ML models deployed in core hyper-automation system **240** (e.g., as part of an AI center thereof). Such RPA robots may send input for execution of the AI/ML model(s) and receive output therefrom via core hyper-automation system **240**.

[0077] One or more of robots **230** may be listeners, as described above. These listeners may provide information to core hyper-automation system **240** regarding what users are doing when they use their computing systems. This information may then be used by core hyper-automation system for process mining, task mining, task capture, etc.

[0078] An assistant / chatbot **250** may be provided on user computing systems to allow users to launch RPA local robots. The assistant may be located in a system tray, for example. Chatbots may have a user interface so users can see text in the chatbot. Alternatively, chatbots may lack a user interface and run in the background, listening using the computing system's microphone for user speech.

[0079] In some embodiments, data labeling may be performed by a user of the computing system on which a robot is executing or on another computing system that the robot provides information to. For instance, if a robot calls an AI/ML model that performs CV on images for VM users, but the AI/ML model does not correctly identify a button on the screen, the user may draw a rectangle around the misidentified or non-identified component and potentially provide text with a correct identification. This information may be provided to core hyper-automation system **240** and then used later for training a new version of the AI/ML model.

[0080] FIG. 3 is an architectural diagram illustrating a deployed RPA system **300**, according to an embodiment of the present invention. In some embodiments, RPA system **300** may be, or may be a part of, RPA system **200** of FIG. 2 and/or hyper-automation system **100** of FIG. 1. Deployed RPA system **300** may be a cloud-based system, an on-premises system, a desktop-based system that offers enterprise level, user level, or device level automation solutions for automation of different computing processes, etc.

[0081] It should be noted that the client side, the server side, or both, may include any desired number of computing systems without deviating from the scope of the invention. On the client side, a robot application **310** includes executors **312**, an agent **314**, and a designer **316**. However, in some embodiments, designer **316** may not be running on the same computing system **310** as executors **312** and agent **314**. Executors **312** are running processes. Several business projects may run simultaneously, as shown in FIG. 3. Agent **314** (e.g., a Windows® service) is the single point of contact for all executors **312** in this embodiment. All messages in this embodiment are logged into conductor **340**, which processes them further via a database server **350**, an AI/ML server **360**, an indexer server **370**, or any combi-



nation thereof. As discussed above with respect to FIG. 2, executors **312** may be robot components.

[0082] In some embodiments, a robot represents an association between a machine name and a username. The robot may manage multiple executors at the same time. On computing systems that support multiple interactive sessions running simultaneously (e.g., Windows® Server **2012**), multiple robots may be running at the same time, each in a separate Windows® session using a unique username. This is referred to as HD robots above.

[0083] Agent **314** is also responsible for sending the status of the robot (e.g., periodically sending a “heartbeat” message indicating that the robot is still functioning) and downloading the required version of the package to be executed. The communication between agent **314** and conductor **340** is always initiated by agent **314** in some embodiments. In the notification scenario, agent **314** may open a WebSocket channel that is later used by conductor **340** to send commands to the robot (e.g., start, stop, etc.).

[0084] A listener **330** monitors and records data pertaining to user interactions with an attended computing system and/or operations of an unattended computing system on which listener **330** resides. Listener **330** may be an RPA robot, part of an operating system, a downloadable application for the respective computing system, or any other software and/or hardware without deviating from the scope of the invention. Indeed, in some embodiments, the logic of the listener is implemented partially or completely via physical hardware.

[0085] On the server side, a presentation layer (web application **342**, Open Data Protocol (ODa) Representative State Transfer (REST) API endpoints **344**, and notification and monitoring **346**), a service layer (API implementation / business logic **348**), and a persistence layer (database server **350**, AI/ML server **360**, and indexer server **370**) are included. Conductor **340** includes web application **342**, ODa REST API endpoints **344**, notification and monitoring **346**, and API implementation / business logic **348**. In some embodiments, most actions that a user performs in the interface of conductor **340** (e.g., via browser **320**) are performed by calling various APIs. Such actions may include, but are not limited to, starting jobs on robots, adding/removing data in queues, scheduling jobs to run unattended, etc. without deviating from the scope of the invention. Web application **342** is the visual layer of the server platform. In this embodiment, web application **342** uses Hypertext Markup Language (HTML) and JavaScript (JS). However, any desired markup languages, script languages, or any other formats may be used without deviating from the scope of the invention. The user interacts with web pages from web application **342** via browser **320** in this embodiment in order to perform various actions to control conductor **340**. For instance, the user may create robot groups, assign packages to the robots, analyze logs per robot and/or per process, start and stop robots, etc.

[0086] In addition to web application **342**, conductor **340** also includes service layer that exposes ODa REST API endpoints **344**. However, other endpoints may be included without deviating from the scope of the invention. The REST API is consumed by both web application **342** and agent **314**. Agent **314** is the supervisor of one or more robots on the client computer in this embodiment.

[0087] The REST API in this embodiment covers configuration, logging, monitoring, and queueing functionality. The configuration endpoints may be used to define and con-

figure application users, permissions, robots, assets, releases, and environments in some embodiments. Logging REST endpoints may be used to log different information, such as errors, explicit messages sent by the robots, and other environment-specific information, for instance. Deployment REST endpoints may be used by the robots to query the package version that should be executed if the start job command is used in conductor **340**. Queueing REST endpoints may be responsible for queues and queue item management, such as adding data to a queue, obtaining a transaction from the queue, setting the status of a transaction, etc.

[0088] Monitoring REST endpoints may monitor web application **342** and agent **314**. Notification and monitoring API **346** may be REST endpoints that are used for registering agent **314**, delivering configuration settings to agent **314**, and for sending/receiving notifications from the server and agent **314**. Notification and monitoring API **346** may also use WebSocket communication in some embodiments.

[0089] The APIs in the service layer may be accessed through configuration of an appropriate API access path in some embodiments, e.g., based on whether conductor **340** and an overall hyper-automation system have an on-premises deployment type or a cloud-based deployment type. APIs for conductor **340** may provide custom methods for querying stats about various entities registered in conductor **340**. Each logical resource may be an OData entity in some embodiments. In such an entity, components such as the robot, process, queue, etc., may have properties, relationships, and operations. APIs of conductor **340** may be consumed by web application **342** and/or agents **314** in two ways in some embodiments: by getting the API access information from conductor **340**, or by registering an external application to use the OAuth flow.

[0090] The persistence layer includes a trio of servers in this embodiment - database server **350** (e.g., a SQL server), AI/ML server **360** (e.g., a server providing AI/ML model serving services, such as AI center functionality) and indexer server **370**. Database server **350** in this embodiment stores the configurations of the robots, robot groups, associated processes, users, roles, schedules, etc. This information is managed through web application **342** in some embodiments. Database server **350** may manage queues and queue items. In some embodiments, database server **350** may store messages logged by the robots (in addition to or in lieu of indexer server **370**). Database server **350** may also store process mining, task mining, and/or task capture-related data, received from listener **330** installed on the client side, for example. While no arrow is shown between listener **330** and database **350**, it should be understood that listener **330** is able to communicate with database **350**, and vice versa in some embodiments. This data may be stored in the form of PDDs, images, XAML files, etc. Listener **330** may be configured to intercept user actions, processes, tasks, and performance metrics on the respective computing system on which listener **330** resides. For example, listener **330** may record user actions (e.g., clicks, typed characters, locations, applications, active elements, times, etc.) on its respective computing system and then convert these into a suitable format to be provided to and stored in database server **350**.

[0091] AI/ML server **360** facilitates incorporation of AI/ML models into automations. Pre-built AI/ML models, model templates, and various deployment options may



make such functionality accessible even to those who are not data scientists. Deployed automations (e.g., RPA robots) may call AI/ML models from AI/ML server **360**. Performance of the AI/ML models may be monitored, and be trained and improved using human-validated data. AI/ML server **360** may schedule and execute training jobs to train new versions of the AI/ML models.

[0092] AI/ML server **360** may store data pertaining to AI/ML models and ML packages for configuring various ML skills for a user at development time. An ML skill, as used herein, is a pre-built and trained ML model for a process, which may be used by an automation, for example. AI/ML server **360** may also store data pertaining to document understanding technologies and frameworks, algorithms and software packages for various AI/ML capabilities including, but not limited to, intent analysis, natural language processing (NLP), speech analysis, different types of AI/ML models, etc.

[0093] Indexer server **370**, which is optional in some embodiments, stores and indexes the information logged by the robots. In certain embodiments, indexer server **370** may be disabled through configuration settings. In some embodiments, indexer server **370** uses ElasticSearch®, which is an open source project full-text search engine. Messages logged by robots (e.g., using activities like log message or write line) may be sent through the logging REST endpoint(s) to indexer server **370**, where they are indexed for future utilization.

[0094] FIG. 4 is an architectural diagram illustrating the relationship **400** between a designer **410**, activities **420**, **430**, **440**, **450**, drivers **460**, APIs **470**, and AI/ML models **480**, according to an embodiment of the present invention. Per the above, a developer uses designer **410** to develop workflows that are executed by robots. The various types of activities may be displayed to the developer in some embodiments. Designer **410** may be local to the user's computing system or remote thereto (e.g., accessed via VM or a local web browser interacting with a remote web server). Workflows may include user-defined activities **420**, API-driven activities **430**, AI/ML activities **440**, and/or and UI automation activities **450**. User-defined activities **420** and API-driven activities **440** interact with applications via their APIs. User-defined activities **420** and/or AI/ML activities **440** may call one or more AI/ML models **480** in some embodiments, which may be located locally to the computing system on which the robot is operating and/or remotely thereto.

[0095] Some embodiments are able to identify non-textual visual components in an image, which is called CV herein. CV may be performed at least in part by AI/ML model(s) **480**. Some CV activities pertaining to such components may include, but are not limited to, extracting of text from segmented label data using OCR, fuzzy text matching, cropping of segmented label data using ML, comparison of extracted text in label data with ground truth data, etc. In some embodiments, there may be hundreds or even thousands of activities that may be implemented in user-defined activities **420**. However, any number and/or type of activities may be used without deviating from the scope of the invention.

[0096] UI automation activities **450** are a subset of special, lower-level activities that are written in lower-level code and facilitate interactions with the screen. UI automation activities **450** facilitate these interactions via drivers **460** that allow the robot to interact with the desired software.

For instance, drivers **460** may include operating system (OS) drivers **462**, browser drivers **464**, VM drivers **466**, enterprise application drivers **468**, etc. One or more of AI/ML models **480** may be used by UI automation activities **450** in order to perform interactions with the computing system in some embodiments. In certain embodiments, AI/ML models **480** may augment drivers **460** or replace them completely. Indeed, in certain embodiments, drivers **460** are not included.

[0097] Drivers **460** may interact with the OS at a low level looking for hooks, monitoring for keys, etc. via OS drivers **462**. Drivers **460** may facilitate integration with Chrome®, IE®, Citrix®, SAP®, etc. For instance, the “click” activity performs the same role in these different applications via drivers **460**.

[0098] FIG. 5 is an architectural diagram illustrating a computing system **500** configured to train a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements using generative AI, according to an embodiment of the present invention. In some embodiments, computing system **500** may be one or more of the computing systems depicted and/or described herein. Computing system **500** includes a bus **505** or other communication mechanism for communicating information, and processor(s) **510** coupled to bus **505** for processing information. Processor(s) **510** may be any type of general or specific purpose processor, including a Central Processing Unit (CPU), an Application Specific Integrated Circuit (ASIC), a Field Programmable Gate Array (FPGA), a Graphics Processing Unit (GPU), multiple instances thereof, and/or any combination thereof. Processor(s) **510** may also have multiple processing cores, and at least some of the cores may be configured to perform specific functions. Multi-parallel processing may be used in some embodiments. In certain embodiments, at least one of processor(s) **510** may be a neuromorphic circuit that includes processing elements that mimic biological neurons. In some embodiments, neuromorphic circuits may not require the typical components of a Von Neumann computing architecture.

[0099] Computing system **500** further includes a memory **515** for storing information and instructions to be executed by processor(s) **510**. Memory **515** can be comprised of any combination of Random Access Memory (RAM), Read Only Memory (ROM), flash memory, cache, static storage such as a magnetic or optical disk, or any other types of non-transitory computer-readable media or combinations thereof. Non-transitory computer-readable media may be any available media that can be accessed by processor(s) **510** and may include volatile media, non-volatile media, or both. The media may also be removable, non-removable, or both.

[0100] Additionally, computing system **500** includes a communication device **520**, such as a transceiver, to provide access to a communications network via a wireless and/or wired connection. In some embodiments, communication device **520** may be configured to use Frequency Division Multiple Access (FDMA), Single Carrier FDMA (SC-FDMA), Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), Orthogonal Frequency Division Multiplexing (OFDM), Orthogonal Frequency Division Multiple Access (OFDMA), Global System for Mobile (GSM) communications, General Packet Radio Service (GPRS), Universal Mobile Telecommunications Sys-



tem (UMTS), cdma2000, Wideband CDMA (W-CDMA), High-Speed Downlink Packet Access (HSDPA), High-Speed Uplink Packet Access (HSUPA), High-Speed Packet Access (HSPA), Long Term Evolution (LTE), LTE Advanced (LTE-A), 802.11x, Wi-Fi, Zigbee, Ultra-Wide-Band (UWB), 802.16x, 802.15, Home Node-B (HnB), Bluetooth, Radio Frequency Identification (RFID), Infrared Data Association (IrDA), Near-Field Communications (NFC), fifth generation (5G), New Radio (NR), any combination thereof, and/or any other currently existing or future-implemented communications standard and/or protocol without deviating from the scope of the invention. In some embodiments, communication device **520** may include one or more antennas that are singular, arrayed, phased, switched, beam-forming, beamsteering, a combination thereof, and or any other antenna configuration without deviating from the scope of the invention.

**[0101]** Processor(s) **510** are further coupled via bus **505** to a display **525**, such as a plasma display, a Liquid Crystal Display (LCD), a Light Emitting Diode (LED) display, a Field Emission Display (FED), an Organic Light Emitting Diode (OLED) display, a flexible OLED display, a flexible substrate display, a projection display, a 4 K display, a high definition display, a Retina® display, an In-Plane Switching (IPS) display, or any other suitable display for displaying information to a user. Display **525** may be configured as a touch (haptic) display, a three dimensional (3D) touch display, a multi-input touch display, a multi-touch display, etc. using resistive, capacitive, surface-acoustic wave (SAW) capacitive, infrared, optical imaging, dispersive signal technology, acoustic pulse recognition, frustrated total internal reflection, etc. Any suitable display device and haptic I/O may be used without deviating from the scope of the invention.

**[0102]** A keyboard **530** and a cursor control device **535**, such as a computer mouse, a touchpad, etc., are further coupled to bus **505** to enable a user to interface with computing system **500**. However, in certain embodiments, a physical keyboard and mouse may not be present, and the user may interact with the device solely through display **525** and/or a touchpad (not shown). Any type and combination of input devices may be used as a matter of design choice. In certain embodiments, no physical input device and/or display is present. For instance, the user may interact with computing system **500** remotely via another computing system in communication therewith, or computing system **500** may operate autonomously.

**[0103]** Memory **515** stores software modules that provide functionality when executed by processor(s) **510**. The modules include an operating system **540** for computing system **500**. The modules further include an AI/ML model training module **545** that is configured to perform all or part of the processes described herein or derivatives thereof. Computing system **500** may include one or more additional functional modules **550** that include additional functionality.

**[0104]** One skilled in the art will appreciate that a “system” could be embodied as a server, an embedded computing system, a personal computer, a console, a personal digital assistant (PDA), a cell phone, a tablet computing device, a quantum computing system, or any other suitable computing device, or combination of devices without deviating from the scope of the invention. Presenting the above-described functions as being performed by a “system” is not intended to limit the scope of the present invention in

any way, but is intended to provide one example of the many embodiments of the present invention. Indeed, methods, systems, and apparatuses disclosed herein may be implemented in localized and distributed forms consistent with computing technology, including cloud computing systems. The computing system could be part of or otherwise accessible by a local area network (LAN), a mobile communications network, a satellite communications network, the Internet, a public or private cloud, a hybrid cloud, a server farm, any combination thereof, etc. Any localized or distributed architecture may be used without deviating from the scope of the invention.

**[0105]** It should be noted that some of the system features described in this specification have been presented as modules, in order to more particularly emphasize their implementation independence. For example, a module may be implemented as a hardware circuit comprising custom very large scale integration (VLSI) circuits or gate arrays, off-the-shelf semiconductors such as logic chips, transistors, or other discrete components. A module may also be implemented in programmable hardware devices such as field programmable gate arrays, programmable array logic, programmable logic devices, graphics processing units, or the like.

**[0106]** A module may also be at least partially implemented in software for execution by various types of processors. An identified unit of executable code may, for instance, include one or more physical or logical blocks of computer instructions that may, for instance, be organized as an object, procedure, or function. Nevertheless, the executables of an identified module need not be physically located together, but may include disparate instructions stored in different locations that, when joined logically together, comprise the module and achieve the stated purpose for the module. Further, modules may be stored on a computer-readable medium, which may be, for instance, a hard disk drive, flash device, RAM, tape, and/or any other such non-transitory computer-readable medium used to store data without deviating from the scope of the invention.

**[0107]** Indeed, a module of executable code could be a single instruction, or many instructions, and may even be distributed over several different code segments, among different programs, and across several memory devices. Similarly, operational data may be identified and illustrated herein within modules, and may be embodied in any suitable form and organized within any suitable type of data structure. The operational data may be collected as a single data set, or may be distributed over different locations including over different storage devices, and may exist, at least partially, merely as electronic signals on a system or network.

**[0108]** Various types of AI/ML models may be trained and deployed without deviating from the scope of the invention. For instance, FIG. 6A illustrates an example of a neural network **600** that has been trained to recognize user interactions with applications, screens, and UI elements, according to an embodiment of the present invention. Neural network **600** includes a number of hidden layers. Both deep learning neural networks (DLNNs) and shallow learning neural networks (SLNNs) usually have multiple layers, although SLNNs may only have one or two layers in some cases, and normally fewer than DLNNs. Typically, the neural network architecture includes an input layer, multiple intermediate layers, and an output layer, as is the case in neural network **600**.



[0109] A DLNN often has many layers (e.g., 10, 50, 200, etc.) and subsequent layers typically reuse features from previous layers to compute more complex, general functions. A SLNN, on the other hand, tends to have only a few layers and train relatively quickly since expert features are created from raw data samples in advance. However, feature extraction is laborious. DLNNs, on the other hand, usually do not require expert features, but tend to take longer to train and have more layers.

[0110] For both approaches, the layers are trained simultaneously on the training set, normally checking for overfitting on an isolated cross-validation set. Both techniques can yield excellent results, and there is considerable enthusiasm for both approaches. The optimal size, shape, and quantity of individual layers varies depending on the problem that is addressed by the respective neural network.

[0111] Returning to FIG. 6A, screenshots, video frames, other information, automation box information, etc. are provided as the input layer and fed as inputs to the J neurons of hidden layer 1. The other information may include, but is not limited to, web browser histories, heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the displays that users are interacting with, locations where the users were looking on the displays, time stamps associated with the screenshots and video frames, text that the users entered, content that the users scrolled past, times that the users stopped on parts of content shown in the displays, what applications the user were interacting with, voice inputs, gestures, emotion information, biometrics, information pertaining to periods of no user activity, haptic information, multi-touch input information, any combination thereof, etc. The automation box information may include time stamped input from user input devices. While all of these inputs are fed to each neuron in this example, various architectures are possible that may be used individually or in combination including, but not limited to, feed forward networks, radial basis networks, deep feed forward networks, deep convolutional inverse graphics networks, convolutional neural networks, recurrent neural networks, artificial neural networks, long/short term memory networks, gated recurrent unit networks, generative adversarial networks, liquid state machines, auto encoders, variational auto encoders, denoising auto encoders, sparse auto encoders, extreme learning machines, echo state networks, Markov chains, Hopfield networks, Boltzmann machines, restricted Boltzmann machines, deep residual networks, Kohonen networks, deep belief networks, deep convolutional networks, support vector machines, neural Turing machines, or any other suitable type or combination of neural networks without deviating from the scope of the invention.

[0112] Hidden layer 2 receives inputs from hidden layer 1, hidden layer 3 receives inputs from hidden layer 2, and so on for all hidden layers until the last hidden layer provides its outputs as inputs for the output layer. While multiple suggestions are shown here as output, in some embodiments, only a single output suggestion is provided. In certain embodiments, the suggestions are ranked based on confidence scores.

[0113] It should be noted that numbers of neurons I, J, K, and L are not necessarily equal. Thus, any desired number of layers may be used for a given layer of neural network 600 without deviating from the scope of the invention. Indeed, in

certain embodiments, the types of neurons in a given layer may not all be the same.

[0114] Neural network 600 is trained to assign confidence score(s) to appropriate outputs. In order to reduce predictions that are inaccurate, only those results with a confidence score that meets or exceeds a confidence threshold may be provided in some embodiments. For instance, if the confidence threshold is 80%, outputs with confidence scores exceeding this amount may be used and the rest may be ignored.

[0115] It should be noted that neural networks are probabilistic constructs that typically have confidence score(s). This may be a score learned by the AI/ML model based on how often a similar input was correctly identified during training. Some common types of confidence scores include a decimal number between 0 and 1 (which can be interpreted as a confidence percentage as well), a number between negative  $\infty$  and positive  $\infty$ , a set of expressions (e.g., “low,” “medium,” and “high”), etc. Various post-processing calibration techniques may also be employed in an attempt to obtain a more accurate confidence score, such as temperature scaling, batch normalization, weight decay, negative log likelihood (NLL), etc.

[0116] “Neurons” in a neural network are implemented algorithmically as mathematical functions that are typically based on the functioning of a biological neuron. Neurons receive weighted input and have a summation and an activation function that governs whether they pass output to the next layer. This activation function may be a nonlinear thresholded activity function where nothing happens if the value is below a threshold, but then the function linearly responds above the threshold (i.e., a rectified linear unit (ReLU) non-linearity). Summation functions and ReLU functions are used in deep learning since real neurons can have approximately similar activity functions. Via linear transforms, information can be subtracted, added, etc. In essence, neurons act as gating functions that pass output to the next layer as governed by their underlying mathematical function. In some embodiments, different functions may be used for at least some neurons.

[0117] An example of a neuron 610 is shown in FIG. 6B. Inputs  $x_1, x_2, \dots, x_n$  from a preceding layer are assigned respective weights  $w_1, w_2, \dots, w_n$ . Thus, the collective input from preceding neuron 1 is  $w_1x_1$ . These weighted inputs are used for the neuron’s summation function modified by a bias, such as:

$$\sum_{i=1}^m (w_i x_i) + bias \quad (1)$$

[0118] This summation is compared against an activation function  $f(x)$  to determine whether the neuron “fires”. For instance,  $f(x)$  may be given by:

$$f(x) = \begin{cases} 1 & \text{if } \sum wx + bias \geq 0 \\ 0 & \text{if } \sum wx + bias < 0 \end{cases} \quad (2)$$

[0119] The output  $y$  of neuron 610 may thus be given by:

$$y = f(x) \sum_{i=1}^m (w_i x_i) + bias \quad (3)$$



**[0120]** In this case, neuron **610** is a single-layer perceptron. However, any suitable neuron type or combination of neuron types may be used without deviating from the scope of the invention. It should also be noted that the ranges of values of the weights and/or the output value(s) of the activation function may differ in some embodiments without deviating from the scope of the invention.

**[0121]** A goal, or “reward function,” is often employed. A reward function explores intermediate transitions and steps with both short-term and long-term rewards to guide the search of a state space and attempt to achieve a goal (e.g., finding the most accurate answers to user inquiries based on associated metrics). During training, various labeled data is fed through neural network **600**. Successful identifications strengthen weights for inputs to neurons, whereas unsuccessful identifications weaken them. A cost function, such as mean square error (MSE) or gradient descent may be used to punish predictions that are slightly wrong much less than predictions that are very wrong. If the performance of the AI/ML model is not improving after a certain number of training iterations, a data scientist may modify the reward function, provide corrections of incorrect predictions, etc.

**[0122]** Backpropagation is a technique for optimizing synaptic weights in a feedforward neural network. Backpropagation may be used to “pop the hood” on the hidden layers of the neural network to see how much of the loss every node is responsible for, and subsequently updating the weights in such a way that minimizes the loss by giving the nodes with higher error rates lower weights, and vice versa. In other words, backpropagation allows data scientists to repeatedly adjust the weights so as to minimize the difference between actual output and desired output.

**[0123]** The backpropagation algorithm is mathematically founded in optimization theory. In supervised learning, training data with a known output is passed through the neural network and error is computed with a cost function from known target output, which gives the error for backpropagation. Error is computed at the output, and this error is transformed into corrections for network weights that will minimize the error.

**[0124]** In the case of supervised learning, an example of backpropagation is provided below. A column vector input  $x$  is processed through a series of  $N$  nonlinear activity functions  $f_i$  between each layer  $i = 1, \dots, N$  of the network, with the output at a given layer first multiplied by a synaptic matrix  $W_i$ , and with a bias vector  $b_i$  added. The network output  $o$ , given by

$$o = f_N(W_N f_{N-1}(W_{N-1} f_{N-2}(\dots f_1(W_1 x + b_1)\dots) + b_{N-1}) + b_N) \quad (4)$$

**[0125]** In some embodiments,  $o$  is compared with a target output  $t$ , resulting in an error

$$E = \frac{1}{2} \|o - t\|^2,$$

which is desired to be minimized.

**[0126]** Optimization in the form of a gradient descent procedure may be used to minimize the error by modifying the synaptic weights  $W_i$  for each layer. The gradient descent procedure requires the computation of the output  $o$  given an input  $x$  corresponding to a known target output  $t$ , and

producing an error  $o - t$ . This global error is then propagated backwards giving local errors for weight updates with computations similar to, but not exactly the same as, those used for forward propagation. In particular, the backpropagation step typically requires an activity function of the form

$$p_j(n_j) = f'_j(n_j),$$

where  $n_j$  is the network activity at layer  $j$  (i.e.,  $n_j = W_j o_{j-1} + b_j$ ) where  $o_j = f_j(n_j)$  and the apostrophe ' denotes the derivative of the activity function  $f$ .

**[0127]** The weight updates may be computed via the formulae:

$$d_j = \begin{cases} (o - t) \circ p_j(n_j), & j = N \\ W_{j+1}^T d_{j+1} \circ p_j(n_j), & j < N \end{cases} \quad (5)$$

$$\frac{\partial E}{\partial W_{j+1}} = d_{j+1} (o_j)^T \quad (6)$$

$$\frac{\partial E}{\partial b_{j+1}} = d_{j+1} \quad (7)$$

$$W_j^{\text{new}} = W_j^{\text{old}} - \eta \frac{\partial E}{\partial W_j} \quad (8)$$

$$b_j^{\text{new}} = b_j^{\text{old}} - \eta \frac{\partial E}{\partial b_j} \quad (9)$$

where  $\circ$  denotes a Hadamard product (i.e., the element-wise product of two vectors),  $^T$  denotes the matrix transpose, and  $o_j$  denotes  $f_j(W_j o_{j-1} + b_j)$ , with  $o_0 = x$ . Here, the learning rate  $\eta$  is chosen with respect to machine learning considerations. Below,  $\eta$  is related to the neural Hebbian learning mechanism used in the neural implementation. Note that the synapses  $W$  and  $b$  can be combined into one large synaptic matrix, where it is assumed that the input vector has appended ones, and extra columns representing the  $b$  synapses are subsumed to  $W$ .

**[0128]** The AI/ML model may be trained over multiple epochs until it reaches a good level of accuracy (e.g., 97% or better using an F2 or F4 threshold for detection and approximately 2,000 epochs). This accuracy level may be determined in some embodiments using an F1 score, an F2 score, an F4 score, or any other suitable technique without deviating from the scope of the invention. Once trained on the training data, the AI/ML model may be tested on a set of evaluation data that the AI/ML model has not encountered before. This helps to ensure that the AI/ML model is not “over fit” such that it performs well on the training data, but does not perform well on other data.

**[0129]** In some embodiments, it may not be known what accuracy level is possible for the AI/ML model to achieve. Accordingly, if the accuracy of the AI/ML model is starting to drop when analyzing the evaluation data (i.e., the model is performing well on the training data, but is starting to perform less well on the evaluation data), the AI/ML model may go through more epochs of training on the training data (and/or new training data). In some embodiments, the



AI/ML model is only deployed if the accuracy reaches a certain level or if the accuracy of the trained AI/ML model is superior to an existing deployed AI/ML model. In certain embodiments, a collection of trained AI/ML models may be used to accomplish a task. For example, one model may be trained to suggest next activities or sequences of activities, another model may perform CV, yet another model may attempt runtime repair if the automation fails, etc.

**[0130]** Some embodiments may use transformer networks such as SentenceTransformers™, which is a Python™ framework for state-of-the-art sentence, text, and image embeddings. Such transformer networks learn associations of words and phrases that have both high scores and low scores. This trains the AI/ML model to determine what is close to the input and what is not, respectively. Rather than just using pairs of words/phrases, transformer networks may use the field length and field type, as well.

**[0131]** Natural language processing (NLP) techniques such as word2vec, BERT, GPT-3, ChatGPT, other LLMs, etc. may be used in some embodiments to facilitate semantic understanding and provide more accurate and human-like answers. Other techniques, such as clustering algorithms, may be used to find similarities between groups of elements. Clustering algorithms may include, but are not limited to, density-based algorithms, distribution-based algorithms, centroid-based algorithms, hierarchy-based algorithms. K-means clustering algorithms, the DBSCAN clustering algorithm, the Gaussian mixture model (GMM) algorithms, the balance iterative reducing and clustering using hierarchies (BIRCH) algorithm, etc. Such techniques may also assist with categorization.

**[0132]** FIG. 7 is a flowchart illustrating a process 700 for training AI/ML model(s), according to an embodiment of the present invention. In some embodiments, the AI/ML model(s) may be generative AI models, per the above. The neural network architecture of AI/ML models typically includes multiple layers of neurons, including input, output, and hidden layers. See FIGS. 6A and 6B, for example. The input layer receives the input(s) and the output layer generates the response(s). The hidden layers in between process the input data and generate intermediate representations of the input that are used to generate the output. These hidden layers can include various types of neurons, such as convolutional neurons, recurrent neurons, and/or transformer neurons.

**[0133]** The training process begins with providing screenshots, video frames, other information, and automation box information, whether labeled or unlabeled, at 710. The AI/ML model is then trained over multiple epochs at 720 and results are reviewed at 730. While various types of AI/ML models may be used, LLMs and other generative AI models are typically trained using a process called “supervised learning”, which is also discussed above. Supervised learning involves providing the model with a large dataset, which the model uses to learn the relationships between the inputs and outputs. During the training process, the model adjusts the weights and biases of the neurons in the neural network to minimize the difference between the predicted outputs and the actual outputs in the training dataset.

**[0134]** One aspect of the models in some embodiments is the use of transfer learning. For instance, transfer learning may take advantage of a pretrained model, such as ChatGPT, which is fine-tuned on a specific task or domain in step 720. This allows the model to leverage the knowl-

edge already learned from the pretraining phase and adapt it to a specific application via the training phase of step 720.

**[0135]** The pretraining phase involves training the model on an initial set of training data that may be more general. During this phase, the model learns relationships in the data. In the fine-tuning phase (e.g., performed during step 720 in addition to or in lieu of the initial training phase in some embodiments if a pretrained model is used as the initial basis for the final model), the pretrained model is adapted to a specific task or domain by training the model on a smaller dataset that is specific to the task. For instance, in some embodiments, the model may be focused on UI automations and take advantage of an object repository. Fine-tuning allows the model to learn the nuances of the task, such as the specific vocabulary and syntax, certain graphical characteristics, certain data formats, etc., without requiring as much data as would be necessary to train the model from scratch. By leveraging the knowledge learned in the pretraining phase, the fine-tuned model can achieve state-of-the-art performance on specific tasks with relatively little additional training data.

**[0136]** If the AI/ML model fails to meet a desired confidence threshold at 740, the training data is supplemented and/or the reward function is modified to help the AI/ML model achieve its objectives better at 750 and the process returns to step 720. If the AI/ML model meets the confidence threshold at 740, the AI/ML model is tested on evaluation data at 760 to ensure that the AI/ML model generalizes well and that the AI/ML model is not over fit with respect to the training data. The evaluation data includes information that the AI/ML model has not processed before. If the confidence threshold is met at 770 for the evaluation data, the AI/ML model is deployed at 780. If not, the process returns to step 750 and the AI/ML model is trained further.

**[0137]** FIG. 8 is an architectural diagram illustrating a system 800 configured to train a generative AI/ML model to recognize applications, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements, according to an embodiment of the present invention. In some embodiments, system 800 may be part of hyper-automation system 100 of FIG. 1. System 800 includes user computing systems, such as desktop computer 802, tablet 804, and smart phone 806. However, any desired computing system may be used without deviating from the scope of invention including, but not limited to, smart watches, laptop computers, etc. In some embodiments, one or more of computing systems 802, 804, 806 may include automation boxes and/or cameras. Also, while three user computing systems are shown in FIG. 8, any suitable number of computing systems may be used without deviating from the scope of the invention. For instance, in some embodiments, dozens, hundreds, thousands, or millions of computing systems may be used.

**[0138]** Each computing system 802, 804, 806 has a recorder process 810 (i.e., a tracking application) running thereon that records screenshots and/or video of the user’s screen or a portion thereof. For instance, a snippet of JavaScript® may be embedded in a web browser as recorder process 810 to track what components the user interacts with, what text the user enters, what locations/components the user clicks with the mouse, what content the user scrolled past, how long the user stopped on a certain part of the content, etc. Scrolling past content may indicate that the content may



have been somewhat close but did not have exactly what the user wanted. Clicks may indicate success.

**[0139]** Recorder processes **810** need not be JavaScript®, and may be any suitable type of application and in any desired programming language without deviating from the scope of the invention. This may allow “generalizing” of recorder processes **810** such that they can track user interactions with multiple applications or any application that the user is interacting with. Using labeled training data from scratch may be difficult since while it may allow the generative AI/ML model to learn to recognize various controls, it does not contain information regarding which controls are commonly used and how. Using recorder processes **810**, a “heat map” could be generated to help bootstrap the generative AI/ML model training process. The heat map may include various information, such as the frequency that a user used applications, the frequency that the user interacted with components of the applications, the locations of the components, the content of the applications/components, etc. In some embodiments, the heat map may be derived from screen analysis, such as detection of typed and/or pasted text, caret tracking, and active element detection for computing systems **802**, **804**, **806**. Some embodiments recognize where on the screen associated with computing systems **802**, **804**, **806** the user has been typing or pasting text, potentially including hot keys or other keys that do not cause visible characters to appear, and provide the physical position on the screen based on the current resolution (e.g., in coordinates) of where one or more characters appeared, where the cursor was blinking, or both. The physical position of the typing or pasting activity and/or of the caret may allow determination of which field(s) the user is typing in or focused on and what the application is for process discovery or other applications.

**[0140]** Per the above, in some embodiments, recorder processes **810** may record additional data to further assist in training generative AI/ML model(s), such as web browser history, heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the screen that the user is interacting with, locations where the user was looking on the screen at different times, time stamps associated with the screenshot/video frames, etc. This may be beneficial for providing key presses and/or other user actions that may not cause a screen change. For instance, some applications may not provide a visual change when the user presses CTRL+S to save a file. However, in certain embodiments, the generative AI/ML model(s) may be trained solely based on captured screen images. Recorder processes **810** may be robots generated via an RPA designer application, part of an operating system, a downloadable application for a personal computer (PC) or smart phone, or any other software and/or hardware without deviating from the scope of the invention. Indeed, in some embodiments, the logic of one or more of recorder processes **810** is implemented partially or completely via physical hardware.

**[0141]** Some embodiments are implemented in a feedback loop process that continuously or periodically compares the current screenshot to the previous screenshot to identify changes. Locations where visual changes occurred on the screen may be identified and OCR may be performed on the location where the change occurred. Results of the OCR may then be compared to the content of a keyboard queue (e.g., as determined by key hooking) to determine

whether a match exists. The locations where the change occurred may be determined by comparing a box of pixels from the current screenshot to a box of pixels in the same location from a previous screenshot.

**[0142]** Images and/or other data (e.g., web browser history, heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the screen that the user is interacting with, locations where the user was looking on the screen at different times, time stamps associated with the screenshot/video frames, voice inputs, gestures, emotions (e.g., whether a user is happy, frustrated, etc.), biometrics (e.g., fingerprints, retinal scans, the user’s pulse, etc.), information pertaining to periods of no user activity (e.g., “dead man switches”), haptic information from a haptic display or touch pad, a heat map with multi-touch inputs, etc.) that are recorded by recorder processes **810** are sent via a network **820** (e.g., a local area network (LAN), a mobile communications network, a satellite communications network, the Internet, any combination thereof, etc.) to a server **830**. In some embodiments, server **830** may be part of a public cloud architecture, a private cloud architecture, a hybrid cloud architecture, etc. In certain embodiments, server **830** may host multiple software-based servers on a single computing system **830**. In some embodiments, server **830** may run a conductor application and the data from recorder processes **810** may be sent periodically as part of the heartbeat message. In certain embodiments, the data may be sent from recorder processes **810** to server **830** once a predetermined amount of data has been collected, after a predetermined time period has elapsed, or both. Server **830** stores the received data from recorder processes **810** in a database **840**.

**[0143]** Server **830** includes multiple AI layers **832** in this embodiment that collectively form a generative AI/ML model. However, in some embodiments, generative AI/ML models may only have a single layer. In certain embodiments, multiple generative AI/ML models may be trained on server **830** and used together to collectively accomplish a larger task. AI layers **832** may employ CV techniques and may perform various functions, such as statistical modeling (e.g., hidden Markov models (HMMs)), and utilize deep learning techniques (e.g., long short term memory (LSTM) deep learning, encoding of previous hidden states, etc.) to identify user interactions. Initially, the generative AI/ML model needs to be trained such that it can perform meaningful analysis of the captured data in database **840**. In some embodiments, users of computing systems **802**, **804**, **806** label the images before they are sent to server **830**. Additionally or alternatively, in some embodiments, labeling occurs subsequently, such as via an application **852** running on computing system **850** that allows users to draw bounding boxes and/or other shapes around graphical elements, provide text labels of what is contained in a bounding box, etc.

**[0144]** The generative AI/ML model goes through a training phase using this data as input and is trained until the generative AI/ML model is sufficiently accurate, but not overfit to the training data. The accuracy that is acceptable may depend on the application. Identification errors can be highlighted by a human reviewer (e.g., by drawing a box around the misidentified element and including the correct identification), and the generative AI/ML model can be retrained using this additional labeled data. Once adequately trained, the generative AI/ML model is able to provide pre-



dictions as output, such as which application(s) and graphical element(s) are recognized as being present in a screen.

**[0145]** However, while this level of training provides information regarding what is present, further information may be needed to determine user interactions, such as comparing two or more consecutive screens to determine that a typed character appeared from one to another, a button was pressed, a menu selection occurred, etc. Thus, after the generative AI/ML model can recognize graphical elements and applications on the screen, in some embodiments, the generative AI/ML model is further trained to recognize labeled user interactions with the applications in the UI to understand such incremental actions taken by the user. Identification errors can again be highlighted by a human reviewer (e.g., by drawing a box around the misidentified element and including the correct identification), and the generative AI/ML model can be trained until its accuracy is sufficiently high to be deployed to understand granular user interactions with the UI.

**[0146]** Once trained to recognize user interactions, the trained generative AI/ML model may be used to analyze video and/or other information from recorder processes **810**. This recorded information may include interactions that multiple/many users tend to perform. These interactions can then be analyzed for common sequences for subsequent automation.

#### AI Layers

**[0147]** In some embodiments, multiple AI layers may be used. Each AI layer is an algorithm (or model) that runs on the data, and the AI model itself may be deep learning neural networks (DLNNs) of trained artificial “neurons” that are trained in training data. Layers may be run in series, in parallel, or a combination thereof.

**[0148]** The AI layers may include, but are not limited to, a sequence extraction layer, a clustering detection layer, a visual component detection layer, a text recognition layer (e.g., OCR), an audio-to-text translation layer, or any combination thereof. However, any desired number and type(s) of layers may be used without deviating from the scope of the invention. Using multiple layers may allow the system to develop a global picture of what is happening in the screens. For example, one AI layer could perform OCR, another could detect buttons, etc.

**[0149]** Patterns may be determined individually by an AI layer or collectively by multiple AI layers. A probability or an output in terms of a user action could be used. For instance, to determine the particulars of a button, its text, where a user clicked, etc., the system may need to know where the button is, its text, the positioning on the screen, etc.

**[0150]** However, it should be noted that various generative AI/ML models may be used without deviating from the scope of the invention. While generative AI/ML models may be trained using neural networks in some embodiments, such as DLNNs, recurrent neural networks (RNNs), GANs, LLMs, VAEs, transformers, any combination thereof, etc., other AI techniques may be used, such as deterministic models, shallow learning neural networks (SLNNs), or any other suitable AI/ML model types and training techniques without deviating from the scope of the invention.

**[0151]** FIG. 9 is a flowchart illustrating a process **900** for training a generative AI/ML model to recognize applica-

tions, screens, and UI elements using CV and to recognize user interactions with the applications, screens, and UI elements, according to an embodiment of the present invention. The process begins with recording screenshots or video frames displays associated with user computing systems and other information at **910**. In some embodiments, the recording is performed by one or more recorder processes. In certain embodiments, the recorder processes are implemented as feedback loop processes that continuously or periodically compare a current screenshot or video frame to a previous screenshot or video frame and identify one or more locations where changes between the current screenshot or video frame and the previous screenshot or video frame occurred. In some embodiments, the recorder processes are configured to perform OCR on the one or more locations where the changes occurred, compare results of the OCR to content of a keyboard queue to determine whether a match exists, and when a match exists, link text associated with the match to a respective location. In some embodiments, the other information includes a web browser history, one or more heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the display that a user is interacting with, locations where the user was looking on the display, time stamps associated with the screenshots or video frames, text that the user entered, content that the user scrolled past, a time that the user stopped on a part of content shown in the display, what application the user is interacting with, or a combination thereof. In certain embodiments, at least part of the other information is captured using one or more automation boxes.

**[0152]** One or more heat maps are generated as part of the other information at **920**. In some embodiments, the one or more heat maps include a frequency that a user used applications, a frequency that the user interacted with components of the applications, locations of the components in the applications, content of the applications and/or components, or a combination thereof. In certain embodiments, the one or more heat maps are derived from display analysis that includes detection of typed and/or pasted text, caret tracking, active element detection, or a combination thereof. The recorded screenshots or video frames, and the other information, are then sent to storage accessible by one or more servers at **930**.

**[0153]** The recorded screenshots or video frames and the other information are accessed at **940** (e.g., via a server configured to train a generative AI/ML model). The generative AI/ML model is initially trained to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information at **950**. In some embodiments, the initial training of the generative AI/ML model is performed without a priori knowledge of the applications, screens, and UI elements in the screenshots or video frames.

**[0154]** After the generative AI/ML model can recognize the applications, screens, and UI elements in the recorded screenshots or video frames with a confidence (e.g., 70%, 95%, 99.99%, etc.), the generative AI/ML model is trained to recognize individual user interactions with the UI elements at **960**. In some embodiments, the individual user interactions include button presses, entry of single characters or character sequences, selection of active UI elements, menu selections, screen changes, or a combination thereof. In certain embodiments, the training of the generative AI/



ML model to recognize the individual user interactions with the UI elements includes comparing two or more consecutive screenshots or video frames and determining that a typed character appeared from one to another, a button was pressed, or a menu selection occurred. The generative AI/ML model is then deployed such that it can be called and used by calling processes (e.g., RPA robots) at 970.

[0155] FIG. 10 is an architectural diagram illustrating an automation box and eye movement tracking system 1000, according to an embodiment of the present invention. System 1000 includes a computing system 1010 that includes eye tracking logic (ETL) 1012 that is configured to process input from a camera 1020 and automation box logic (ABL) 1014 that is configured to process input from an automation box 1060. In some embodiments, computing system 1010 may be or include computing system 500 of FIG. 5. In certain embodiments multiple cameras may be used.

[0156] Camera 1020 records video of a user while the user is interacting with computing system 1010 via a mouse 1040 and a keyboard 1050. Computing system 1010 converts the recorded camera video into video frames. ETL processes these frames identifies the user's eyes and interpolates the location where the user is looking to a location on a display 1030. Any suitable eye tracking technique(s) may be used without deviating from the scope of the invention, such as those described in U.S. Pat. Application Publication No 2018/0046248, U.S. Pat. No. 7,682,026, etc. Time stamps may be associated with the video frames of the user such that they can be matched with screenshot frames of what is displayed on display 1030 at that time.

[0157] An automation box 1060 also includes automation box logic 1062 in this embodiment that receives input from mouse 1040 and keyboard 1050. In some embodiments, automation box 1060 may have similar hardware to computing system 1010 (e.g., processor(s), memory, a bus, etc.). This input may then be passed along to computing system 1010. While mouse 1040 and keyboard 1050 are shown in FIG. 10, any suitable input device(s) may be used without deviating from the scope of the invention, such as a touchpad, buttons, etc. In some embodiments, only computing system 1010 or automation box 1060 include automation box logic. A reason for the latter may be to record user interactions and send them directly to a server (e.g., a cloud-based server) for subsequent processing via network 1070. In such embodiments, screenshot frames may also be sent from computing system 1010 to automation box 1060 and then on to the server via network 1070. Alternatively, computing system 1010 may send the screenshots itself via network 1070. Such embodiments may provide a plug-and-play tracking solution that can be plugged into computing system 1010, relay keyboard and mouse information to computing system 1010 for its operations, and also relay keyboard and mouse click information to a remote server for subsequent training of a generative AI/ML model.

[0158] In some embodiments, automation box 1060 may include actuation logic that runs an automation and simulates inputs. This may allow automation box 1060 to provide simulated key presses, mouse movements and clicks, etc. to computing system 1010 as though this information was actually coming from a human user interacting with these components. The UI screenshots and other information may then be used to train the generative AI/ML model. Another advantage of such embodiments is that the generative AI/ML model can be trained when the user is away

from computing system 1010, potentially allowing for larger amounts of training information to be captured more quickly, and thus, also potentially allowing for the generative AI/ML model to be trained more quickly.

[0159] In certain embodiments, an "information box" may be implemented as software on computing system 810 and may function in a similar manner to recorder processes 810 of FIG. 8. Such embodiments may store screenshot frames, mouse click information, and key press information. In certain embodiments, eye tracking information may also be tracked. This information may then be sent to a server via network 1070, and gaze tracking may potentially be performed remotely rather than on computing system 1010.

[0160] The process steps performed in FIG. 9 may be performed by a computer program, encoding instructions for the processor(s) to perform at least part of the process(es) described in FIG. 9, in accordance with embodiments of the present invention. The computer program may be embodied on a non-transitory computer-readable medium. The computer-readable medium may be, but is not limited to, a hard disk drive, a flash device, RAM, a tape, and/or any other such medium or combination of media used to store data. The computer program may include encoded instructions for controlling processor(s) of a computing system (e.g., processor(s) 510 of computing system 500 of FIG. 5) to implement all or part of the process steps described in FIG. 9, which may also be stored on the computer-readable medium.

[0161] The computer program can be implemented in hardware, software, or a hybrid implementation. The computer program can be composed of modules that are in operative communication with one another, and which are designed to pass information or instructions to display. The computer program can be configured to operate on a general purpose computer, an ASIC, or any other suitable device.

[0162] It will be readily understood that the components of various embodiments of the present invention, as generally described and illustrated in the figures herein, may be arranged and designed in a wide variety of different configurations. Thus, the detailed description of the embodiments of the present invention, as represented in the attached n, is not intended to limit the scope of the invention as claimed, but is merely representative of selected embodiments of the invention.

[0163] The features, structures, or characteristics of the invention described throughout this specification may be combined in any suitable manner in one or more embodiments. For example, reference throughout this specification to "certain embodiments," "some embodiments," or similar language means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases "in certain embodiments," "in some embodiment," "in other embodiments," or similar language throughout this specification do not necessarily all refer to the same group of embodiments and the described features, structures, or characteristics may be combined in any suitable manner in one or more embodiments.

[0164] It should be noted that reference throughout this specification to features, advantages, or similar language does not imply that all of the features and advantages that may be realized with the present invention should be or are in any single embodiment of the invention. Rather, language



referring to the features and advantages is understood to mean that a specific feature, advantage, or characteristic described in connection with an embodiment is included in at least one embodiment of the present invention. Thus, discussion of the features and advantages, and similar language, throughout this specification may, but do not necessarily, refer to the same embodiment.

**[0165]** Furthermore, the described features, advantages, and characteristics of the invention may be combined in any suitable manner in one or more embodiments. One skilled in the relevant art will recognize that the invention can be practiced without one or more of the specific features or advantages of a particular embodiment. In other instances, additional features and advantages may be recognized in certain embodiments that may not be present in all embodiments of the invention.

**[0166]** One having ordinary skill in the art will readily understand that the invention as discussed above may be practiced with steps in a different order, and/or with hardware elements in configurations which are different than those which are disclosed. Therefore, although the invention has been described based upon these preferred embodiments, it would be apparent to those of skill in the art that certain modifications, variations, and alternative constructions would be apparent, while remaining within the spirit and scope of the invention. In order to determine the metes and bounds of the invention, therefore, reference should be made to the appended claims.

**1. A system, comprising:**

one or more user computing systems comprising respective recorder processes; and

a server configured to train a generative artificial intelligence (AI) / machine learning (ML) model to recognize applications, screens, and user interface (UI) elements using computer vision (CV) and to recognize user interactions with the applications, screens, and UI elements, wherein

the respective recorder processes are configured to:

record screenshots or video frames of a display associated with the respective user computing system and other information, and

send the recorded screenshots or video frames, and the other information, to storage accessible by the server, and

the server is configured to:

initially train the generative AI/ML model to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information, and

after the generative AI/ML model can recognize the applications, screens, and UI elements in the recorded screenshots or video frames with a confidence, train the generative AI/ML model to recognize individual user interactions with the UI elements.

**2. The system of claim 1, wherein the individual user interactions comprise button presses, entry of single characters or character sequences, selection of active UI elements, menu selections, screen changes, voice inputs, gestures, providing biometric information, haptic interactions, or a combination thereof.**

**3. The system of claim 1, wherein the training of the generative AI/ML model to recognize the individual user**

interactions with the UI elements comprises comparing two or more consecutive screenshots or video frames and determining that a typed character appeared from one screenshot to another, a button was pressed, or a menu selection occurred.

**4. The system of claim 1, wherein the other information comprises a web browser history, one or more heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the display that a user is interacting with, locations where the user was looking on the display, time stamps associated with the screenshots or video frames, text that the user entered, content that the user scrolled past, a time that the user stopped on a part of content shown in the display, what application the user is interacting with, voice inputs, gestures, emotion information, biometrics, information pertaining to periods of no user activity, haptic information, multi-touch input information, or a combination thereof.**

**5. The system of claim 1, wherein**

the one or more user computing systems or the server are configured to generate one or more heat maps, the other information comprising the one or more heat maps, and the one or more heat maps comprise a frequency that a user used applications, a frequency that the user interacted with components of the applications, locations of the components in the applications, content of the applications and/or components, or a combination thereof.

**6. The system of claim 5, wherein the one or more user computing systems or the server are configured to derive the one or more heat maps from display analysis comprising detection of typed and/or pasted text, caret tracking, active element detection, or a combination thereof.**

**7. The system of claim 1, wherein the respective recorder processes are implemented as feedback loop processes that continuously or periodically compare a current screenshot or video frame to a previous screenshot or video frame and identify one or more locations where changes between the current screenshot or video frame and the previous screenshot or video frame occurred.**

**8. The system of claim 7, wherein the respective recorder processes are further configured to:**

perform optical character recognition (OCR) on the one or more locations where the changes occurred;

compare results of the OCR to content of a keyboard queue to determine whether a match exists; and

when a match exists, link text associated with the match to a respective location.

**9. The system of claim 1, further comprising:**

an automation box operably connected to a user computing system of the one or more user computing systems, the automation box configured to:

receive input from one or more user input devices,

associate time stamps with the input, and

send the time stamped input to storage accessible by the server, wherein

the server is configured to use the time stamped input for the initial training of the generative AI/ML model.

**10. The system of claim 1, wherein server is configured to perform the initial training of the generative AI/ML model without a priori knowledge of the applications, screens, and UI elements in the screenshots or video frames.**

**11. The system of claim 1, wherein the generative AI model is a large language model (LLM), a generative adversarial network (GAN), a variational autoencoder (VAE), or a transformer.**

**12. A non-transitory computer-readable medium storing a computer program configured to train a generative artificial**



intelligence (AI)/machine learning (ML) model to recognize applications, screens, and user interface (UI) elements using computer vision (CV) and/or to recognize user interactions with the applications, screens, and UI elements, the computer program configured to cause at least one processor to:

- access recorded screenshots or video frames of displays associated with one or more computing systems and access other information associated with the one or more computing systems; and

- initially train the generative AI/ML model to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information, wherein

- the initial training of the generative AI/ML model is performed without a priori knowledge of the applications, screens, and UI elements in the screenshots or video frames.

**13.** The non-transitory computer-readable medium of claim **12**, wherein after the generative AI/ML model can recognize the applications, screens, and UI elements in the recorded screenshots or video frames with a confidence, the computer program is further configured to cause the at least one processor to:

- train the generative AI/ML model to recognize individual user interactions with the UI elements.

**14.** The non-transitory computer-readable medium of claim **13**, wherein the training of the generative AI/ML model to recognize the individual user interactions with the UI elements comprises comparing two or more consecutive screenshots or video frames and determining that a typed character appeared from one to another, a button was pressed, or a menu selection occurred.

**15.** The non-transitory computer-readable medium of claim **13**, wherein the individual user interactions comprise button presses, entry of single characters or character sequences, selection of active UI elements, menu selections, screen changes, voice inputs, gestures, providing biometric information, haptic interactions, or a combination thereof.

**16.** The non-transitory computer-readable medium of claim **12**, wherein the other information comprises a web browser history, one or more heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the display that a user is interacting with, locations where the user was looking on the display, time stamps associated with the screenshots or video frames, text that the user entered, content that the user scrolled past, a time that the user stopped on a part of content shown in the display, what application the user is interacting with, voice inputs, gestures, emotion information, biometrics, information pertaining to periods of no user activity, haptic information, multi-touch input information, or a combination thereof.

**17.** The non-transitory computer-readable medium of claim **12**, wherein the computer program is further configured to cause the at least one processor to:

- generate one or more heat maps, the other information comprising the one or more heat maps, wherein

- the one or more heat maps comprise a frequency that a user used one or more applications, a frequency that the user interacted with components of the one or more applications, locations of the components in the one or more applications, content of the one or more applications and/or components, or a combination thereof.

**18.** The non-transitory computer-readable medium of claim **17**, wherein the one or more heat maps are derived from display analysis comprising detection of typed and/or pasted text, caret tracking, active element detection, or a combination thereof.

**19.** The non-transitory computer-readable medium of claim **12**, wherein the generative AI model is a large language model (LLM), a generative adversarial network (GAN), a variational autoencoder (VAE), or a transformer.

**20.** A computer-implemented method for training a generative artificial intelligence (AI)/machine learning (ML) model to recognize applications, screens, and user interface (UI) elements using computer vision (CV) and to recognize user interactions with the applications, screens, and UI elements, the method comprising:

- accessing recorded screenshots or video frames of displays associated with one or more computing systems and accessing other information associated with the one or more computing systems;

- initially training the generative AI/ML model to recognize the applications, screens, and UI elements that are present in the recorded screenshots or video frames using the recorded screenshots or video frames and the other information; and

- after the generative AI/ML model can recognize the applications, screens, and UI elements in the recorded screenshots or video frames with a confidence, training the generative AI/ML model to recognize individual user interactions with the UI elements.

**21.** The computer-implemented method of claim **20**, wherein the initial training of the generative AI/ML model is performed without a priori knowledge of the applications, screens, and UI elements in the screenshots or video frames.

**22.** The computer-implemented method of claim **20**, wherein the training of the generative AI/ML model to recognize the individual user interactions with the UI elements comprises comparing two or more consecutive screenshots or video frames and determining that a typed character appeared from one to another, a button was pressed, or a menu selection occurred.

**23.** The computer-implemented method of claim **20**, wherein the individual user interactions comprise button presses, entry of single characters or character sequences, selection of active UI elements, menu selections, screen changes, voice inputs, gestures, providing biometric information, haptic interactions, or a combination thereof.

**24.** The computer-implemented method of claim **20**, wherein the other information comprises a web browser history, one or more heat maps, key presses, mouse clicks, locations of mouse clicks and/or graphical elements on the display that a user is interacting with, locations where the user was looking on the display, time stamps associated with the screenshots or video frames, text that the user entered, content that the user scrolled past, a time that the user stopped on a part of content shown in the display, what application the user is interacting with, voice inputs, gestures, emotion information, biometrics, information pertaining to periods of no user activity, haptic information, multi-touch input information, or a combination thereof.

**25.** The computer-implemented method of claim **20**, further comprising:

- generating one or more heat maps, the other information comprising the one or more heat maps, wherein



the one or more heat maps comprise a frequency that a user used one or more applications, a frequency that the user interacted with components of the one or more applications, locations of the components in the one or more applications, content of the one or more applications and/or components, or a combination thereof, and

the one or more heat maps are derived from display analysis comprising detection of typed and/or pasted text, caret tracking, active element detection, or a combination thereof.

**26.** The computer-implemented method of claim **20**, wherein the generative AI model is a large language model (LLM), a generative adversarial network (GAN), a variational autoencoder (VAE), or a transformer.

\* \* \* \* \*