

Autopilot

Automating behavioral experiments with lots of Raspberry Pis

JONNY L. SAUNDERS, MICHAEL WEHR @

May 9, 2022

Neuroscience needs behavior, and behavioral experiments require the coordination of large numbers of heterogeneous hardware components and data streams. Currently available tools strongly limit the complexity and reproducibility of experiments. Here we introduce Autopilot, a complete, open-source Python framework for behavioral neuroscience that distributes experiments over networked swarms of Raspberry Pis. Autopilot enables qualitatively greater experimental flexibility by allowing arbitrary numbers of hardware components to be combined in arbitrary experimental designs. Research is made reproducible by documenting all data and task design parameters in a human-readable and publishable format at the time of collection. Autopilot provides an order-of-magnitude performance improvement over existing tools while also being an order of magnitude less costly to implement. Autopilot's flexible, scalable architecture allows neuroscientists to design the next generation of experiments to investigate the behaving brain.

version: v2-20-gc4f7bfb



[HOMEPAGE](#)



[REPOSITORY](#)



[DOCUMENTATION](#)

Contents

1	<i>Introduction</i>	5
	1.1 <i>Existing Systems for Behavioral Experiments</i>	7
	1.2 <i>Limitations of Existing Systems</i>	8
2	<i>Design</i>	11
	2.1 <i>Efficiency</i>	11
	2.2 <i>Flexibility</i>	13
	2.3 <i>Reproducibility</i>	15
3	<i>Program Structure</i>	19
	3.1 <i>Directory Structure</i>	19
	3.2 <i>Data</i>	20
	3.3 <i>Tasks</i>	22
	3.4 <i>Hardware</i>	27
	3.5 <i>Transforms</i>	28
	3.6 <i>Stimuli</i>	29
	3.7 <i>Agents</i>	29
	3.8 <i>Networking</i>	31
	3.9 <i>GUI & Plots</i>	33
4	<i>Tests</i>	35
	4.1 <i>GPIO Latency</i>	35
	4.2 <i>Sound Latency</i>	37
	4.3 <i>Network Latency</i>	37
	4.4 <i>Network Bandwidth</i>	38
	4.5 <i>Distributed Go/No-go Task</i>	39

5 *Limitations and Future Directions* 41

6 *Glossary* 43

Bibliography 47

1

Introduction

ANIMAL BEHAVIOR experiments need precision and patience, so we make computers do them for us. The complexity of contemporary behavioral experiments, however, presents a stiff methodological challenge. For example, researchers might wish to measure pupil dilation[41, 41], respiration[37], and running speed[36], while tracking the positions of body parts in 3 dimensions[35] and recording the activity of large ensembles of neurons[24], as subjects perform tasks with custom input devices such as a steering wheel[9] while immersed in virtual reality environments using stimuli synthesized in real time[51, 11]. Coordinating the array of necessary hardware into a coherent experimental design—with the millisecond precision required to study the brain—can be daunting.

Historically, researchers have developed software to automate behavior experiments as-needed within their lab or relied on purchasing proprietary software (eg. [16]). Open-source alternatives have emerged recently, often developed in tandem with hardware peripherals available for purchase [18, 43]. However, the diverse hardware and software requirements for behavioral experiments often lead researchers to cobble together multiple tools to perform even moderately complex experiments. Understandably, most software packages do not attempt to simultaneously support custom hardware operation, behavioral task logic, stimulus generation, and data acquisition. The difficulty of designing and maintaining lab-idiosyncratic systems thus defines much of the everyday practice of science. Idiosyncratic systems can hinder reproducibility, especially if the level of detail reported in a methods section is sparse[54]. Additionally, development time and proprietary software are expensive, as are the custom hardware peripherals that are required to use most available open-source behavior software, stratifying access to state-of-the-art techniques according to inequitable funding distributions.

Technical challenges are never merely technical: they reflect and are structured by underlying *social* challenges in the organization of scientific labor and knowledge work. Lab infrastructure occupies a space between technology intended for individual users and for large organizations: that of *groupware*¹ [21, 22]. Experimental frameworks then face the joint challenge of technical competency while also embedding in and supporting existing cultures of practice. Behind every line of code is an unwritten wealth of technical knowledge needed to make use of it, as well as an unspoken set of beliefs about how it is to be used — labs aren’t born fresh on release day ready to retool at a moment’s notice, they’re held together by decades of duct tape and run on ritual. The boundaries of this “contextual knowledge” extend fluidly beyond individual labs, structuring disciplinary, status, and role systems in scientific work[7]. Given their position at the intersection of scientific theory, technical work, data production, and social organization; experimental frameworks are an elusive design challenge: but also an underexplored means of realizing some of our loftier dreams of open, accessible, and collaborative science.

Here we present Autopilot, a complete open-source software and hardware frame-

¹ “Our original definition of groupware was ‘intentional group processes plus software to support them.’ It has both *computer* and *human* components: software of the computer and ‘software’ of the people using it. [...] Recently this definition has been extended to include other more expressly cultural factors including myth, values and norms. The computer software should reflect and support a group’s purpose, process and culture.”

Peter and Trudy Johnson-Lenz (1991)[22]

work for behavioral experiments. We leverage the power of distributed computing using the surprisingly capable Raspberry Pi 4² to allow researchers to coordinate arbitrary numbers of heterogeneous hardware components in arbitrary experimental designs.

Autopilot takes a different approach than existing systems to overcome the technical challenges of behavioral research: *just use more computers*. Specifically, the advent of inexpensive single-board computers (ie. the Raspberry Pi) that are powerful enough to run a full Linux operating system allows a unified platform to run on every Pi or other computer in the system so that they can work together seamlessly. At the core of its architecture is a networking protocol (Section 3.8) that is fast enough to stream electrophysiological or imaging data and flexible enough to make the mutual coordination of hardware straightforward.

This distributed design also makes Autopilot extremely scalable, as the Raspberry Pi's \$35 price tag makes it an order of magnitude less costly than comparable systems (Section 2.3). Its low cost doesn't come at the expense of performance or usability: Autopilot also has an order of magnitude greater measurement precision and an order of magnitude lower latency than comparable systems (Sections 2.1 and 4).

Autopilot balances experimental flexibility with support. Its task design infrastructure is flexible enough to perform arbitrary experiments, but also provides support for data management, plotting task progress, and custom training regimens. We try to bridge multiple modalities of use: use its modular framework of tools out of the box, or use its [complete low-level API documentation](https://docs.auto-pi-lot.com)³ to hack at it until it does what you need. Rather than relying on costly proprietary hardware modules, users can take advantage of the wide array of peripherals and extensive community support available for the Raspberry Pi. Autopilot is designed to be *permissive*: build your whole experiment with it or just use its networking modules, adapt it to existing hardware, integrate your favorite analysis tool. We want Autopilot to *play nice* with other software libraries and existing practices, rather than forcing you to retool your lab around it.

Finally, we have designed Autopilot to help scientists do reproducible research and be good stewards of the human knowledge project. Experiments are not written as scripts that are reliant on the particularities of each researcher's hardware configuration. Instead, we have designed the system to encourage users to write reusable, portable experiments that can be incorporated into a public central library while also allowing space to iterate and refine without needing to learn complicated programming best-practices to contribute. Every parameter that defines an experiment is automatically saved in publication-ready format, removing ambiguity in reported methods and facilitating exact replication with a single file. Its plugin system is built atop a densely-linked [semantic wiki](https://wiki.auto-pi-lot.com)⁴ that fluidly combines human- and computer-readable, communally editable technical knowledge that surrounds your experiments with the software that performs them.

We begin by defining the requirements of a complete behavioral system and evaluating two current examples (Sections 1.1 and 1.2). We then describe Autopilot's design principles (Section 2) and how they are implemented in the program's structure (Section 3). We close with a demonstration of its current capabilities and our plans to expand them (Sections 4 and 5).

² See Table 3.2

³ <https://docs.auto-pi-lot.com>

⁴ <https://wiki.auto-pi-lot.com>

We would like to acknowledge and thank Lucas Ott for doing much of the behavioral training, Brynna Paros and Nick Sattler for their help with constructing our behavioral boxes, Matt Smear and Reese Findley for loaning us their Bpod for far longer than they intended to, Erik Flister whose Ratrix software inspired some of the design features of Autopilot [33], several artists on [FlatIcon.com](https://flaticon.com) ([Freepik](#), [Nikita Golubev](#), [Those Icons](#)) whose work served as stems for many of the figures, and the Janet Smith House for the endless support and relentless criticism of the figures. This material is based on work supported by NIH NIDCD R01 DC-015828, NSF Graduate Research Fellowship No. 1309047, and a University of Oregon Incubating Interdisciplinary Initiatives award.

1.1 Existing Systems for Behavioral Experiments

At minimum, a complete system to automate behavioral experiments has 6 requirements:

1. **Hardware** to interact with the experimental subject, including **sensors** (eg. photodiodes, cameras, rotary encoders) to receive input and **actuators** (eg. lights, motors, solenoids) to provide feedback.
2. Some capability to synthesize and present sensory **stimuli**. Ideally both discrete stimuli, like individual tone pips or grating patches, and continuous stimuli, like those used in virtual reality experiments, should be possible.
3. A framework to coordinate hardware and stimuli as a **task**. Task definition should be flexible such that it facilitates rather than constrains experimental design.
4. A **data management** system that allows fine control of data collection and format. Data should be human readable and include complete metadata that allows independent analysis and reproduction. Ideally the program would also allow some means of realtime data processing of sensor values for use in a task.
5. Some means of **visualizing data** as it is collected in order to observe task status. It should be possible to customize visualization to the needs and structure of the task.
6. Finally, a **user interface** to control task operation. The UI should make it possible for someone who does not program to operate the system.

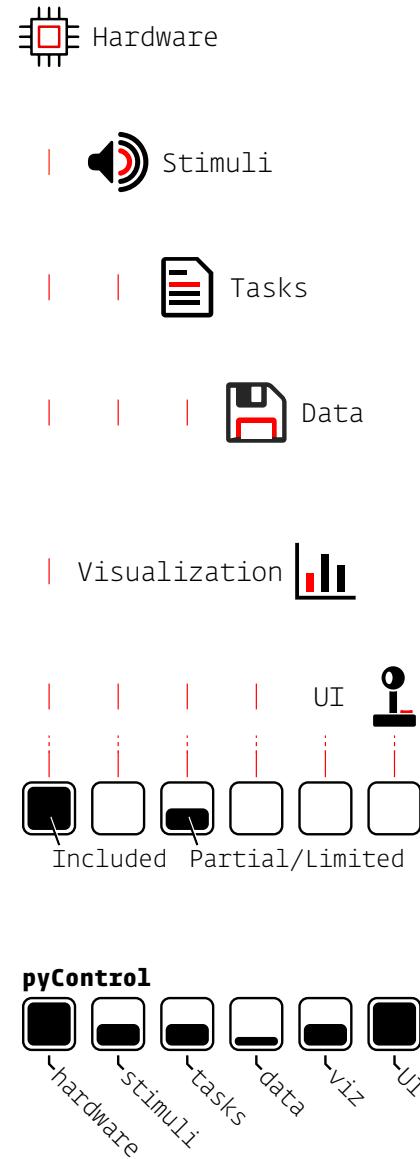
We will briefly describe two other systems that are “complete” as described above: pyControl and Bpod.

pyControl

pyControl[3] is a behavioral framework built in Python by the Champalimaud Foundation. It uses the [micropython microcontroller \(“pyboard”\)](#) as its primary hardware device along with several extension boards [sold by openephys](#). The pyboard has four I/O ports, or eight with a multiplexing expander board. Schematics are available for many other hardware components like solenoid valve drivers and rotary encoders. Multiple pyboards can be connected to a computer via USB and run independent tasks simultaneously with a GUI.

There is limited support for some parametrically defined sound stimuli, presented from a separate amplifier connected using the I2C protocol. Visual stimuli are unsupported.

Like most behavioral software, pyControl uses a finite-state machine formalism to define its tasks. A task is a set of discrete states, each of which has a set of events that transition the task from one state to another. pyControl also allows timed transitions between states, and one function that is called on every event for a rough sort of parallelism. pyControl also allows the use of external variables to control state logic, making these state machines more flexible than strict finite state machines.



```
D 0 2
D 8976 3
D 8976 1
P 8976 Print Statement
D 10162 3
D 10163 2
```

Figure 1.1: pyControl data is stored as plain text, each line having a type (like **Data** or **Print**), timestamp, and state

All events and states are stored alongside timestamps as a plain text log file, one file per subject per session (Figure 1.1). Analog data are stored in a custom binary serialization that alternates 4-byte data and timestamp integers.

There is only one plot type available in the GUI, a raster plot of events, and no facility for varying the plot by task type. The GUI is otherwise quite capable, including the ability to batch run subjects, redefine task variables, and configure hardware.

Bpod

Bpod is primarily a collection of hardware designs and an assembly service run by [Sanworks LLC](#). Similar to pyControl, each Bpod behavior box is based on a finite-state machine microcontroller with four I/O ports. Additional hardware modules provide extended functionality. Bpod is controlled using its own [MATLAB package](#), though there are at least two other third-party software packages, [BControl](#) and [pyBpod](#), that can control Bpod hardware. A task is implemented as a MATLAB script that constructs a new state machine for each trial, uploads it to the Bpod, and waits for the trial to finish. As a result, only one Bpod can be used per host computer, or at least per MATLAB session. Data are stored as trial-split events in a MATLAB structure.

There are a few basic plots for two-alternative forced choice tasks, but there doesn't seem to be a prescribed way to add additional plots. Bpod has a reasonably complete GUI for managing the hardware and running tasks, but it is relatively technical (Figure 1.2).

For brevity we have omitted many other excellent tools that perform some subset of the operations of a complete behavioral system, or are otherwise have a substantial difference in scope.⁵

1.2 Limitations of Existing Systems

We see several limitations with these and other behavioral systems:

- **Hardware** - Both Pycontrol and Bpod strongly encourage users to purchase a limited set of hardware modules and add-ons from their particular hardware ecosystem. If a required part is not available for purchase, neither system provides a clear means of interacting with custom hardware, requiring the user to 'tack on' loosely-integrated components. There is also a hard limit on the *number* of hardware peripherals that can be used in any given task, as there is no ability to use additional pyboards or Bpod state machines. The microcontrollers used in these systems also impose strong limits on their software: neither run a full, high-level programming language⁶. We will discuss this further in section 2.2. A broader limitation of existing systems is the difficulty of flexibly integrating the diverse hardware and analytical tools necessary to perform the next generation of behavioral neuroscience experiments that study "naturalistic, unrestrained, and minimally shaped behavior"[\[14\]](#).
- **Stimuli** - Stimuli are not tightly integrated into either of these systems, requiring

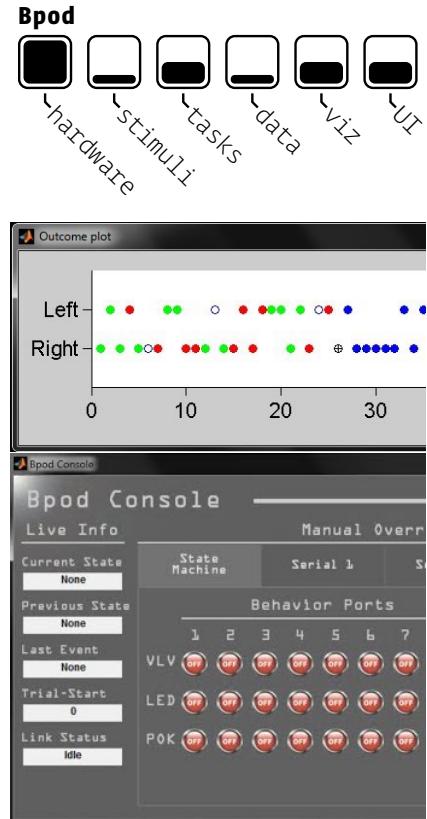


Figure 1.2: A Bpod event plot (above) showing the results of individual behavioral trials, and the Bpod GUI (below).

⁵ Other tools:

- Bonsai[\[30\]](#) - [site](#), [git](#)
- Expyriment[\[28\]](#) - [site](#), [git](#)
- PsychoPy[\[40\]](#) - [site](#), [git](#)
- OpenSesame[\[31\]](#) - [site](#), [git](#)
- SMiLE - [docs](#)
- ArControl[\[12\]](#) - [git](#)
- and see [OpenBehavior](#)

⁶ Bpod runs [custom firmware](#) written in C++ on a [Teensy 3.6](#) microcontroller. pyControl's pyboard implements [micropython](#), a subset of Python that excludes canonical libraries like [numpy](#)[\[52\]](#) or [scipy](#)[\[23\]](#)

the user to write custom routines for their synthesis, presentation, and description in the resulting data. Neither are capable of delivering visual stimuli. Since the publication of the initial version of this manuscript, Bpod has added support for the HiFiberry that we also describe here[44], but the sound generation API appears to be unchanged, with a single method for generating sine waves. Some parametric audio stimuli are included in the pyControl source code but we were unable to find any documentation or examples of their use.

- **Tasks** - Tasks in both systems require a large amount of code and effort duplication. Neither system has a notion of reusable tasks or task ‘templates,’ so every user needs to rewrite every task from scratch. Bpod’s structure in particular tends to encourage users to write long task scripts that are difficult to read (Figure 1.3) because much of its codebase is ‘backend’ code for compiling and communicating with the state machine, so users have to write basic routines like stimulus creation themselves. Another factor that contributes to the difficulty of task design in these systems is the need to work around the limitations of finite state machines, which we discuss further in section 3.3.

- **Data** - Data storage and formatting is basic, requiring extensive additional processing to make it human readable. For example, to determine whether a subject got a trial correct in an example Bpod experiment, one would use the following code:

```
SessionData.RawEvents.Trial{1,1}.States.Punish(1) ~= NaN
```

As a result, data format is idiosyncratic to each user, making data sharing dependent on manual annotation and metadata curation from investigators.

- **Visualization & GUI** - The GUIs of each of these systems are highly technical, and are not designed to be easily used by non-programmers, though pyControl’s documentation offsets much of this difficulty. Visualization of task progress is quite rigid in both systems, either a timeseries of task states or plots specific to two-alternative forced choice tasks. There is no obvious way to adapt plots to specific tasks.

- **Documentation** - Writing good documentation is challenging, but particularly for infrastructural systems where a user is likely to need to modify it to suit their needs it is important that it be possible to understand its lower-level workings. PyControl has relatively good user documentation for how to use the system, but no API-level documentation. Bpod’s documentation is a bit more scattered, and though it does have documentation for a subset of its functions, there is little indication of how they work together or how someone might be able to modify them.

- **Reproducibility** - As of November 2020, pyControl has versioned task files that append a hash to each version of a task and save it along with any produced data. PyControl’s most recent releases have explicit version numbers, but these don’t appear to be saved along with the data. Bpod stores neither code nor task versions in its data. Neither system saves experimental parameter changes by default—and the GUIs of both allow parameters to be changed at will—and so critical data could be lost and experiments made unreplicable unless the user writes custom code to save them. Bpod has an undocumented plugin system, but neither system has a formal system for sharing plugins or task code, requiring work to be duplicated across all users of the system.

```

for currentTrial = 1:MaxTrials
% new state matrix every trial
sma = NewStateMatrix();

% add states and transitions
sma = AddState(sma,
    'Name', 'Wait', ...
    'Timer', 0, ...
    'StateChangeConditions', ...
    {'Port2In', 'Delay'}, ...
    'OutputActions', ...
    {'AudioPlayer1','*'});

% add more states ...

% upload and run task
SendStateMatrix(sma);
RawEvents = RunStateMatrix;

% manually gather data and params
BpodSystem.Data = AddTrialEvents(
    BpodSystem.Data, RawEvents);

% plotting in the main loop
UpdateSideOutcomePlot( ... );
UpdateTotalRewardDisplay( ... );

% manually save data
SaveBpodSessionData;
end

```

Figure 1.3: Bpod’s general task structure.

- **Integration and Extension** - Integration with other systems that might handle some out-of-scope function is tricky in both of these example systems. All systems have some limitation, so care must be taken to provide points by which other systems might interact with them. One particularly potent example is the use of Bpod in the International Brain Laboratory’s standardized experimental rig[29], which relies on a single-purpose [93 page PDF](#) to describe how to use the [iblrig](#) library, which consists of a large amount of single-purpose code for stitching together pybpod with [bonsai](#) for controlling video acquisition. Even if a system takes that amount of additional work to integrate with another, hopefully the system allows it to be done in a way such that it can be reused and shared with others in the future so they can be spared the trouble. The relatively sparse [documentation](#) and the high proportion of ibl-specific code present in the repository make that seem unlikely.

Some of these limitations are cosmetic—fixable with additional code or hardware—but several of the most crucial are intrinsic to the design of these systems.

These systems, among others, have pioneered the development of modern behavioral hardware and software, and are to be commended for being open-source and highly functional. One need look no further for evidence of their usefulness than to their adoption by many labs worldwide. At the time that these systems were developed, a general-purpose single-board computer with performance like the Raspberry Pi 4 was not widely available. The above two systems are not unique in their limitations, but are reflective of broader constraints of developing experimental tools. We are only able to articulate the design principles that differentiate Autopilot by building on their work.

And Autopilot, of course, also has many of its own weaknesses

2

Design

AUTOPILOT DISTRIBUTES EXPERIMENTS across a network of Raspberry Pis,¹ a type of inexpensive single-board computer.

¹ Raspberry Pi model 4B, see [Table 3.2](#)

Autopilot has three primary design principles:

1. **Efficiency** - Autopilot should minimize computational overhead and maximize use of hardware resources.
2. **Flexibility** - Autopilot should be transparent in all its operations so that users can expand it to fit their existing or desired use-cases. Autopilot should provide clear points of modification and expansion to reduce local duplication of labor to compensate for its limitations.
3. **Reproducibility** - Autopilot should maximize system transparency and minimize the potential for the black-box of local reprogramming. Autopilot should maximize the information it stores about its operation as part of normal data collection.

2.1 Efficiency

Though it is a single board, the Raspberry Pi operates more like a computer than a microcontroller. It most commonly runs a custom Linux distribution, Raspbian, allowing Autopilot to use Python across the whole system. Using an interpreted language like Python running on Linux has inherent performance drawbacks compared to compiled languages running on embedded microprocessors. In practice these drawbacks are less profound than they appear on paper: Python's overhead is negligible on modern processors², jitter and performance can be improved by [wrapping compiled code](#), etc. While we view the gain in accessibility and extensibility of a widely used high-level language like Python as outweighing potential performance gains from using a compiled language, Autopilot is nevertheless designed to maximize computational efficiency.

² and improvements to CPython in Python 3.11 and onwards will bring overhead close to zero[4]

Concurrency

Most behavioral software is single-threaded (Figure 2.1), meaning the program will only perform a single operation at a time. If the program is busy or waiting for an input, other operations are blocked until it is finished.

Autopilot distributes computation across multiple processes and threads to take advantage of the Raspberry Pi's four CPU cores. Every object in Autopilot does its work in separate **threads**. Specifically, Autopilot spawns separate threads to process messages and events, an architecture described more fully in [section 3.8](#). Threading does not offer true concurrency³, but does allow Python to distribute computa-

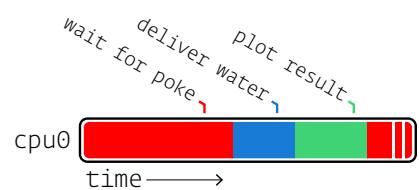


Figure 2.1: A single-threaded program executes all operations sequentially, using a single process and cpu core.

³ See David Beazley's '[Understanding the Global Interpreter Lock](#)' and associated [visualizations](#).

tional time between operations so that, for example, waiting for an event does not block the rest of the program, and events are not missed because the program is busy (Figure 2.2).

Critical operations that are computationally intensive or cannot be interrupted are given their own dedicated **processes**. Linux allows individual cores of a processor to be reserved for single processes, so individual Raspberry Pis are capable of running four truly parallel processing streams. For example, all Raspberry Pis in an Autopilot swarm create a messaging client to handle communication between devices which runs on its own processor core so no messages are missed. Similarly, if an experiment requires sound delivery, a realtime **sound engine** in a separate process (Figure 2.3) also runs on its own core.

Since even moderately complex experiments can consume more resources than are available on a single processor, the topmost layer of concurrency in Autopilot is to use additional **computers**. Autopilot uses the Raspberry Pi as a low-cost hardware controller, but only its GPIO control system is unique to them: the rest of the code can be used on any type of computer, so computationally expensive or GPU-intensive operations can be offloaded to any number of high performance machines. Computers divide labor *autonomously* (see ?? and 3.7), so for example a computer running a task can send and receive messages from one running the GUI and plots, but does not *depend* on that input as it would in a system that couples a microcontroller with a managing computer. The ability to coordinate multiple, autonomous computers with heterogeneous responsibilities and capabilities in a shared task is Autopilot’s definitive design decision.

Leveraging Low-Level Libraries

Autopilot uses Python as a “glue” language, where it wraps and coordinates faster low-level compiled code[53]. Performance-critical components of Autopilot are thin wrappers around fast C libraries (Table 2.1). As Autopilot’s API matures, we intend to replace any performance-limiting Python code like its sound server and networking operations with compiled code exposed to python with tools like the C Foreign Function Interface (CFFI).

Since Autopilot coordinates its low-level components in parallel rather putting everything inside one “main loop,” Autopilot actually has *better* temporal resolution than single-threaded systems like Bpod or pyControl, despite the realtime nature of their dedicated processors (Table 2.2).

Caching

Finite-state machines are only aware of the current state and the events that transition it to future states. They are thus incapable of exploiting the often predictable structure of behavioral tasks to precompute future states and precache stimuli. Further, to change task parameters between trials (eg. changing the rewarded side in a two-alternative forced-choice task), state machines need to be fully reconstructed and reuploaded to the device that runs them each time.

Autopilot precomputes and caches as much as possible. Rather than wait “inside” a state, Autopilot prepares each of the next possible events and saves them for immediate execution when the appropriate trigger is received. Static stimuli are prepared

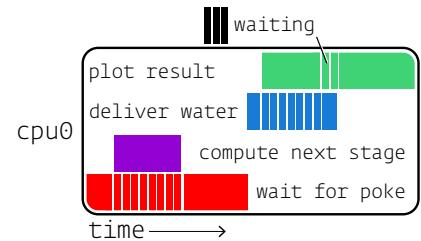


Figure 2.2: A multi-threaded program divides computation time of a single process and cpu core across multiple operations so that, for example, waiting for input doesn’t block other operations.

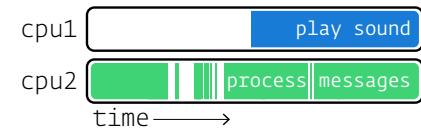


Figure 2.3: A multi-process program is truly concurrent, allowing multiple cpu cores to operate in parallel.

Table 2.1: A few libraries Autopilot uses

jack	realtime audio
pigpio	GPIO control
ZeroMQ	networking
Qt	GUI

Table 2.2: Using pigpio as a dedicated I/O process gives autopilot greater measurement precision

	Precision
Autopilot (pigpio)	$5\mu s$
Bpod	$100\mu s$
pyControl	$1000\mu s$

once at the beginning of a behavioral session and stored in memory. Before their presentation, they are buffered to minimize latency.

By providing full low-level documentation, we let researchers choose the balance between ease of use and performance themselves: it's possible to just call a sound's `play()` method, explicitly buffered with its `buffer()` method, or generated on the fly with its `play_continuous()` method. Similarly, messages can be sent with a networking node's `send()` method, or prepared beforehand by explicitly making a `Message` and calling its `serialize()` method.

Autopilot's efficient design lets it access the best of both worlds—the speed and responsiveness of compiled code on dedicated microprocessors and the accessibility and flexibility of interpreted code.

2.2 Flexibility

Single-language

Behavior software that uses dedicated microprocessors must have some routine for compiling the high-level abstraction of the experiment into machine code. This gives those systems a theoretical advantage in processing speed, but the compiler becomes the bottleneck of complexity: only those things that can be compiled can be included in the experiment. This may in part contribute to the ubiquity of state-machine formalisms in behavior software.

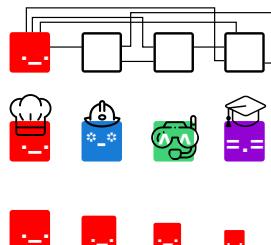
Because Python is used throughout the system, extending Autopilot's functionality is straightforward. Task design (see section 3.3) is effectively arbitrary—anything that can be expressed in Python is a valid task. This also allows Autopilot to easily be extended to make use of external libraries (eg. our integration with DeepLabCut-Live[25] and our `planned` integration with OpenEphys).

Modularity

Although Autopilot deeply integrates with the Raspberry Pi's hardware, we have also worked to make its components modular. There is a tension between providing a full-featured behavioral system and the flexibility of its components — as additional features are added to a system, they can constrain the functionality of existing components that they rely on. To address this tension, we have continuously worked to decouple Autopilot into subcomponents with clear inheritance hierarchies and APIs that can be used quasi-independently.

Modularity has 3 primary advantages:

1. **Modularity makes code more flexible** by reducing the constraints imposed by unstructured code interdependencies
2. **Modularity makes code more intelligible** by logically distributing tasks to discrete classes
3. **Modularity reduces effort-duplication** by allowing multiple, similar classes to be created with inheritance rather than copying and pasting.



There is no such thing as “incompatible hardware” with Autopilot because the classes that control hardware are independent from the code that provides other core functionality. In systems without modular design, hardware implementation is spread across the codebase; for example to add a new type of hardware output to a Bpod system, one would need to write new firmware for it in C, modify Bpod’s existing firmware, hunt through the code to modify how states are added and state machines are assembled, add its controls explicitly to the GUI, and so on.

Tasks specify what type of hardware is needed to run them, but are agnostic about the way the hardware is implemented, making their descriptions more portable. Tasks that have the same structure but differ in hardware (eg. a freely moving two-alternative forced choice task in which a mouse visits several IR sensors, or a head-fixed two-alternative forced choice task in which a mouse runs on a wheel to indicate its choice) can be implemented by a trivial subclass that modifies the hardware description rather than completely rewriting the task.

Plugins & Code Transparency

We call Autopilot a software framework because in addition to providing classes and methods to run experiments out of the box, it also provides explicit structure that scaffolds any additional code that is needed by the user. Our goal is to clearly articulate in the documentation how modules should interact so that anyone can write code that works on any apparatus.

As groupware intended to be used differently by lab members with different responsibilities, Autopilot is designed for users with a range of programming expertise, from those who only want to interact with a GUI, to those who wish to fundamentally rewrite core operations for their particular experiment. As such, it is extensively documented: this paper provides a high-level introduction to its design and structure, its user guide describes how to use the program and provides examples, and its API-level documentation describes in granular detail how the program actually works⁴. Nothing is “off-limits” to the user—there isn’t any hidden, undocumented hardware code behind the curtain. We want users to be able to understand how and why everything works the way it does so that Autopilot can be adapted and expanded to any use-case.

A broader goal of Autopilot is to build a library of flexible task prototypes that can be tweaked and adapted, hopefully reducing the number of times the wheel is reinvented. We have attempted to nudge users to write reusable tasks by designing Autopilot such that rather than writing tasks as local unstructured scripts, they use its plugin system that scaffolds development by extending any of its basic types. Plugins are registered using a form in the Autopilot Wiki which makes them available to anyone while also embedding them in a semantically annotated information system that allows giving explicit credit to contributors, programmatically linking to any derivative publications that use the plugin, and further documentation of any tasks, hardware, or other extensions included within the plugin. Inheriting from parent classes give plugins structure and a set of basic features⁵ while also being maximally permissive — anything can be overridden and modified.

⁴ The user guide and API documentation are available at docs.auto-pi-lot.com

⁵ Like inheriting from the `GPIO` class gives GPIO plugins a systematic means of interacting with the underlying pigpiod daemon.

Message Handling

Modular software needs a well-defined protocol to communicate between modules, and Autopilot's is heavily influenced by the concurrency philosophy⁶ of ZeroMQ[20]. All communication between computers and modules happens with ZeroMQ messages, and handling those messages is the main way that Autopilot handles events. A key design principle is that Autopilot components should not “share state”—they can communicate, but they are not *dependent* on one another. While this may seem like a trivial detail, having networking and message-handling at its core has three advantages that make Autopilot a fundamental departure from previous behavioral software.

First, new software modules can be added to any system by simply dropping in a standalone networking object. There is no need to dramatically reorganize existing code to make room for new functionality. Instead new modules can receive, process, and send information by just connecting to a parent module in the swarm. For example, each `plot` opens a network connection to stream incoming task data independently from the stream that is saving the data.

Second, Autopilot can be made to interact with other software libraries that use ZeroMQ. For example, The OpenEphys GUI for electrophysiology [can send and receive ZMQ messages](#) to execute actions such as starting or stopping recordings. Interaction with other software is also useful in the case that some expensive computation needs to happen mid-task. For example, one could send frames captured from a video camera on a Raspberry Pi to a GPU computing cluster for tracking the position of the animal. Since ZeroMQ messages are just TCP packets it is also possible to communicate over the internet for remote control or to communicate with a data server.

Third, making every component network-capable allows tasks to be distributed over multiple Raspberry Pis. Chaining multiple Pis distributes the computational load, allowing, for example, one Raspberry Pi to record and process video while another runs a sound server and delivers rewards. Autopilot expands with the complexity of your task, simultaneously eliminating limitations on quantity of hardware peripherals while ensuring latency is minimal. More interestingly, distributing tasks allows the arbitrary construction of what we call “behavioral topologies,” which we describe in [section 3.7](#).

2.3 Reproducibility

We take a broad view on reproducibility: including not only the ability to share data and recreate experiments, but also integrating into a broader ecosystem of tools that reduces labor duplication and encourages sharing and organizing technical knowledge. For us, reproducibility means building a set of tools that make every experiment and every technique available to anyone, anywhere.

Standardized task descriptions

The implementation and fine details of a behavioral experiment matter. Seemingly trivial details like milliseconds of delay between trial phases and microliters of reward volume can be the difference between a successful and unsuccessful task (Fig-

⁶ “ZeroMQ [...] has a subversive effect on how you develop network-capable applications. [...] message processing rapidly becomes the central loop, and your application soon breaks down into a set of message processing tasks.”

“If there’s one lesson we’ve learned from 30+ years of concurrent programming, it is: *just don’t share state.*”

-The ZeroMQ Guide

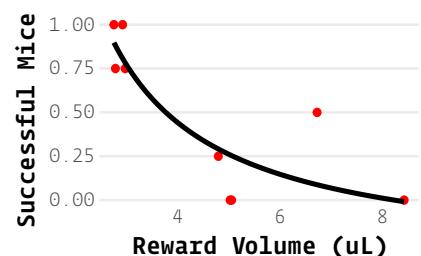


Figure 2.4: “Minor” details have major effects. Proportion of mice (each point, n=4) that were successful learning the first stage of the speech task described in [45] across 10 behavior boxes with variable reward sizes. A 2 μ L difference in reward size had a surprisingly large effect on success rate.

ure 2.4). *Reporting* those details can thus be the difference between a reproducible and unreproducible result. Researchers also often use “auxiliary” logic in tasks—such as methods for correcting response bias—that are never completely neutral to the interpretation of results. These too can be easily omitted due to brevity or memory in plain-English descriptions of a task, such as those found in Methods sections. Even if all details of an experiment were faithfully reported, the balkanization of behavioral software into systems peculiar to each lab (or even to individuals within a lab) makes actually performing a replication of a behavior result expensive and technically challenging. Widespread use of experimental tools that are not explicitly designed to preserve every detail of their operation presents a formidable barrier to rigorous and reproducible science[54].

Autopilot splits experiments into a) the **code** that runs the experiment, which is intended to be standardized and shared across implementations, and b) the **parameters** (Figure 2.5) that define your particular experiment and system configuration. For example, two-alternative forced choice tasks have a shared structure regardless of the stimulus modality, but only your task plays pitch-shifted national anthems. This division of labor, combined with Autopilot’s structured plugin system, help avoid the ubiquitous problem of rig-specific code and hard-coded variables making experimental code only useful on the single rig it was designed for: enabling the possibility of a shared library of tasks as described in section 2.2.

The practice of reporting exactly the parameter description used by the software to run the experiment removes any chance for incompleteness in reporting. Because all task parameters are included in the produced data files, tasks are fully portable and can be reimplemented exactly by anyone that has comparable hardware to yours.

Self-Documenting Data

A major goal of the open science movement is to normalize publishing well-documented and clearly-formatted data alongside every paper. Typically, data are acquired and stored in formats that are lab-idiomatic or ad-hoc, which, over time, sprout entire software libraries needed just to clean and analyze it. Idiosyncratic data formats hinder collaboration within and between labs as the same cleaning and analysis operations gain multiple, mutually incompatible implementations, duplicating labor and multiplying opportunities for difficult to diagnose bugs. Over time these data formats and their associated analysis libraries can mutate and become incompatible with prior versions, rendering years of work inaccessible or uninterpretable. In one worst-case scenario, the cleaning process unearths some critically missing information about the experiment, requiring awkward caveats in the Methods section or months of extra work redoing it. In another, the missing information or bugs in analysis code are never discovered, polluting scientific literature with inaccuracies.

The best way to make data publishable is to avoid cleaning data altogether and *design good data hygiene practices into the data acquisition process*. Autopilot automatically stores all the information required to fully reconstruct an experiment, including any changes in task parameters or code version that happen throughout training as the task is refined.

Autopilot data is stored in **HDF5** files, a hierarchical, high-performance file format. HDF5 files support metadata throughout the file hierarchy, allowing annotations to natively accompany data. Because HDF5 files can store nearly all commonly used

```
{
  "step_name" : "tone_discrim",
  "task_type" : "2AFC",
  "bias_mode" : 0,
  "punish_sound" : false,
  "stim" : {
    "sounds" : {
      "L" : {
        "duration" : 100,
        "frequency" : 10000,
        "type" : "tone",
        "amplitude" : 0.01
      },
      "R" : {"...": "..."}}
    },
  "reward" : {
    "type" : "volume",
    "volume" : 20},
  "graduation" : {
    "type" : "accuracy",
    "threshold" : 0.75,
    "window" : 400},
```

Figure 2.5: Task parameters are stored as portable JSON, formatting has been abbreviated for clarity.

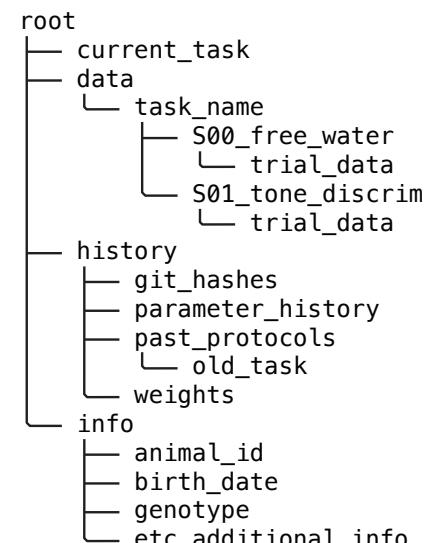


Figure 2.6: Example data structure. All information necessary to reconstruct an experiment is automatically stored in a

data types, data from all collection modalities—trialwise behavioral data, continuous electrophysiological data, imaging data, etc.—can be stored together from the time of its acquisition. Data is always stored with the full conditions of its collection, and is ready to analyze and publish immediately (Figure 2.6). No Autopilot-specific scripts are needed to import data into your analysis tool of choice—anything that can read HDF5 files can read Autopilot data⁷.

As of v0.5.0, we have built a formal data modeling system into Autopilot, allowing for unified declaration of data for experimental subjects, task parameters, and resulting data with verifiable typing and human-readable annotations. These abstract data models can be used with multiple storage interfaces, allowing export to, for example, the Neurodata Without Borders standard[42], further enabling Autopilot data to be immediately incorporated into existing processing pipelines (see section 3.2).

⁷ Though our `Subject` class provides a simplified interface to access and manipulate Autopilot data

Testing & Continuous Integration

Open-source scientific software explodes prior limitations to access and inspection imposed by proprietary tools. It also exposes the research process to bugs in software written by semi-amateurs that can yield errors in the resulting data, analysis, and interpretation[49, 15, 8, 34]. Autopilot tries to bring best practices in software development to experimental software, including a set of automated tests for continuous integration.

We are still formalizing our contribution process, and our tests are still far from achieving full coverage⁸, but we currently require tests and documentation for all new code added to the library. Writing good tests is hard, and we are in the process of building a set of hardware simulators and test fixtures to ease contribution.

Tests are effectively provable statements about how a program functions (Figure 2.7), which are particularly important for a library that aspires to be baseline lab infrastructure like Autopilot. Tests make it possible to use and contribute to the library with confidence: all tests are run on every commit, making it possible to determine if some new contribution breaks existing code without manually reading and testing every line. As we work to complete our test coverage, we hope to provide researchers with a tool that they can trust and elevates the verifiability of scientific results at large.

Expense

Autopilot is an order of magnitude less expensive than comparable behavioral systems (Table 2.3). We think the expense of a system is important for two reasons: scientific equity and statistical power.

The distribution of scientific funding is highly skewed, with a large proportion of research funding concentrated in relatively few labs[26]. Lower research costs benefit all scientists, but lower instrumentation costs directly increase the accessibility of state-of-the-art experiments to labs with less funding. Since well-funded labs also tend to be concentrated at a few (well-funded) institutions, lower research costs also broaden the base of scientists outside traditional research institutions that can stay at the cutting edge[5, 13, 39].

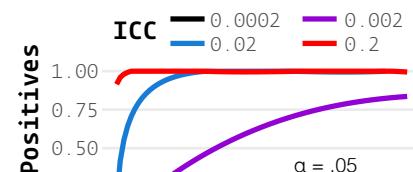
⁸ Coverage statistics for Autopilot are available on coveralls.io at <https://coveralls.io/github/wehr-lab/autopilot>

```
def test_set_gpio():
    """
    The `set` method of a Digital_Out
    object sets the pin state
    """
    pin = Digital_Out(pin=17)

    # Turn GPIO pin on
    pin.set(True)
    assert pin.state == True

    # Turn GPIO pin off
    pin.set(False)
    assert pin.state == False
```

Figure 2.7: A test like `test_set_gpio` is a provable statement about the functionality of a program, in this case that “the `Digital_Out.set()` method sets the state of a GPIO pin.”



Neuroscience also stands to benefit from the lessons learned from the replication crisis in Psychology[46]. In neuroscience, underpowered experiments are the rule, rather than the exception[10]. Statistical power in neuroscience is arguably even worse than it appears, because large numbers of observations (eg. neural recordings) from a small number of animals are typically pooled, ignoring the nested structure of observations collected within individual animals. Increasing the number of cells recorded from a small number of animals dramatically increases the likelihood of Type I errors (Figure 2.8)—indeed, for values of within-animal correlation typical of neuroscientific data, high numbers of observations make Type I errors more likely than not[1]. For this reason, perhaps paradoxically, recent technical advances in multiphoton imaging and silicon-probe recordings will actually make statistical rigor in neuroscience *worse* if we don't use analyses that account for the multilevel structure of the data and correspondingly record from the increased number of animals that they require.

Although the expense of multi-photon imaging and high-density electrophysiology will always impose an experimental bottleneck, behavioral training time is often the greater determinant of study sample size. Typical behavioral experiments require daily training sessions often carried out over weeks and months, while far fewer imaging or electrophysiology sessions are carried out per animal. Training large cohorts of animals in parallel is thus the necessary basis of a well-powered imaging or electrophysiology experiment.

	Autopilot	pyControl	Bpod
Behavior CPU	\$45	\$270	\$925
Nosepoke (3x)	\$216	\$369	\$810
Total for One	\$261	\$639	\$1735
Five Systems	\$1305	\$3195	\$8675
Host CPU(s)	\$1000	\$1000	\$5000
Total for Five	\$2305	\$4195	\$13625
Total for Ten	\$3610	\$8390	\$27350

Table 2.3: Cost for Basic 2AFC System

“Nosepoke” includes a solenoid valve, IR sensor, water tube, LED, housing, and any necessary driver PCBs. For PyControl and Autopilot, we included the cost of one [Lee LHDA0531115H](#) solenoid valve per nosepoke (\$63.35). For PyControl, we estimated a typical USB hub with 5 ports to control 5 pyControl systems from one computer. We note that the Bpod and PyControl systems both include cost of assembly for the control CPUs and nosepokes, but also that Autopilot does not require assembly for its control CPU and its default nosepoke is a snap-together 3D printed part and PCB without surface mounted components that can be assembled by an amateur in roughly half an hour.

3

Program Structure

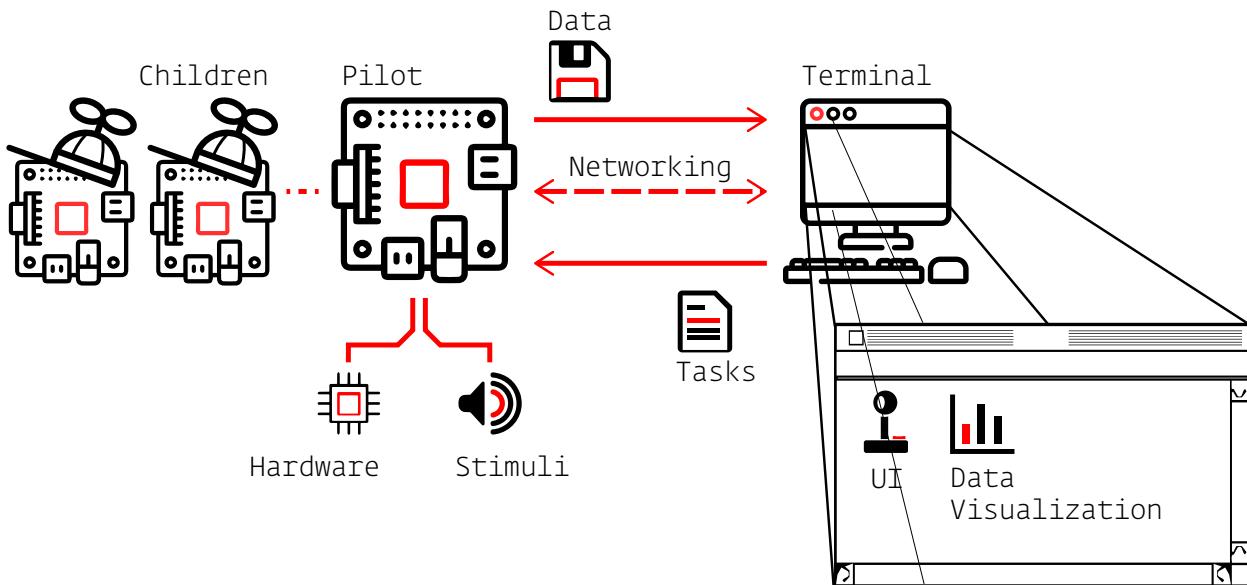


Figure 3.1: Overview of major Autopilot components

AUTOPILOT CONSISTS OF SOFTWARE AND HARDWARE MODULES that are configured to create a behavioral topology. Independent agents linked by flexible networking objects fill different roles within a topology, such as hosting the user interface, controlling hardware, or delivering stimuli. This infrastructure is ultimately organized to perform a behavioral task.

3.1 Directory Structure

On setup, Autopilot creates a user directory that contains all local files that define its operation (Figure 3.2). The subdirectories include:

- **calibration** — Calibration for hardware objects like audio or solenoids that, for example, map opening durations to volumes of liquids dispensed
- **data** — Data for experimental subjects
- **launch_autopilot.sh** — Launch script that includes launching external processes like the jack audio daemon (will be removed and integrated into a more formal agent structure in future versions)
- **logs** — Every Autopilot object is capable of full debug logging, neatly formatted by object type and instance ID and grouped within module-level logging files.

```
./autopilot
├── calibration
├── data
│   ├── subject_1.h5
│   └── subject_2.h5
├── launch_autopilot.sh
└── logs
    ├── core.terminal.log
    │   └── plugins.my_plugin.log
    ├── pilot_db.json
    └── plugins
        └── my_plugin
            └── my_task.py
├── prefs.json
└── protocols
    ├── 2afc_easy.json
    └── 2afc_hard.json
└── sounds
```

Figure 3.2: Example user directory structure, typically in `~/autopilot`.

Logs are both written to disk, and output to `stderr` using the [rich logging handler](#)

for clean and readable inspection during program operation (Figure 3.3). Logs can be parsed back into python objects to make it straightforward to diagnose problems or recover data in the case of an error.

```
[22-03-17T20:10:39] INFO [core.terminal] parent, module-level logger created: core.terminal loggers.py:159
[22-03-17T20:10:39] INFO [core.terminal.Terminal] Logger created: core.terminal.Terminal loggers.py:163
[22-03-17T20:10:39] INFO [core.terminal.Terminal] successfully loaded pilot_db.json file from terminal.py:437
[22-03-17T20:10:39] /Users/jonny/autopilot/pilot_db.json
[22-03-17T20:10:39] DEBUG [core.terminal.Terminal] OrderedDict([('added_pilot', OrderedDict([('subjects', ['myid']), ('ip', ''), ('prefs', OrderedDict())])), ('platform', OrderedDict([('subjects', ['test_nafc']), ('ip', '192.168.0.101'), ('prefs', OrderedDict())]))])
[22-03-17T20:10:39] INFO [core.gui] parent, module-level logger created: core.gui loggers.py:159
[22-03-17T20:10:39] INFO [core.gui.Control_Panel] Logger created: core.gui.Control_Panel loggers.py:163
[22-03-17T20:10:39] INFO [core.plots] parent, module-level logger created: core.plots loggers.py:159
[22-03-17T20:10:39] INFO [core.plots.Plot_Widget] Logger created: core.plots.Plot_Widget loggers.py:163
[22-03-17T20:10:39] INFO [core.plots.Plot] Logger created: core.plots.Plot loggers.py:163
```

Figure 3.3: Logs printed to `stderr` are formatted and colorized by the [rich logging handler](#)

- **pilot_db.json** — A .json file that stores information about associated Pilots, including the contents of their prefs files, which hash/version of Autopilot they are running, and any Subjects that are associated with them.
- **plugins** — Plugins, which are any Python files that contain subclasses of Autopilot objects, that are automatically made available by Autopilot’s [registry](#) system (eg. `autopilot.get('hardware', 'My_Hardware')` would retrieve a custom hardware object). Plugins can be documented and made available to other Autopilot users by registering them on the [wiki](#)
- **prefs.json** — Configuration options for this particular Autopilot instance, including configurations of local hardware objects, audio output, etc. In the future this will likely be broken into multiple files for different kinds of preferences¹.
- **protocols** — [Protocols](#), which consist of parameterizations of individual Tasks as well as criteria for graduating between them. These are also stored in individual subject data files, and updated whenever the source protocol files change.
- **sounds** — Any sound files that are requested by the [File](#) sound class.

¹ with care for backwards compatibility

3.2 Data

As of v0.5.0, Autopilot uses [pydantic](#) to create explicitly typed and schematized data models. Submodules include data abstract modeling tools that define base model types like Tables, Groups, and sets of Attributes. These base modeling classes are then built into a few core data models like subject Biography information, Protocol declaration, and the Subject data model itself that combines them. Modeling classes then have multiple interfaces that can be used to create equivalent objects in other formats, like pytables for hdf5 storage, pandas dataframes for analysis, or exported to Neurodata Without Borders.

For example, consider a simplified version of the Biography model:

```

1   _____ data - Biography _____
2   from autopilot.data.modeling import Data, Attributes
3   from typing import Optional, Union
4
5   class Enclosure(Data):
6       """Where does the subject live?"""
7       box: Optional[Union[str, int]] = Field(
8           default=None,
9           description="The box this Subject lives in"
10      )
11      room: Optional[Union[str, int]] = Field(
12          default=None,
13          description="The room number that the animal is housed in"
14      )
15
16   class Biography(Attributes):
17       """Biography of an Experimental Subject"""
18       id: str = Field(...,
19                      description="The identifying name of this subject.")
20
21       dob: datetime = Field(...,
22                             description="The Subject's date of birth")
23
24       enclosure: Optional[Enclosure] = None
25
26   @property
27   def age(self) -> timedelta:
28       """Difference between now and :attr:`.dob`"""
29       return datetime.now() - self.dob

```

A new subject could then be created with this biography like this, storing it in the HDF5 file and returning an exact copy when requested:

```

1   _____ data - New Subject _____
2   from autopilot.data import Subject
3
4   bio = Biography(
5       id="my_subject",
6       dob="2022-01-01T00:00:00",
7       enclosure=Enclosure(box=100, room="Building 200")
8   )
9   sub = Subject.new(bio)
10  assert sub.info == bio

```

The models are declared using a combination of python type hints² and `Field` objects that provide defaults and descriptions. Because these models can be recursive, as in the case of using the `Enclosure` model as a type within the `Biography` model, we can build expressive, flexible, but still strict representations of complex data.

Out of the box, pydantic models can create explicit and interoperable `schemas` in

² Python `type hints` are colon delimited annotations like this: `x:int = 1` that indicate the type (integer, string, etc.) of the variable. Though typically Python does not, Pydantic both validates that a type matches its hint and coerces it to the correct type if possible.

[JSON Schema](#) and [OpenAPI](#) formats, and Autopilot extends them with additional interfaces and representations. Autopilot can create a GUI form for filling in fields for models, for example, to create a new Subject or declare parameters for a task (Figure 3.2). Attribute models that consist of scalar key-value pairs can be reliably stored and retrieved from metadata attribute sets in HDF5 groups, but Autopilot knows that Table models should be created as HDF5 tables as they will have multiple values for each field. An additional Trial_Data class that inherits from Table can be exported to NWB trial data, and the Subject.get_trial_data method uses the model to load trial data and convert it to a correctly typed pandas[32] DataFrame.

Though the data modeling system is new³, we have laid the groundwork for Autopilot’s plugin system to allow researchers to declare custom schema for all data produced by Autopilot, and to preserve both interoperability and reproducibility by combining them with datasets potentially produced by multiple incompatible tools (see 5).

3.3 Tasks

Behavioral experiments in Autopilot are centered around **tasks**. Tasks are Python classes that describe the parameters, coordinate the hardware, and perform the logic of the experiment. Tasks may consist of one or multiple **stages**, completion of which constitutes a **trial** (Figure 3.5). Stages are analogous to states in the finite state machine formalism.

Multiple tasks are combined to make **protocols**, in which animals move between tasks according to “graduation” criteria like accuracy or number of trials. Training an animal to perform a task typically requires some period of shaping where they are familiarized to the apparatus and the structure of the task. For example, to teach animals about the availability of water from “nosepoke” sensors, we typically begin with a “free water” task that simply gives them water for poking their nose in them. Having a structured protocol system prevents shaping from relying on intuition or ad hoc criteria.

Task Components

The following is a basic two-alternative choice (2AFC) task—a sound is played and an animal is rewarded for poking its nose in a designated target nosepoke. While simple, it is included here in full to show how one can program a task, including an explicit data and plotting structure, in roughly 60 lines of generously spaced Python.

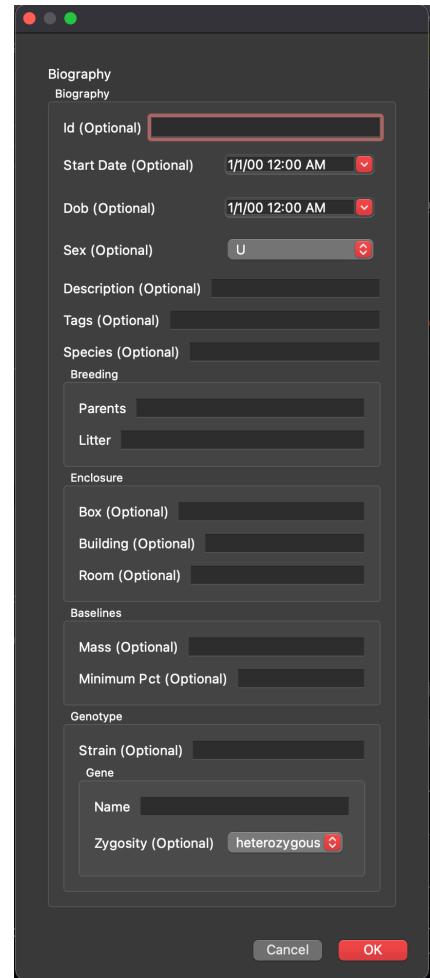


Figure 3.4: An Autopilot Data model can automatically generate a GUI form to fill in its properties, in this example to define a new experimental Subject’s biography.

³ Released as an alpha version at the time of writing

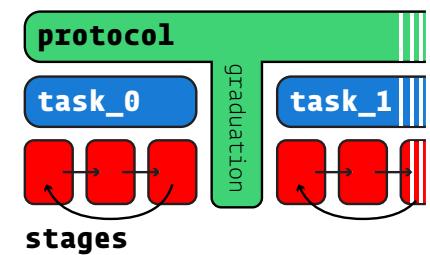


Figure 3.5: Protocols consist of one or multiple tasks, tasks consist of one or multiple stages. Completion of all of a task’s stages constitutes a trial, and meeting some graduation criterion like accuracy progresses a subject between tasks.

Every task begins by describing four elements:

- 1) the task's parameters, 2) the data that will be collected, 3) how to plot the data, and 4) the hardware that is needed to run the task.

```

1      task - parameters
2      class Nafc(Task):
3          class Params(Task_Parms):←
4              stim: Sounds = Field(...,
5                  description = "Sound Stimuli") }
6              reward: units.mL = Field(...,
7                  description↑= "Reward Volume (mL)"
8          )
9
9      class TrialData(Trial_Data): ←
10         target: Side = Field(...,
11             description="Side (L, R) of the correct response")
12         correct: bool = Field(...,
13             description="Response matched target")
14
15     PLOT = {} ←
16     PLOT['data'] = {'target' : 'point',
17                     'correct' : 'rollmean'},
18     # n trials to roll window over
19     PLOT['params'] = {'roll_window' : 50}
20
21     HARDWARE = { ←
22         'POKES':{
23             'L': 'Digital_In',
24             'R': 'Digital_In'
25         },
26         'PORTS':{
27             'C': 'Solenoid', ←
28         }
29     }

```

1) A `Task_Parms` model defines what parameters are needed to run the task.
 We use `Field` objects as before, and can also use some special types like `Sounds` to declare complex parameters
`units` work like numbers but avoid ambiguity, so eg. the `Solenoid` class below knows this is a volume, rather than a duration

2) A `Trial_Data` model defines what data will be returned from the task.

3) A `PLOT` dictionary maps the data output to graphical elements in the GUI. (In future versions this will be incorporated into the `Fields` of `TrialData`)

4) A `HARDWARE` dictionary that describes what hardware will be needed to run the task.

The specific implementation of the hardware (eg. where it is connected, how to interact with it) is independent of the task. The task just knows about a PORT named 'C' that is a `Solenoid`.

Created tasks receive some common methods, like input/trigger handling and networking, from an inherited metaclass. Python inheritance can also be used to make small alterations to existing tasks⁴ rather than rewriting the whole thing. The GUI will use the `Params` model and the `PLOT` dictionary to generate forms for parameterizing the task within a protocol and display the data as it is collected. The `Subject` class will use the `TrialData` model to create HDF5 tables to store the data, and the `Task` metaclass will instantiate the described `HARDWARE` objects from their system-specific configuration in the `prefs.json` file so they are available in the rest of the class like `self.hardware['POKES']['L'].state`

⁴ An example of subclassing a generic 'Task' class is included in Autopilot's [user guide](#)

Stage Methods

The logic of tasks is described in one or a series of methods (stages). The order of stages can be cyclical, as in this example, or can have arbitrary logic governing the transition between stages.

```

task - methods
30 def __init__(self, params:'Nafc.Params'):
31     self.stim_mgr = Stim_Manager(params.stim)
32     self.reward    = Reward_Manager(params.reward)
33
34     stage_list  = [self.discrim, self.reinforcement]
35     self.stages = itertools.cycle(stage_list)
36
37     self.init_hardware()
38     next(self.stages)()
39
40 def discrim(self):
41     target, wrong, stim = self.stim_mgr.next()
42     self.target = target
43
44     self.triggers[target] = [
45         self.hardware['PORTS']['C'].open,
46         lambda: next(self.stages())]
47     self.triggers[wrong] = lambda: next(self.stages())
48
49     self.node.send('DATA', {'target':target})
50
51     stim.play()
52
53 def reinforcement(self, response):
54     if response == self.target:
55         self.node.send('DATA', {'correct':True})
56     else:
57         self.node.send('DATA', {'correct':False})
58
59     next(self.stages())

```

- In Python, `def` defines new methods. The `__init__` method is called when a new object is initialized
- Managers control stimulus and reward delivery, so users can, for example, continually synthesize new stimuli or implement adaptive rewards
- Stages are combined into an object that (in this case) continually cycles through them when its `next()` method is called.
- This starts the task by retrieving the first stage and then calling it.
- The stimulus manager returns which port will be the target and the sound to be played.
- A sequence of triggers is set: if the target port is poked, a reward will be delivered and the next stage will be called. A `Lambda` function indicates not to call the method *now*, but only when triggered.
- The task has a networking object that asynchronously streams data back to the user-facing terminal
- In this example, the response port is passed from the trigger handling function. If it matches the stored target variable, the animal answered correctly.
- Finally, the task is repeated by calling the next stage.

Autopilot is not prescriptive about how tasks are written. The same task could have two separate methods for correct and incorrect answers rather than a single reinforcement method, or only a single stage that blocks the program while it waits for a response.

Publishing data from this task requires no additional effort: a hash that uniquely identifies the code version (as well as any local changes) is automatically stored at the time of collection, as is a JSON-serialized version of the parameter model (Figure 3.6). If this task was incorporated into the central task library, anyone using Autopilot would be able to exactly replicate the experiment from the published data.

```
{
  "step_name": "Simple 2AFC",
  "stim": {
    "sounds": {
      "L": {
        "type": "tone",
        "frequency": 4000
      },
      "R": {
        "type": "tone",
        "frequency": 8000
      }
    },
    "reward": 10
}
```

Figure 3.6: Simplified example of parameters for the above task

The limitations of finite state machines

The 2AFC task described above could be easily implemented in a finite-state machine. However, the difficulty of programming a finite-state machine is subject to combinatoric explosion with more complex tasks. Specifically, finite-state machines can't handle any task that requires any notion of "state history."

As an example, consider a maze-based task. In this task, the animal has to learn a particular route through a maze—it is not enough to reach the endpoint, but the animal has to follow a specific path to reach it (Figure 3.7). The arena is equipped with an actimeter that detects when the animal enters each area.

In Autopilot, we would define a hardware object that logs positions from the actimeter with a `store_position()` method. If the animal has entered the target position ("i" in this example), a `task_trigger()` that advances the task stage is called. The following code is incomplete, but illustrates the principle.

```
1      maze - hardware
2
3      class Actimeter(Hardware):
4          def __init__(self):
5              # ... some code to access the hardware ...
6              self.positions = []
7              self.target_position = "i"
8
9
10         def store_position(self, position):
11             self.positions.append(position)
12
13
14         if position == self.target_position:
15             self.finished_cb(self.positions) ← See line 18 below
16             self.positions = []
```

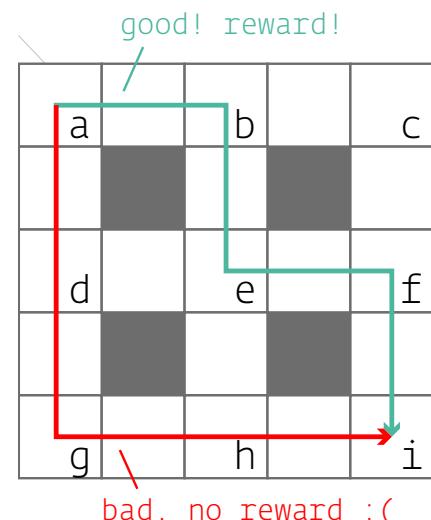


Figure 3.7: The subject must reach point i but only via the correct (green) path.

The task follows, with parameters and network methods for sending data omitted for clarity.

```
1      maze - task
2
3      class Maze(Task):
4          def __init__(self):
5              self.target_path = ['a', 'b', 'e', 'f', 'i']
6
7              self.actimeter = Actimeter()
8              self.actimeter.finished_cb = self.finished ← The actimeter is given a reference to the
9
10         def finished(self, positions):           Maze task's finished() method, which it
11             if positions == self.target_path:    calls when the target position is reached
12                 self.reward()                  The sequence of positions is compared to
13
14
15
16
17
18
19
20
21
22
```

How would such a task be programmed in a finite-state machine formalism? Since the path matters, each "state" needs to consist of the current position and all the positions before it. But, since the animal can double back and have arbitrarily many state transitions before reaching the target corner, this task is impossible to represent with a finite-state machine, as a full representation would necessitate infinitely many

states (this is one example of the *pumping lemma*, see [27]).

Even if we dramatically simplify the task by 1) assuming the animal never turns back and visits a space twice, and 2) only considering paths that are less than or equal to the length of the correct path, the finite state machine would be as complex as figure 3.8.

While finite-state machines are relatively easy to implement and work well for simple tasks, they quickly become an impediment to even moderately complex tasks. Even for 2AFC tasks, many desirable features are difficult to implement with a finite state machine, such as: (1) graduation to a more difficult task depending on performance history, (2) adjusting reward volume based on learning rate, (3) selecting or synthesizing upcoming stimuli based on patterns of errors[6], etc.

Some of these problems are avoidable by using extended versions of finite state machines that allow for extra-state logic, but require additional complexity in the code running the state machines to accomodate, and with enough exceptions the clean systematicity that is the primary benefit of finite state machines is lost. Autopilot attempts to avoid these problems by providing *tools* to program tasks without *requiring a specified format*, balancing the increased complexity by scaffolding the broader ecosystem of the experiment like its output data, hardware control, etc. When possible, we have tried to avoid forcing people to change the way they think about their work to fit our “little universe”⁵ and instead try to provide a set of tools that let researchers decide how they want to use them.

⁵ We take inspiration from Aaron Swartz’ description of another engineering project, the Semantic Web, that became too precious about its formalisms:

“Instead of the “let’s just build something that works” attitude that made the Web (and the Internet) such a roaring success [...] they formed committees to form working groups to write drafts of ontologies that carefully listed (in 100-page Word documents) all possible things in the universe and the various properties they could have, and they spent hours in Talmudic debates over whether a washing machine was a kitchen appliance or a household cleaning device. [...] And instead of spending time building things, they’ve convinced people interested in these ideas that the first thing we need to do is write *standards*. (To engineers, this is absurd from the start—standards are things you write *after* you’ve got something working, not before!)”[50]

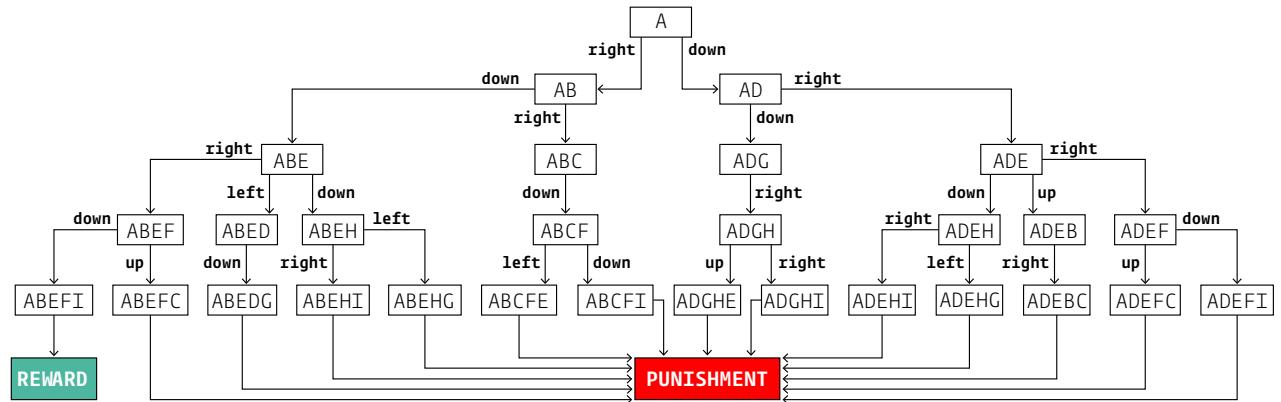


Figure 3.8: State transition tree for a simplified maze task.

3.4 Hardware

The Raspberry Pi can interface with nearly all common hardware, and has an extensive collection of guides, tutorials, and an active forum to support users implementing new hardware. There is also an enormous amount of existing hardware for the Raspberry Pi, including sound cards, motor controllers, sensor arrays, ADC/DACs, and touchscreen displays, largely eliminating the need for a separate ecosystem of purpose-built hardware (Table 3.1).

Autopilot controls hardware with an extensible inheritance hierarchy of Python classes intended to be built into a library of hardware controllers analogously to tasks. Autopilot uses `pigpio` to interact with its GPIO pins, giving Autopilot 5 μ s measurement precision and enabling protocols that require high precision (such as Serial, PWM, and I2C) for nearly all of the pins. Currently, Autopilot also has a family of objects to control cameras (both the [Raspberry Pi Camera](#) and [high-speed GENICAM-compliant cameras](#)), i2c-based [motion](#) and [heat](#) sensors, and [USB mice](#). In the [future](#) we intend to replace the external `pigpio` daemon and other performance-critical hardware operations with low-level interfaces written in Rust.

To organize and make available the vast amount of contextual knowledge needed to build and use experimental hardware, we have made a densely linked and publicly editable semantic wiki. The Autopilot wiki contains, among others, reference information for off-the-shelf parts, schematics for 2D and 3D-printable components, and guides for building experimental apparatuses and custom parts. The wiki combines unrestrictive freeform editing with structured, computer-readable semantic properties, and we have defined a collection of schemas for commonly documented items coupled with submission forms for ease of use. For example, the wiki page for the [Lee Company solenoid](#) we use has fields from a generic Part schema like a datasheet, price, and voltage, but also that it's a 3-way, normally-closed solenoid.

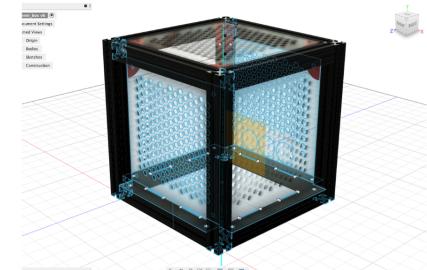
The wiki's blend of structure and freedom breaks apart typically monolithic hardware documentation into a collaborative, multimodal technical knowledge graph. Autopilot can access the wiki through its API, and we intend to tighten their integration over time, including automatic configurations for common parts, usage and longevity benchmarks, detecting mutually incompatible parts, and automatically resolving any additional plugins or dependencies needed to use a part.

	Raspberry Pi 4B	Teensy 3.6	pyboard
CPU Clock	1.5GHz	180MHz	168MHz
CPU Cores	4	1	1
Architecture	ARMv8-A, 64-bit	ARMv7 32-bit	ARMv7 32-bit
RAM Size	2, 4, or 8GB	256KB	192KB
Storage	MicroSD (any size)	1024KB	1024KB
GPU	Broadcom VideoCore VI	—	—
GPIO Pins	40	58	29
USB Ports	2x USB 2.0, 2x USB 3.0	2x USB 2.0	1x USB 2.0
Ethernet	1Gbps	100Mbps	—
WiFi	2.4/5 GHz b/g/n/ac	—	—
Camera	15-pin Serial Interface	—	—
Bluetooth	✓	—	—

Table 3.1: Cost of common peripherals. The native hardware of the Raspberry Pi, low-level hardware control of Autopilot, and availability of inexpensive off-the-shelf components compatible with the raspi make most custom-built peripherals unnecessary.

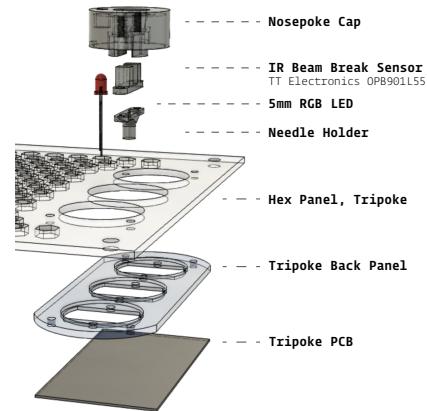
Device	Raspi	Bpod
HiFiBerry DAC2 Pro	\$45	\$445
ADC	\$30	\$495
I2C	\$0	\$225
Ethernet	\$0	\$285
Rotary Encoder	\$0	\$145

Autopilot Behavior Box



A modular box for training mouse-sized animals

Modality	Enclosures
Build Guide Type	Construction Build Guide
Creator	Jonny Saunders
Version	2
Submitted Date	2021-06-10



Autopilot Tripoke

Figure 3.9: Two examples of parts with assembly guides available on the autopilot wiki: A modular behavior box with magnetic snap-in panels (top), and a three-nosepoke panel (bottom).

Table 3.2: Specifications of reviewed behavior hardware. BPod's state machine uses the Teensy 3.6 microcontroller, and PyControl uses the Micropython Pyboard.

3.5 Transforms

In v0.3.0, we introduced the `transform` module, a collection of tools for transforming data. The raw data off a sensor is often not in itself useful for performing an experiment: we want to compare it to some threshold, extract positions of objects in a video feed, and so on. Transforms are like building blocks, each performing some simple operation with a standard object structure, and then composed into a pipeline (Figure 3.10). Pipelines are portable, and can be created on the fly from a JSON representation of their arguments, so it's easy to offload expensive operations to a more capable machine for distributed realtime experimental control (See [25]).

In addition to computing derived values, we use transforms in a few ways, including

- **Bridging Hardware** — Different hardware devices use different data types, units, scales, so transforms can `rescale` and convert values to make them compatible.
- **Integrating External Tools** — The number of exciting analytical tools for realtime experiments keep growing, but in practice they can be hard to use together. The transform module gives a scaffolding for writing wrappers around other tools and exposing them to each other in a shared framework, as we did with DeepLabCut-Live[25], making closed-loop pose tracking available to the rest of Autopilot's ecosystem. We don't need to rally thousands of independent developers to agree to write their tools in a shared library, instead we want to make wrapping them easy.
- **Extending Objects** — Transforms can be used to augment existing and create new objects. For example, a `motion sensor` uses the `spheroid` transform to calibrate its accelerometer, and the `gammatone filter`⁶ extends the `Noise` sound to make a gammatone `filtered noise` sound.

Like Tasks and Hardware, the transform module provides a scaffolding for writing reference implementations of algorithms commonly needed for realtime behavioral experiments. For example, neuroscientists often want to quickly measure a research subject's velocity or orientation, which is possible with inexpensive inertial motion sensors (IMUs), but since anything worth measuring will be swinging the sensor around with wild abandon the readings first need to be rotated back to a geocentric coordinate frame. Since the readings from an accelerometer are noisy, we found a few whitepapers describing using a Kalman filter for fusing the accelerometer and gyroscope data for a more accurate orientation estimate ([2, 38]), but couldn't find an implementation — so we `wrote one`. We integrated it into the IMU class (Figure 3.11) and since it's an independent transform, it's available to anyone even if they use nothing else from Autopilot.

Transforms were made to be composed, so we broke it into independent sub-operations: A `Kalman` filter, `rotation`, and a `spheroid correction` to calibrate accelerometers. Then we combined it with the DLC-Live transform for a fast but accurate motion estimate from position, velocity, and acceleration measurements from three independent sensors. Since each step of the transformation is exposed in a clean API, it was straightforward to `extend the Kalman filter` to accomodate the the wildly different sampling rates of the camera and IMU. It's still got its quirks, but that's the purpose of plugins — to make the code `available and documented` without formally integrating it in the library.



New in
v0.3.0:
Transform

```
____ Transform - DLC Live! ____  
from autopilot import transform as t  
  
# track points on a human body  
dlc = t.image.DLC()  
    model_zoo="full_human")  
# select one of the knees  
dlc += t.selection.DLCSlice(  
    select="knee2")  
# Test if it's in an ROI  
dlc += t.logical.Condition(  
    minimum=(0,0),  
    maximum=(128,128))  
  
# Process frames with the pipeline  
# set pin High if knee2 in ROI  
while True:  
    pin.set(dlc.process(  
        cam.q.get()))
```

Figure 3.10: Transforms can be chained together (here with the in-place addition operator `+=`) to make pipelines that encapsulate the logical relationship between some input and a desired output. Here `pin` is a `Digital_Out` object, and `cam` is a `PiCamera` with queue enabled.

⁶a thin wrapper around `scipy's signal.gammatone`

```
____ Geocentric Velocity ____  
# rotate input in x and y  
# by some pitch and roll  
reorient = t.geometry.Rotate(  
    dims='xy')  
# select the z axis  
reorient += t.selection.Slice(  
    select=2)  
# remove gravity  
reorient += t.math.Add(  
    -9.8)  
  
# using I2C_9DOF ...  
angle = imu.rotation  
z_accel = reorient.process(  
    imu.acceleration, angle)
```

Figure 3.11: Using the `IMU_Orientation` transform built into the IMU's `rotation` property, a processing chain to reorient the accelerometer reading and subtract gravity for geocentric z-axis acceleration.

3.6 Stimuli

A hardware object would control a speaker, whereas stimulus objects are the individual sounds that the speaker would play. Like tasks and hardware, Autopilot makes stimulus generation portable between users, and is released with a family of common sounds like tones, noises, and sounds from files. The logic of sound presentation is contained in an inherited metaclass, so to program a new stimulus a user only needs to describe how to generate it from its parameters (Figure 3.12). Sound stimuli are better developed than visual stimuli in the current version of Autopilot, but we present a proof-of-concept visual experiment (Section 4.5) using `psychopy`[40].

Autopilot controls the realtime audio server `jack` from an independent Python process that dumps samples directly into `jack`'s buffer (Figure 3.13), giving it a trigger-to-playback latency very near the theoretical minimum (Section 4.2). Sounds can be pre-buffered in memory or synthesized on demand to play continuous sounds. Because the realtime server is independent from the logic of sound synthesis and storage, stimuli can be controlled independently from different threads without interrupting audio or dropping frames.

We use the [Hifiberry Amp 2](#), a combined sound card and amplifier, which is capable of 192kHz/24Bit audio playback. Autopilot and Jack can output to any sound hardware, however, including the builtin audio of the Raspberry Pi if fidelity isn't important. There are no external video cards for the Raspberry Pi, but its embedded video card is capable of presenting video and visual stimuli (Section 4.5) especially if the other computationally demanding parts of the task are distributed to other Raspberry Pis (Section 3.7). If greater video performance is needed, Autopilot is capable of running on typical desktops as well as other single-board computers with GPUs (as we did with the Nvidia Jetson in [25]).

3.7 Agents

All of Autopilot's components can be organized into a single system as an “agent,” the executable that coordinates everyday use. An agent encapsulates:

- **Runtime Logic** — an initialization routine that starts any needed system processes and any subsequent operations that define the behavior of the agent.
- **Networking Station** — Agents have networking objects called `Stations` that are intended to be the “load bearing” networking objects (described more [below](#)).
- **Callbacks** — An action vocabulary that maps different types of messages to methods for handling them. Called `listens` to disambiguate from other types of callbacks.
- **Dependencies** — Required packages, libraries, and system reconfigurations needed to operate. Python dependencies are currently defined for agents as groups of [optional packages](#)⁷, and system configuration is done with `scripts` which shorthand common operations like [compiling OpenCV](#) with optimizations for the raspi or [enabling a soundcard](#) (Though see [5.5](#)).

All of which amount to a definition of an agent's *role* in the swarm.

```
____ An Autopilot Tone _____
my_tone = sounds.Tone(
    frequency = 500,
    duration = 200)
my_tone.play()
```

Figure 3.12: Autopilot stimuli are parametrically defined and inherit all the playback logic that makes them easy to integrate in tasks

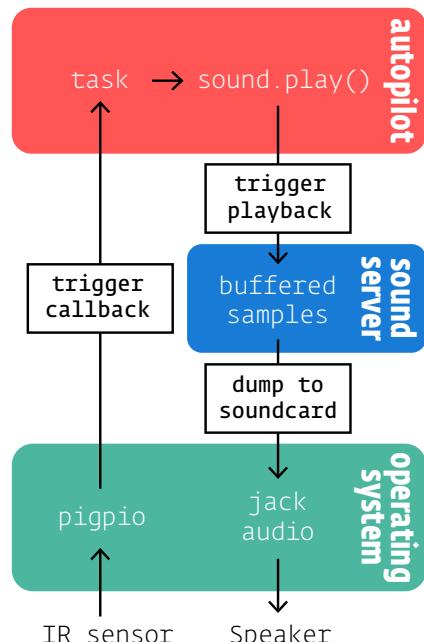


Figure 3.13: Our sound server keeps audio samples buffered until a `.play()` method is called, and then dumps them directly into the `jack` audio daemon.

⁷ As of v0.5.0, Autopilot is packaged with `Poetry`, so they are `[tool.poetry.extras]` entries within the `pyproject.toml` file, installed with pip like `pip install auto-pi-lot[pilot]` or poetry like `poetry install -E pilot`

There are currently two agents in Autopilot:

- **Terminal** - The user-facing control agent.
- **Pilot** - A Raspberry Pi that runs tasks, coordinates hardware, and optionally coordinates a set of child Pis.

Terminal agents serve as a root node (see Section 3.8) in an Autopilot swarm. The terminal is the only agent with a **GUI**, which is used to control its connected pilots and visualize incoming task data. The terminal also manages data and keeps a registry of all active experimental subjects. The terminal is intended to make the day-to-day use of an Autopilot swarm manageable, even for those without programming experience. The terminal GUI is described further in Section 3.9.

Pilot agents are the workhorses of Autopilot—the agents that run the experiments. Pilots are intended to operate as always-on, continuously running system services. Pilots make a network connection to a terminal and wait for further instructions. They maintain the system-level software used for interfacing with the hardware connected to the Raspberry Pi, receive and execute tasks, and continually return data to the terminal for storage.

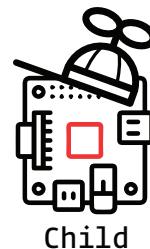
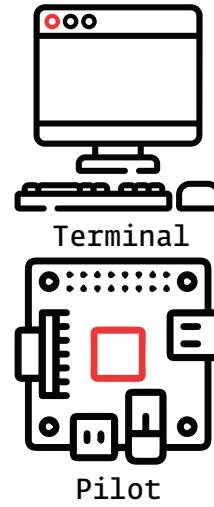
Each agent runs autonomously, and so a Pilot can run a task without a Terminal and store data locally, a Terminal can be used without Pilots to define protocols and manage subjects, and so on. This decoupling lets each agent have more freedom in its behavior at the expense of the complexity of configuring and maintaining them (see 5). All interaction is based on the “listen” callbacks known by the agents, so to start a task a Terminal will send a Pilot a “START” message containing information about a Task class that it is to run along with its parameterization. The Pilot then attempts to run the task, sends a message to the Terminal alerting it to a “STATE” change, and begins streaming data back to it in messages with a “DATA” key.

Each pilot is capable of mutually coordinating with one or many **Copilots**⁸. We are still experimenting with, and thus openminded to the best way to structure multi-pilot tasks. As a sketch, we imagine that one might specify “slots” for agents in a task and configure which agent takes on which slot in the same way that one would configure how hardware is controlled in a single pilot. For now, in the multi-pilot tasks we have seen and programmed, as part of the task initialization method one main Pilot sends a “START” message to the other Pilots, mimicking the way that a Terminal starts a task. A single “Task” then consists of multiple Task classes that define each Agent’s responsibilities. This open-endedness is part of the fun of using Autopilot, though the dizziness of freedom does generate its own kind of anxiety that can be paradoxically constraining⁹ We expand on this further in 5.5, as it is one of the major points of future improvement.

Behavioral topologies

We think one of the most transformative features of Autopilot’s distributed structure is the control that users have over what we call “behavioral topology.” The logic of hardware and task operation within an agent, the distribution of labor between agents performing a task, and the pattern of connectivity and command within a swarm of agents constitute a topology.

Below we illustrate this idea with a few examples:



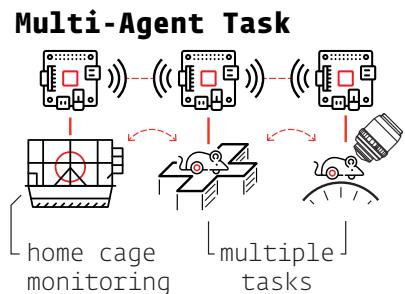
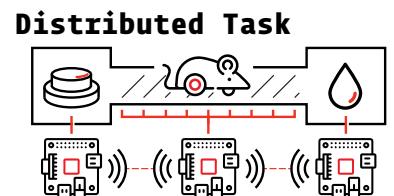
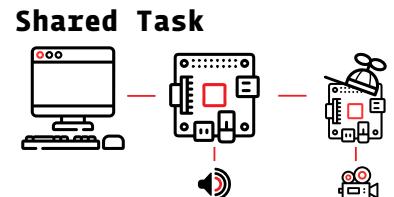
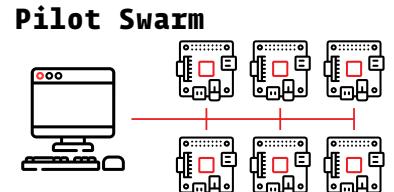
⁸ A previous version described a third, subordinate “Child” agent that performed auxiliary operations in a task. It became clear relatively quickly that such a hierarchy was unnecessary and overcomplicating, and that distribution of labor within a task is better served by a fluid combination of multiple Pilots than designating one of them to be the leader. We now refer one among multiple agents performing a task together as a “copilot.”

⁹ I’d make a very loose analogy to [19], if the reader allows it

- **Pilot Swarm** - The first and most obvious topological departure from traditional behavioral instrumentation is the use of a single computer to independently coordinate tasks in parallel. Our primary installation of Autopilot is a cluster of 10 behavior boxes that can independently run tasks dispatched from a central terminal which manages data and visualization. This topology highlights the expandability of an Autopilot system: adding new pilots is inexpensive, and the single central terminal makes controlling experiments and managing data simple.
- **Shared Task** - Tasks can be shared across a set of copilots to handle tasks with computationally intensive operations. For example, in an open-field navigation task, one pilot can deliver position-dependent sounds while another records and analyzes video of the arena to track the animal's position. The terminal only needs to be configured to connect to the parent pilot, but the other copilot is free to send data to the Terminal marked for storage in the subject's file as well.
- **Distributed Task** - Many pilots with overlapping responsibilities can cooperate to perform distributed tasks. We anticipate this will be useful when the experimental arenas can't be fully contained (such as natural environments), or when experiments require simultaneous input and output from multiple subjects. Distributed tasks can take advantage of the Pi's wireless communication, enabling, for example, experiments that require many networked cameras to observe an area, or experiments that use the Pis themselves as an interface in a multisubject augmented reality experiment.
- **Multi-Agent Task** - Neuroscientific research often consists of multiple mutually interdependent experiments, each with radically different instrumentation. Autopilot provides a framework to unify these experiments by allowing users to rewrite core functionality of the program while maintaining integration between its components. For example, a neuroethologist could build a new “**Observer**” agent that continually monitors an animal's natural behavior in its home cage to calibrate a parameter in a task run by a pilot. If they wanted to manipulate the behavior, they could build a “**Compute**” agent that processes Calcium imaging data taken while the animal performs the task to generate and administer patterns of optogenetic stimulation. Accordingly, passively observed data can be combined with multiple experimental datasets from across the subject's lifespan. We think that unifying diverse experimental data streams with interoperable frameworks is the best way to perform experiments that measure natural behavior in the fullness of its complexity in order to understand the naturally behaving brain[14].

3.8 Networking

Agents use two types of object to communicate with one another: core **station** objects and peripheral **node** objects (Figure 3.14). Each agent creates one station in a separate process that handles all communication *between* agents. Stations are capable of forwarding data and maintaining agent state so the agent process is not unnecessarily interrupted. Nodes are created by individual modules run within an agent—eg. tasks, plots, hardware—that allow them to send and receive messages within an agent, or make connections directly to other nodes on other agents after the station discovers their network addresses. Messages are TCP packets¹⁰, so there



¹⁰ Autopilot uses ZeroMQ[20] and tornado to send and process messages

is no distinction between sending messages within a computer, a local network, or over the internet¹¹.

Both types of networking objects are tailored to their hosts by a set of callback functions — **listens** — that define how to handle each type of message. Messages have a uniform key-value structure, where the key indicates the listen used to process the message and the value is the message payload. This system makes adding new network-enabled components trivial:

```
1   A new networked LED
2
3   class LED_RGB(Hardware):
4       def __init__(self):
5           # call self.color for a 'COLOR' message
6           self.listens = {'COLOR': self.color}
7           self.node = networking.Node(
8               id      = 'BEST_LED',
9               listens = self.listens)
10
11
12   def color(msg):
13       self.set_color(msg.value)
14
15
16   # elsewhere in the code, we change the color to red!
17   node.send(to='BEST_LED', key='COLOR', value=[255, 0, 0])
```

Messages are serialized¹² with JSON¹³, and can handle arrays, including on-the-fly compression with `blosc`. Net Nodes can create additional sockets to stream data that is stashed in a `queue`, and can take advantage of message batching and compressing multiple arrays together when latency is less critical.

Network connectivity is currently treelike by default (Figure 3.15) — each independent networking object can have many children but at most one parent. This structure makes an implicit assumption about the anisotropy of information flow: ‘higher’ nodes don’t need to send messages to the ‘lowest’ nodes, and the ‘lowest’ nodes send all their messages to one or a few ‘higher’ nodes. It enforces simplified delegation of responsibilities in both directions: a terminal shouldn’t need to know about every hardware object connected to all of its connected pilots, it just sends messages to the pilots, who handle it from there. A far-downstream node shouldn’t need to know exactly how to send its data back to the terminal, so it pushes it upstream until it reaches a node that does.

In the time since we released Autopilot it became clear that this treelike structure is useful for getting started quickly with the default full system configuration, but an unnecessary encumbrance for experimenting with different configurations. Alongside the **formalization** of the Agent classes, we intend to replace it with a peer-to-peer addressing system where every Autopilot has a unique ID and can message any other object in a given swarm with an “agent:id” address scheme. See 5.9 for further discussion.

¹¹ Though automatically configuring the use of faster protocols like IPC for communication within an agent is part of our **development goals**

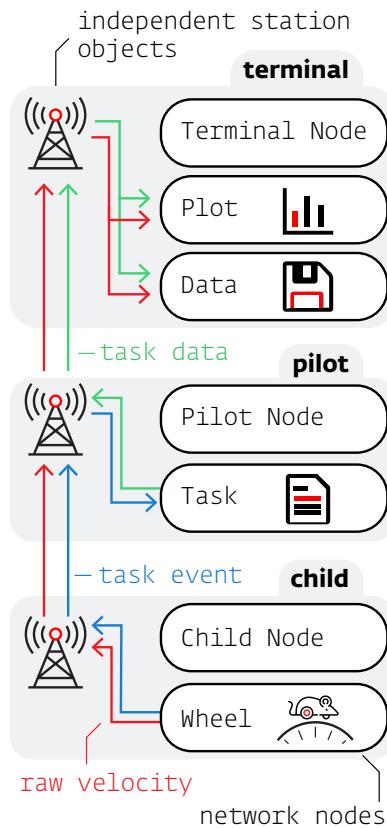


Figure 3.14: Autopilot segregates data streams efficiently—eg. raw velocity (red) can be plotted and saved by the terminal while only the task-relevant events (blue) are sent to the pilot. The pilot then sends trial-summarized data to the terminal (green).

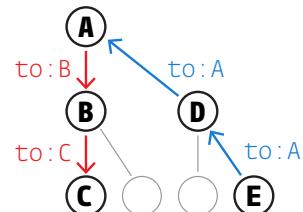


Figure 3.15: Treelike network structure—downstream messages are addressed by successive nodes, but upstream messages can always be pushed until the target is found.

¹² converted to binary suitable for sending between computers

¹³ though we have been investigating faster serialization libraries like `msgpack`

3.9 GUI & Plots

The terminal's GUI controls day-to-day system operation¹⁴. It is intended to be a nontechnical frontend that can be used by those without programming experience.

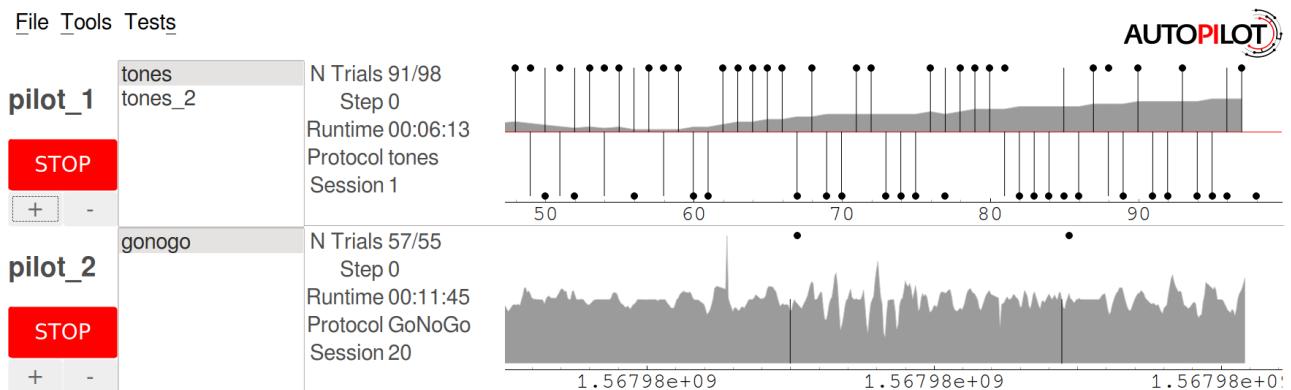
For each pilot, the terminal creates a control panel that manages subjects, task operation, and plots incoming data. Subjects can be managed through the GUI, including creation, protocol assignment, and metadata editing. Protocols can also be created from within the GUI. The GUI also has a set of basic maintenance and informational routines in its menus, like calibrating water ports or viewing a history of subject weights.

The simple callback design and network infrastructure makes adding new GUI functionality straightforward, and in the future we intend to extend the plugin system such that plugins can provide additional menu actions, plots, and utilities.

Plotting

Realtime data visualization is critical for monitoring training progress and ensuring that the task is working correctly, but each task has different requirements for visualization. A task that has a subject continuously running on a ball requires a continuous readout of running velocity, whereas a trial-based task only needs to show correct/incorrect responses as they happen. Autopilot approaches this problem by assigning the data returned by the task to graphical primitives like points, lines, or shaded areas as specified in a task's PLOT dictionary (taking inspiration from Wilkinson's grammar of graphics[56]).

The GUI is now some of the oldest code in the library, and we are in the process of decoupling some of its functionality from its visual representation and moving to a model where it is a thinner wrapper around the **data modeling tools**. Following the lead of formal models with strict typing will, for example, make plotting more fluid where the researcher can map incoming data to the set of graphical elements that are appropriate for its type. We discuss this further in 5.13



¹⁴ Autopilot uses PySide, a wrapper around Qt, to build its GUI.

Trial Plot

```
{"data": {
    "target" : "point",
    "response" : "segment",
    "correct" : "rollmean"
},
"roll_window" : 50}
```

Continuous Plot

```
{"data": {
    "target" : "point",
    "response" : "segment",
    "velocity" : "shaded"
},
"continuous": true}
```

Figure 3.16: PLOT parameters for Figure 3.17. In both, “target” and “response” data are mapped to “point” and “segment” graphical primitives, but timestamps rather than trial numbers are used for the x-axis in the “continuous” plot (Figure 3.17, bottom). Additional parameters can be specified, eg. the trial plot (Figure 3.17, top) computes rolling accuracy over the past 50 trials

Figure 3.17: Screenshot from a terminal GUI running two different tasks with different plots concurrently. pilot_1 runs 2 subjects: (tones and tones_2). See Figure 3.16 for plot description

4

Tests

WE HAVE BEEN TESTING AND REFINING AUTOPILOT since we built our swarm of 10 training boxes in the spring of 2019. In that time 178 mice¹ have performed over 6 million trials on a range of tasks. While Autopilot is still relatively new, it is by no means untested.

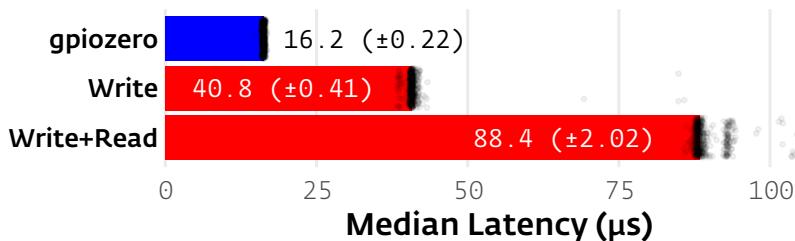
In this section we will present a set of basic performance benchmarks while also showing several of the different ways that Autopilot can be used. The code for all of the following tests is available as a [plugin](#) that is further documented on the [wiki](#), and runs on a prerelease of v0.5.0 (specific commit hashes will be referenced where relevant).

4.1 GPIO Latency

Neurons compute at millisecond timescales, so any task that links neural computation to behavior needs to have near-millisecond latency. We start by characterizing Autopilot’s GPIO control latency in “script mode” — using the GPIO control classes on their own, without using any of the rest of Autopilot’s modules.

Output Latency

We first tested the software measured latency between when a command to write a value to a GPIO pin is issued and when it completes (Figure 4.1). By default, the pigpio interface we use to control GPIO pins issues a command and then confirms the request was successful by querying the pigpio daemon for the status of the pin (Write+Read). We [extended](#) pigpio to just issue the command without confirmation to estimate the true time between when the command is issued and when the voltage of the pin changes (Write). Each of these operations takes roughly $40\mu s$ with minimal jitter (Median \pm IQR — Write Only: $40.8\mu s \pm 0.41$, Write and Read: $88.4\mu s \pm 2.02$, n=100,000 each).



Pigpio is useful as a general purpose controller because of its ability to run scripts within its daemon, use hardware PWM via direct memory access, and consistently poll for pin state, but takes a latency penalty because the python interface communicates with it through a local TCP socket. To demonstrate the flexibility of Autopilot

¹ All procedures were performed in accordance with National Institutes of Health guidelines, as approved by the University of Oregon Institutional Animal Care and Use Committee.

Table 4.1: Latency Test Materials

Hardware	Raspberry Pi 4
Soundcard	Hifiberry Amp2
IR Break Sensor	TT Electronics OPB901L55
Speaker	HiVi RT1.3WE
Oscilloscope	Tektronix TDS 2004B
Router	
Python	version
Software	RaspiOS version
Analysis	
R	(version)
ggplot2	(version)
dplyr	(version)

Table 4.2: GPIO Latency Materials

Autopilot	hash
Plugin	hash
Code	test_gpio.py

Figure 4.1: Software latency from GPIO write to completion of command. Values are presented as medians \pm IQR with n=100,000 tests for each. A random subsample of 500 (for tractability of plotting) of each type of test are presented (black points) after filtering to the bottom 99th percentile to exclude extreme outliers. Commands sent using pigpio (red) took roughly $40\mu s$ each (write and read are effectively two separate commands, Write Only: $40.8\mu s \pm 0.41$, Write and Read: $88.4\mu s \pm 2.02$). The prototype gpizero wrapper (blue) using RPi.GPIO as its backend was faster, taking $16.2\mu s \pm 0.22$ to complete.

in incorporating additional software libraries, we wrote a [thin wrapper](#) around [gpi-ozero](#), which can use [RPi.GPIO](#) to directly write to the GPIO registers. For simple output, this wrapper proved to be faster ($16.2 \mu s \pm 0.22$, n=100,000), and with 63 lines of code is now available in the plugin accompanying this paper to be used, repurposed, and extended.

Roundtrip Input/Output Latency

Output commands usually aren't issued in isolation, but as a response to some external or task-driven trigger. We measured the roundtrip latency from a 5V square pulse from an external function generator to when an output pin was flipped from low to high on an oscilloscope.

Typically that is as much methodological detail as you would expect in a scientific paper, but actually making those measurements via oscilloscope requires knowing how to set up such a test as well as how to extract the measured data afterwards — which is not altogether trivially available technical knowledge. As an example of how integrating semantically linked documentation with experimental tools enables a fundamentally deeper kind of reproducibility and methodological transparency, we instead documented these operations, including a code sample and a guide to unlocking additional features on our oscilloscope on the [autopilot wiki](#). The code to extract traces from the oscilloscope is also included in this paper's [plugin](#), which links to the oscilloscope page with a [[Controls Hardware :: Rigol DS1054Z]] tag, so it is possible to bidirectionally find code examples from the oscilloscope page as well as find further documentation about the hardware used in this paper from the plugin page. The same can be true for any hardware used by any plugin in any paper using Autopilot.

We used the `assign_cb` method of the [Digital_In](#) class to test the typical roundtrip latency that Autopilot objects can deliver. This gave us a median $474 \mu s$ (IQR: $52.5 \mu s$) latency (Red bar in Figure 4.3). GPIO callbacks are flexible, and can use arbitrary python functions, but if all that's needed is to trigger one pin off of another with some simple logic like a parametric digital waveform or static "on" time, pigpio also allows us to directly program pin to pin logic as a "pigs" script (literally Figure 4.2) that runs within the pigpio daemon. The pigs script gave us roughly three orders of magnitude lower latency (Median \pm IQR: $380\text{ns} \pm 140$, blue bar in 4.3).

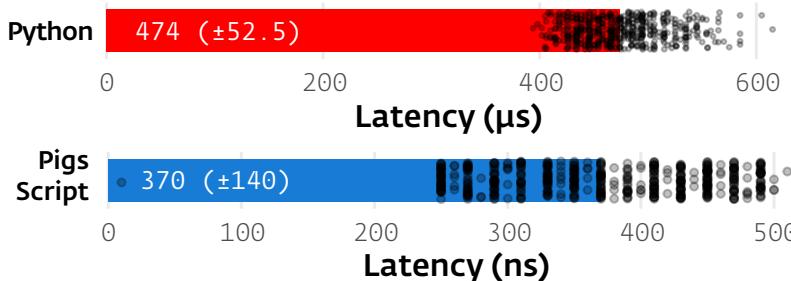


Table 4.3: GPIO Latency Materials

Function Generator	hash
Oscilloscope	hash

```
____ Pigs Trigger Script ____
" ".join([
    "tag 999",
    f'r {pin_in.pin_bcm}',
    f"jz 998", # if off, goto 998
    # else, turn on
    f"w {pin_out.pin_bcm} 1",
    "jp 999", # then goto 999
    "tag 998",
    f"w {pin_out.pin_bcm} 0",
    "jp 999"
])
```

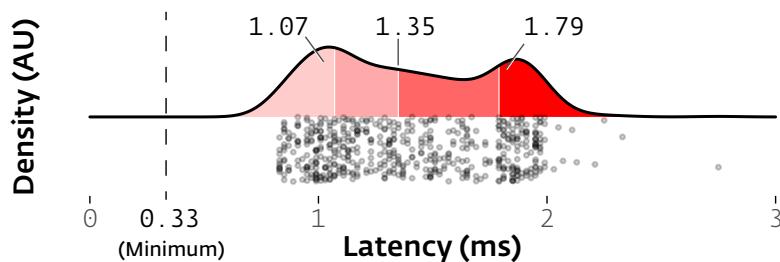
Figure 4.2: The pigs script used to trigger one pin (pin_out), from another (pin_in). At the expense of a little bit of complexity having to write a script in its scripting language, we are able to reduce latency by three orders of magnitude

Figure 4.3: Roundtrip latency from external trigger to digital output using two methods: Red (Median \pm IQR: $474 \mu s \pm 52.5$) – Typical Autopilot callback function given to a [Digital_In](#) object that turns a [Digital_Out](#) pin on for 1ms when an input pin changes state. Blue (380ns \pm 140) – Pigs script that runs entirely within the pigpio daemon. For each, black points represent individual measurements (n=525). Note the difference in x-axis scale

4.2 Sound Latency

We measured end to end, hardware input to sound output latency by measuring the delay between an external digital input and the onset of a 10kHz pure tone (Table 4.1). Sound playback was again triggered by the `Digital_In` class's callback method, and sound samples were buffered in a `deque` held in a separate process by the jack audio client between each trial. A `Digital_Out` pin was wired to the `Digital_In` pin in order to deliver the trigger pulse (but the `Digital_Out` pin was uninvolved in the software trigger for sound output).

Autopilot's `jack` audio backend was configured with a 192kHz sampling rate with a buffer with two periods of 32 samples each for theoretical minimum latency of 0.33ms². We observed a median 1.35ms (± 0.72 IQR) latency across 521 samples — roughly 4x the theoretical minimum (Figure 4.4). This suggests that Autopilot eliminates most perceptible end-to-end latency, which is necessary for tasks that require realtime feedback. One clear future direction is to write the sound processing loop in a compiled language exposed with a foreign function interface (FFI) to decrease both latency and jitter.



² A previous version of this paper included benchmarking and comparison to Bpod and pyControl's sound onset latency, but since then both packages have changed substantially, including Bpod creating a new `hifi sound module` based off HiFiBerry hardware very similar to the card used here, making those benchmarks obsolete. In this version we have omitted comparative benchmarks in favor of allowing the maintainers of those packages to publish their own benchmarks due to the cost of acquiring the new hardware, the version dependency of any benchmarks, and the distraction from characterizing the system at hand.

Figure 4.4: Autopilot has a median 1.35ms ($\pm .72$ IQR) latency between an external trigger and sound onset. Individual trials (dots, n=521) are shown beneath a density plot (red area under curve) colored by quartile (shades, numbers above are median, first, and third quartile). This latency is roughly 4x the theoretical minimum (0.33ms, dashed line).

4.3 Network Latency

To support data-intensive tasks like those that require online processing of video or electrophysiological data, the networking modules at the core of Autopilot need high bandwidth and low latency.

To test the latency of Autopilot's networking modules, we switch from "script mode" to "Task mode."³ Tasks are useful for encapsulating multistage routines across multiple devices that would be hard to coordinate with scripts alone. Our `Network_Latency` task consists of one "leader" raspi sending timestamped messages to a "follower" raspi which returns the timestamp marking when it received the message. The two pis were directly connected to one another (rather than routing each message through agent-level `Station` objects) after the leader pi initiated the follower with a multi-hop "START" message routed through a Terminal agent containing the task and networking parameters. We measured latency using software timestamps while synchronizing the clocks of the two pis with Chrony, an `NTP` daemon previously measured to synchronize Raspberry Pis within dozens of microseconds[48]⁴, with the leader pi hosting an `NTP` server and the follower pi synchronizing its clock solely from the leader. We documented this [on the wiki](#) too, since synchronization is a universal problem in multi-computer experiments.

³ Link to task code

⁴ Our sync is likely to be near to or better than that reported in [48]: in addition to a quiet network, we configured chrony to poll more frequently and tolerate a smaller error than default

> note that this latency includes message serialization and deserialization, which take

$200\mu\text{s}$

4.4 Network Bandwidth

To demonstrate yet another modality of use, we tested network capacity using Autopilot’s `Bandwidth_Test` widget, an action available from the Terminal GUI’s tests menu. This test requests that a set of selected pilots send messages at a range of selected frequencies and payload sizes back to the terminal. The messages pass through four networking objects en route: the stations and network nodes running the test for both the terminal and pilots (See Figure 3.14). Delay is measured as the duration between the creation of the message at the sender and the processing of the message at the receiver. The Pis and terminal were synchronized on common NTP servers to align timestamps.

First we tested the limits of our terminal’s ability to receive messages from the 10 pilots that it controls. Our terminal is a modest desktop (complete with a vintage 2012 CPU, see Table 4.4) with ethernet connections to 10 Raspberry Pi 3b’s through a network switch. We first tested the rate at which the Pi 3b’s and our terminal could send and process typical (255 Byte) messages without a data payload (Figure 4.5, top). A single Pi was capable of sending at a maximum rate of 707 Hz without exceeding its nominal mean delay of 4.9 (± 0.47) ms. Adding additional Pis did not cause increased delay until the total sending rate surpassed roughly 2000 Hz. These are the rate limits of sending and receiving messages, respectively.

As we increased the size of each individual message by including payloads of generated data (Figure 4.5, bottom), the rate of messaging decreased, but the total throughput (message rate (Hz) * size (Bytes)) saturated linearly as a multiple of the number of sending Pis. The Raspberry Pi 3b has a shared USB/Ethernet Bus, and thus appears to have a relatively limited 11.8MB/s throughput.

Fortunately, the Raspberry Pi 4 has an independent gigabit ethernet bus. On a Raspberry Pi 4, Autopilot has a 41MB/s maximum throughput and a 1,919Hz maximum messaging rate (Figure 4.6). We observed a slightly higher messaging delay with the Raspberry Pi 4 (6.9ms vs. 4.9ms Raspberry Pi 3B+). We note that the NTP synchronization method we used to measure delays has a margin of error on the order of milliseconds.

Autopilot’s networking modules are capable of supporting the infrastructure of next-generation behavioral neuroscience experiments. Our humble terminal was capable of receiving the full 114.6MB/s of 10 Pis without sign of saturation, and a Raspberry Pi 4 is capable of sending data at 41MB/s. This bandwidth makes Autopilot capable of streaming raw Calcium imaging⁵ and electrophysiological data from modern high-density probes⁶. The delay between sending and processing messages over 4 hops in a network (4.9ms) is less than the latency with which comparable systems (Figure 4.4) process triggers when connected directly via serial.

Finally, while Autopilot typically operates in a “TCP-like” protocol—resending messages until they have been confirmed as received—these tests were run with an optional “UDP-like” protocol which does not check for confirmation. Across the approximately 2.5 million messages sent during these tests only 537 were dropped (and only during tests which saturated rate or bandwidth capacity), giving Autopilot a delivery rate of 99.98% in “UDP” mode. By design, delivery rate is guaranteed to be 100% in “TCP” mode.

Table 4.4: Terminal Specs

CPU	AMD FX-4300
GPUS/Pi PIS	3.8GHz
Memory	8GB
Ethernet	1Gbit/s
Switch	NETGEAR GSS116E

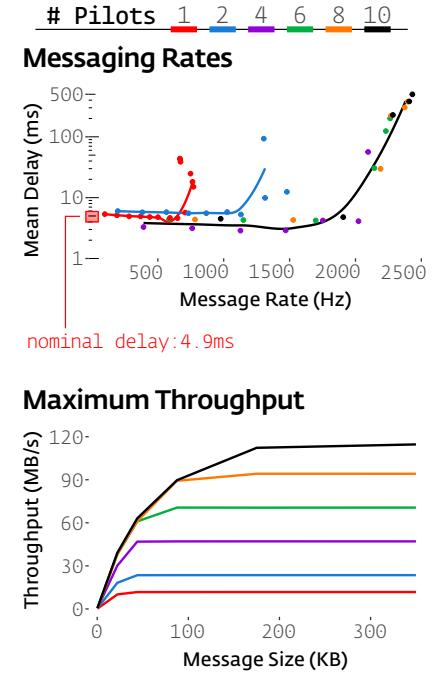


Figure 4.5: Network latency (top) and throughput (bottom) tests. Each point in the latency test represents the mean rate and delay of 5,000 255 Byte messages. Throughput (bottom) was calculated as the product of message rate and message size, and is displayed for a test that requested different numbers of pilots (colors) to send messages of different size to the terminal.

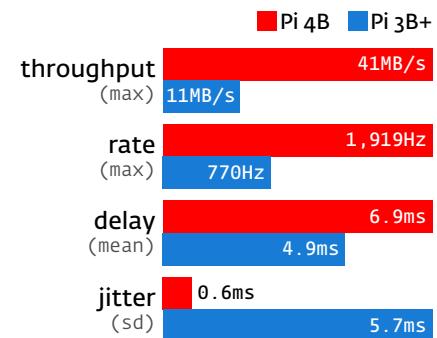


Figure 4.6: The Raspberry Pi 4’s gigabit ethernet bus markedly improves network performance.

⁵ 2-Photon: 5.9MB/s
(12 bits * 512x512 resolution * 15Hz)

⁶ Neuropixels: 14.4MB/s[24]
(10 bits * 30kHz * 384 channels)

4.5 Distributed Go/No-go Task

We designed a visual go/no-go task as a proof of concept for distributing task elements across multiple Pis, and also for the presentation of visual stimuli (Figure 4.7). The code for this task is described in greater detail in the user guide.

In this task, a head-fixed subject would⁷ be running on a wheel in front of a display with a lick-detecting water port able to deliver reward. Above the port is an LED. Whenever the LED is green, if the subject drops below a threshold velocity for a fixation period, a grating stimulus at a random orientation is presented on the monitor. After a random delay, there is a chance that the grating changes orientation by a random amount. If the subject licks the port in trials when the orientation is changed, or refrains from licking when it is not, the subject is rewarded.

One “parent” pilot controlled the operation of the task, including the coordination of its child⁸. The parent was connected to the LED and solenoid valve for reward delivery, as well as a monitor⁹ to display the gratings¹⁰. The child continuously streamed velocity data (measured with a USB optical mouse against the surface of the wheel) back to the terminal for storage (see also Figure 3.14, which depicts the network topology for this task). The child waited for a message from the parent to initiate measuring velocity, and when a rolling average of recent velocities fell below a given threshold the child sent a TTL trigger back to the parent to start displaying the grating. This split-pilot topology allows us to poll the subject velocity continuously (at 125Hz in this example) without competing for resources with psychopy’s rendering engine.

We measured trigger (TTL pulse from the child) to visual stimulus onset latency using the measurement cursors of our oscilloscope as before. To detect the onset of the visual stimulus, we used a high-speed optical power meter¹¹ attached to the top-left corner of our display monitor. The stimulus was a drifting Gabor grating drawn to fill half the horizontal and vertical width of the screen (960 x 540px), with a spatial frequency of 4cyc/960px and temporal (drift) frequency of 1Hz.

We observed a bimodal distribution of latencies (Quartiles: 28, 30, 36ms, n=50, Figure 4.8), presumably because onsets of visual stimuli are quantized to the refresh rate (60Hz, 16.67ms) of the monitor. This range of latencies corresponds to the second and third frame after the trigger is sent (2/3 of observations fall in the 2nd frame, 1/3 in the 3rd frame). We observed a median framerate of 36.2 FPS (IQR: 0.7) across 50 trials (8863 frames, Figure 4.9).

We further tested the Pi’s framerate by using Psychopy’s `timeByFrames` test—a script that draws stimuli without any Autopilot components running—to see if the framerate limits were imposed by the hardware of the Raspberry Pi or overhead from Autopilot (Table 4.5). We tested a series of Gabor filters and random dot stimuli (dots travel in random directions with equal velocity, default parameters) at different screen resolutions and stimulus complexities. The Raspberry Pi was capable of moderately high framerates (>60 FPS) for smaller, lower resolution stimuli, but struggled (<30 FPS) for full HD, fullscreen stimuli.

Autopilot is appropriate for realtime rendering of simple stimuli, and the proof-of-concept API we built around Psychopy doesn’t impose discernible overhead (Mean framerate for a 960 x 540px grating at 1080p in Autopilot: 36.2 fps, vs. `timeByFrames`: 35.0 fps). In the future we will investigate prerendering and caching complex stimuli in order to increase performance. A straightforward option for higher-performance

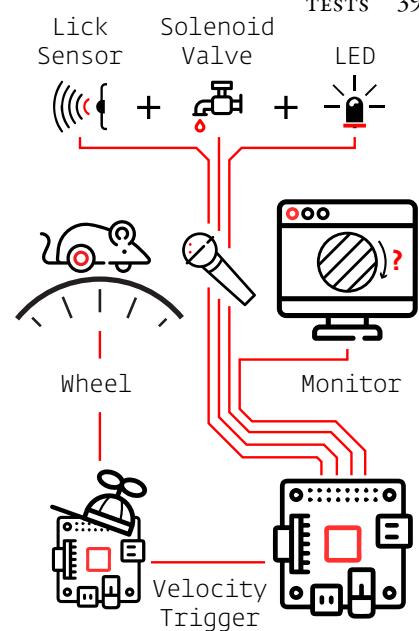


Figure 4.7: Hardware distribution for the distributed go/no-go task

⁷ No mice were trained on this task

⁸ Both Raspberry Pi 4s

⁹ Acer S230HL - (1920x1080px, 60Hz)

¹⁰ Visual stimuli were presented with Psychopy (v3.1.5) using the `glfw` (v1.8.3) backend while Autopilot was run in a dedicated X11 display server.

¹¹ Thorlabs PM100D

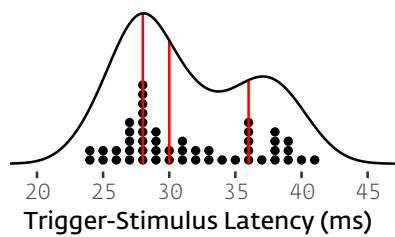


Figure 4.8: Stacked dots are a histogram of individual observations (n=50) underneath the probability density (black line), red lines indicate quartiles.

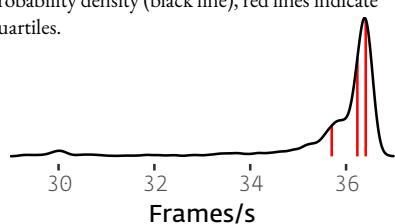


Figure 4.9: Probability density of framerates for 960 x 540px grating rendered at 1080p. Red lines indicate quartiles

video would be to deploy an Autopilot agent running on a desktop computer with a high-performance GPU, or to use a single-board computer with a GPU like the [NVIDIA Jetson \(\\$99\)](#).

Stimulus	Resolution	Size / # Dots	Mean FPS	σ FPS
Gabor Filter	1280 x 720	300 x 300px	106.4	5.5
Gabor Filter	1920 x 1080	300 x 300px	75.2	3.5
Gabor Filter	1280 x 720	640 x 360px	53.5	2.2
Gabor Filter	1920 x 1080	960 x 540px	35.0	1.0
Gabor Filter	1280 x 720	720 x 720px	41.5	2.2
Gabor Filter	1920 x 1080	1080 x 1080px	20.1	0.7
Random Dots	1280 x 720	100 dots	98.0	3.8
Random Dots	1920 x 1080	100 dots	67.6	3.0
Random Dots	1280 x 720	1000 dots	20.9	0.25
Random Dots	1920 x 1080	1000 dots	19.5	0.36

Table 4.5: Tests performed over 1000 frames with PsychoPy's `timeByFrames` test.

5

Limitations and Future Directions

WHILE WE BELIEVE THAT Autopilot’s order of magnitude increase of performance and decrease in expense, and its qualitative improvements in task design flexibility due to its distributed architecture are already useful contributions to behavioral neuroscience, we do not view Autopilot as “finished.” We view Autopilot—like all open-source software—as an evolving project. We are invested in its development, and will be continually working to fix bugs, make its use more elegant, and add new features in collaboration with its users.

We expect that as the codebase matures and other researchers use Autopilot in new, unexpected ways that some fundamental elements of its structure may evolve. We have built version logging into the structure of the system so that changes will not compromise the replicability of experiments (see **Versioning and Containerization** below). While there will inevitably be changes between versions, these will be both transparently documented and announced in release notes in order to alert users and describe how to adapt as needed. Accordingly, potential users should not let the limitations and future directions described below cause them to worry about early adoption or to wait for a stable version—the cost to start using Autopilot is low, and in our experience implementing experiments is already easier and more straightforward than comparable behavior systems.

We see several limitations in the launch version of Autopilot that we will improve on in future versions:

1. **Python 3** - We began developing Autopilot while there was still a case to be made for using Python 2. Now, given Python 2’s impending [end of life](#) in 2020, we will transition Autopilot to Python 3 by the end of 2019. We have already started transitioning with the `Subject` data class and don’t see the transition as a great obstacle.
2. **Synchronization** - Currently, there is no synchronization engine built into Autopilot. To ensure time-sensitive operations distributed over multiple Raspberry Pis are synchronized (ie. generate near-identical timestamps), we will add the ability for agents to [generate and follow a clock signal with pigpio](#). This synchronization engine will also allow alignment of Autopilot data with external software, such as the proprietary software often used for imaging data acquisition.
3. **Integration with Other Software** - We will make Autopilot capable of natively recording electrophysiological data by integrating with Open Ephys[47]. We also are interested in tightly integrating other recent tools like DeepLabCut[35] and MoSeq[57] to make Autopilot a unified platform for complex and naturalistic behavioral experiments.
4. **Transformations** - To enable the use of computer vision and other analytical tools within tasks we have begun building a data transformation module. This module will provide a framework to perform high-level data transformations—eg. images from a camera to positions of tracked objects—that convert raw data from hardware objects to processed data useful for designing complex tasks.
5. **Agents** - The Agent infrastructure is still immature—the terminal, pilot, and child agents are written as independent classes, rather than with a shared inheritance structure. We will be designing a common Agent class schema so that they are easier to design and deploy. We also plan to expand the available agents, specifically by introducing Observer and Compute agents. Observers will be designed for passive observation without supervision from a terminal, eg. for monitoring animals continuously in their home cages. Compute agents will run on high-performance computers in order to facilitate computationally intensive operations like GPU-dependent image analysis, online spike-sorting, etc. A mature agent framework will provide a much more streamlined path to the complex multi-agent experiments alluded to in Section 3.7.
6. **Data** - We plan on transitioning our data model to implementing the Neurodata Without Borders[42] standard. Since the Neurodata Without Borders standard is implemented in HDF5 and structurally similar to our data model, this transition should be straightforward. We also plan on adding support for a NoSQL [mongoDB](#) database backend to improve reliability, scalability, and performance of data storage and retrieval. Since our data model is standardized, we will ensure all data storage backends are mutually compatible so data stored in a database can be exported to HDF5 files and vice versa. Currently

Autopilot only automatically logs changes in task parameters and code version, but in the future we will expand our logging facility to include detailed data on systemwide preferences and connected hardware.

7. **Versioning and Containerization** - While Autopilot version and local changes are logged in collected data by default, there is no way to specify that a task should be run using a particular version automatically (ie. the user has to manually check out the specific git commit before running Autopilot). We intend on supporting task parameterizations that specify particular versions of Autopilot. We also will expand Autopilot's version logging system to include the versions of all the other packages in the environment. In our view, the best way to support reproducible software environments is to use a container system like [Docker](#), so we will be building infrastructure to generate containers from task parameterizations.
8. **Tasks** - We look forward to collaborating with other researchers to expand the available library of tasks. While the two-alternative forced choice and go/no-go tasks we have implemented are common, we designed Autopilot to be capable of performing *any* behavioral experiment. For example: we have already started a collaboration to build a freely-moving, jumping-based behavior that relies on 16 hardware components and data streams, and have future plans to build hardware and stimulus management extensions for human psychophysical tasks performed in an fMRI.
9. **Mesh Networking** - The tree structure of Autopilot's networking was built to enforce simplicity of its messaging protocol, but it limits the ability for data to be shared efficiently between a large number of pilots because communication has to be routed through a hub terminal. We will implement a true mesh network architecture by implementing a distributed hash table, allowing agents to directly communicate with one another without explicit configuration. We also will implement a peer-to-peer data protocol akin to BitTorrent to allow efficient distribution of data across a swarm of agents.
10. **Web Interface** - We would like to make a web-compatible UI that allows tasks to be administered and monitored from any computer. A web interface would make continuous experiments much easier to manage—we specifically intend this improvement (along with the Observer agent) to facilitate active sensory enrichment[[55](#), [17](#)] and developmental experiments.
11. **Platform Independence** - We have not rigorously tested Autopilot on operating systems other than Raspbian and Ubuntu Linux, though we know the terminal agent and its GUI works on macOS.
12. **Unit Tests** - At release, Autopilot has no unit tests. To make the codebase easier to maintain, we aim to reach 100% coverage by the first stable release of the program (v1.0).
13. **Save the GUI Plz**
14. **Plugins** - we want to be permissive, but we still need a way of specifying dependencies and all that. Also need to still complete all the actual management stuff like getting plugins and etc.
15. **Wiki** - generate data models to upload to wiki,

6

Glossary

Agent	3.7 The executable part of Autopilot. A set of startup routines (eg. opening a GUI or starting an audio server), runtime behavior (eg. opening as a window or running as a background system process), and event handling methods (ie. listens) that constitute the role of the particular Autopilot instance in the swarm .
Child	3.7 An agent that performs some auxiliary, supporting role in a task —primarily used for offloading some hardware responsibilities from a pilot .
Graduation	3.3 Moving between successive tasks in a protocol when some criterion is met.
Listen	3.8 A method belonging to the station or node of a particular agent that defines how to process a particular type of message (ie. a message with a particular key).
Node	3.8 A networking object that some module (eg. hardware, tasks , GUI routines) or method (eg. a listen) uses to communicate with other nodes . Messages to other agents in the swarm are relayed through their Station
Pilot	3.7 An agent that runs on a Raspberry Pi, the primary experimental agent of Autopilot. Typically runs as a system service, receives tasks from a terminal and runs them. Can organize a group of children if requested by the task .
Protocol	3.3 A (.json) file that contains a list of task parameters and the graduation criteria to move between them. The tasks in a protocol are also known as its levels .
Stage	3.3 Stages are methods that implement the logic of a task . They can be used analogously to states in a finite-state machine (eg. wait for trial initiation, play stimulus, etc.) or asynchronously (whenever x input is received, rotate stimulus by y degrees).
Station	3.8 Each agent has a single station , a networking object that is run in its own process and is responsible for communication between agents . The station also routes messages from children or other nodes .
Swarm	Informally, a group of connected agents .
Task	3.3 A formalized description of an experiment: the parameters it takes, the data that it collects, the hardware it needs, and a collection of stages that describe what happens during the experiment.
Terminal	3.7 A user-facing agent that provides a GUI for operating and maintaining a swarm .
Topology	3.7 A particular combination of agents , their designated responsibilities, and the networking connections between them invoked by a task (eg. task requires one pilot to record video, one to process the video, and one to administer reward) or by usage (eg. 10 pilots are connected to a single terminal and are typically used to run 10 independent tasks, though they could run shared tasks together).
Trial	3.3 If a task is structured such that its stages form a repeating series, a trial is a single completion of that series.

Bibliography

- [1] Emmeke Aarts, Matthijs Verhage, Jesse V. Veenvliet, Conor V. Dolan, and Sophie van der Sluis. A solution to dependency: Using multilevel analysis to accommodate nested data. 17(4):491–496. [2.8](#), [2.3](#)
- [2] Fatemeh Abyarjoo, Armando Barreto, Jonathan Cofino, and Francisco R. Ortega. Implementing a Sensor Fusion Algorithm for 3D Orientation Detection with Inertial/Magnetic Sensors. pages 305–310. [3.5](#)
- [3] Thomas Akam, Andy Lustig, James M Rowland, Sampath KT Kapanaiah, Joan Esteve-Agraz, Mariangela Panniello, Cristina Márquez, Michael M Kohl, Dennis Kätsel, Rui M Costa, and Mark E Walton. Open-source, Python-based, hardware and software for controlling behavioural neuroscience experiments. 11:e67846. [1.1](#)
- [4] Tim Anderson. Guido van Rossum aiming to make CPython 2x faster in 3.11. [2](#)
- [5] Jeremy Ashkenas, Haeyoun Park, and Adam Pearce. Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago. [2.3](#)
- [6] Ji Hyun Bak, Jung Yoon Choi, Athena Akrami, Ilana Witten, and Jonathan W Pillow. Adaptive optimal training of animal behavior. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 1947–1955. Curran Associates, Inc. [3.3](#)
- [7] STEPHEN R. BARLEY and BETH A. BECHKY. In the Backrooms of Science: The Work of Technicians in Science Labs. 21(1):85–126. [1](#)
- [8] Jayanti Bhandari Neupane, Ram P. Neupane, Yuheng Luo, Wesley Y. Yoshida, Rui Sun, and Philip G. Williams. Characterization of Leptazolines A–D, Polar Oxazolines from the Cyanobacterium Leptolyngbya sp., Reveals a Glitch with the “Willoughby–Hoye” Scripts for Calculating NMR Chemical Shifts. 21(20):8449–8453. [2.3](#)
- [9] Christopher P. Burgess, Armin Lak, Nicholas A. Steinmetz, Peter Zatka-Haas, Charu Bai Reddy, Elina A. K. Jacobs, Jennifer F. Linden, Joseph J. Paton, Adam Ranson, Sylvia Schröder, Sofia Soares, Miles J. Wells, Lauren E. Wool, Kenneth D. Harris, and Matteo Carandini. High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice. 20(10):2513–2524. [1](#)
- [10] Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: Why small sample size undermines the reliability of neuroscience. 14(5):365–376. [2.3](#)
- [11] Anna R. Chambers, Kenneth E. Hancock, Kamal Sen, and Daniel B. Polley. Online stimulus optimization rapidly reveals multidimensional selectivity in auditory cortical neurons. 34(27):8963–8975. [1](#)
- [12] Xinfeng Chen and Haohong Li. ArControl: An Arduino-Based Comprehensive Behavioral Platform with Real-Time Performance. 11. [5](#)
- [13] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. 1(1):e1400005. [2.3](#)
- [14] Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational Neuroethology: A Call to Action. 104(1):11–24. [1.2](#), [3.7](#)
- [15] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. 113(28):7900–7905. [2.3](#)
- [16] Chance Elliott, Vipin Vijayakumar, Wesley Zink, and Richard Hansen. National Instruments LabVIEW: A Programming Environment for Laboratory Automation and Measurement. 12(1):17–24. [1](#)

- [17] Navzer D. Engineer, Cherie R. Percaccio, Pritesh K. Pandya, Raluca Moucha, Daniel L. Rathbun, and Michael P. Kilgard. Environmental Enrichment Improves Response Strength, Threshold, Selectivity, and Latency of Auditory Cortex Neurons. 92(1):73–82. [5.10](#)
- [18] Open Ephys. pyControl. [1](#)
- [19] Jo Freeman. The Tyranny of Structurelessness. [9](#)
- [20] Pieter Hintjens. ZeroMQ: Messaging for Many Applications. O'Reilly Media, 1 edition edition. [2.2](#), [10](#)
- [21] Peter Johnson-Lenz and Trudy Johnson-Lenz. Groupware: Coining and defining it. 19(2):34. [1](#)
- [22] Peter Johnson-Lenz and Trudy Johnson-Lenz. Post-mechanistic groupware primitives: Rhythms, boundaries and containers. 34(3):395–417. [1](#)
- [23] Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open Source Scientific Tools for Python. [6](#)
- [24] James J. Jun, Nicholas A. Steinmetz, Joshua H. Siegle, Daniel J. Denman, Marius Bauza, Brian Barbarits, Albert K. Lee, Costas A. Anastassiou, Alexandru Andrei, Çağatay Aydin, Mladen Barbic, Timothy J. Blanche, Vincent Bonin, João Couto, Barundeb Dutta, Sergey L. Gratiy, Diego A. Gutnisky, Michael Häusser, Bill Karsh, Peter Ledochowitsch, Carolina Mora Lopez, Catalin Mitelut, Silke Musa, Michael Okun, Marius Pachitariu, Jan Putzeys, P. Dylan Rich, Cyrille Rossant, Wei-Lung Sun, Karel Svoboda, Matteo Carandini, Kenneth D. Harris, Christof Koch, John O'Keefe, and Timothy D. Harris. Fully integrated silicon probes for high-density recording of neural activity. 551(7679):232–236. [1](#), [2.8](#), [6](#)
- [25] Gary A Kane, Gonçalo Lopes, Jonny L Saunders, Alexander Mathis, and Mackenzie W Mathis. Real-time, low-latency closed-loop feedback using markerless posture tracking. 9:e61909. [2.2](#), [3.5](#), [3.5](#), [3.6](#)
- [26] Yarden Katz and Ulrich Bernhard Matter. On the Biomedical Elite: Inequality and Stasis in Scientific Knowledge Production. [2.3](#)
- [27] Dexter C. Kozen. Limitations of Finite Automata. In Dexter C. Kozen, editor, Automata and Computability, Undergraduate Texts in Computer Science, pages 67–71. Springer New York. [3.3](#)
- [28] Florian Krause and Oliver Lindemann. Expyriment: A Python library for cognitive and neuroscientific experiments. 46(2):416–428. [5](#)
- [29] The International Brain Laboratory, Valeria Aguillon-Rodriguez, Dora E. Angelaki, Hannah M. Bayer, Niccolò Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle A. Chapuis, Anne K. Churchland, Yang Dan, Eric E. J. Dewitt, Mayo Faulkner, Hamish Forrest, Laura M. Haetzler, Michael Hausser, Sonja B. Hofer, Fei Hu, Anup Khanal, Christopher S. Krasniak, Inês Laranjeira, Zachary F. Mainen, Guido T. Meijer, Nathaniel J. Miska, Thomas D. Mrsic-Flogel, Masayoshi Murakami, Jean-Paul Noel, Alejandro Pan-Vazquez, Cyrille Rossant, Joshua I. Sanders, Karolina Z. Socha, Rebecca Terry, Anne E. Urai, Hernando M. Vergara, Miles J. Wells, Christian J. Wilson, Ilana B. Witten, Lauren E. Wool, and Anthony Zador. Standardized and reproducible measurement of decision-making in mice. page 2020.01.17.909838. [1.2](#)
- [30] Gonçalo Lopes, Niccolò Bonacchi, João Frazão, Joana P. Neto, Bassam V. Atallah, Sofia Soares, Luís Moreira, Sara Matias, Pavel M. Itskov, Patrícia A. Correia, Roberto E. Medina, Lorenza Calcaterra, Elena Dreosti, Joseph J. Paton, and Adam R. Kampff. Bonsai: An event-based framework for processing and controlling data streams. 9. [5](#)
- [31] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. OpenSesame: An open-source, graphical experiment builder for the social sciences. 44(2):314–324. [5](#)
- [32] Wes McKinney. Pandas: A foundational Python library for data analysis and statistics. 14(9):1–9. [3.2](#)
- [33] Philip Meier, Erik Flister, and Pamela Reinagel. Collinear features impair visual detection by rats. 11(3). [1](#)
- [34] Greg Miller. A Scientist's Nightmare: Software Problem Leads to Five Retractions. 314(5807):1856–1857. [2.3](#)

- [35] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. 14(7):2152–2176. 1, 5.3
- [36] Christopher M. Niell and Michael P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. 65(4):472–479. 1
- [37] Ana Parabucki, Alexander Bizer, Genela Morris, Antonio E. Munoz, Avinash D. S. Bala, Matthew Smear, and Roman Shusterman. Odor Concentration Change Coding in the Olfactory Bulb. 6(1). 1
- [38] Photis Patonis, Petros Patias, Ilias N. Tziavos, Dimitrios Rossikopoulos, and Konstantinos G. Margaritis. A Fusion Method for Combining Low-Cost IMU/Magnetometer Outputs for Use in Applications on Mobile Devices. 18(8). 3.5
- [39] J. M. Pearce, J. C. Molloy, S. Kuznetsov, and S. Dosemagen. Expanding Equitable Access to Experimental Research and STEM Education by Supporting Open Source Hardware Development. 2.3
- [40] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, and Erik Kastman. PsychoPy2: Experiments in behavior made easy. 51(1):195–203. 5, 3.6
- [41] Jacob Reimer, Matthew J. McGinley, Yang Liu, Charles Rodenkirch, Qi Wang, David A. McCormick, and Andreas S. Tolias. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. 7:13289. 1
- [42] Oliver Rübel, Andrew Tritt, Benjamin Dichter, Thomas Braun, Nicholas Cain, Nathan Clack, Thomas J. Davidson, Max Dougherty, Jean-Christophe Fillion-Robin, Nile Graddis, Michael Grauer, Justin T. Kiggins, Lawrence Niu, Doruk Ozturk, William Schroeder, Ivan Soltesz, Friedrich T. Sommer, Karel Svoboda, Ng Lydia, Loren M. Frank, and Kristofer Bouchard. NWB:N 2.0: An Accessible Data Standard for Neurophysiology. 2.3, 5.6
- [43] Josh Sanders. Sanworks - BPod. 1
- [44] Sanworks, LLC. 8 reasons to use Bpod's new HiFi module. 1.2
- [45] Jonny L. Saunders and Michael Wehr. Mice can learn phonetic categories. 145(3):1168–1177. 2.4
- [46] Patrick E. Shrout and Joseph L. Rodgers. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. 69(1):487–510. 2.3
- [47] Joshua H. Siegle, Aarón Cuevas López, Yogi A. Patel, Kirill Abramov, Shay Ohayon, and Jakob Voigts. Open Ephys: An open-source, plugin-based platform for multichannel electrophysiology. 14(4):045003. 5.3
- [48] Eduardo Soares, Pedro Brandão, and Rui Prior. Analysis of Timekeeping in Experimentation. In 2020 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), pages 1–6. 4.3, 4
- [49] David A. W. Soergel. Rampant software errors may undermine scientific results. 3. 2.3
- [50] Aaron Swartz. Aaron Swartz's A Programmable Web: An Unfinished Work. 3(2):1–64. 5
- [51] Kay Thurley and Aslı Ayaz. Virtual reality systems for rodents. 63(1):109–119. 1
- [52] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. 13(2):22–30. 6
- [53] Guido van Rossum. Glue It All Together With Python. 2.1
- [54] Matthew B. Wall. Reliability starts with the experimental tools employed. 113:352–354. 1, 2.3
- [55] Deborah L. Wells. Sensory stimulation as environmental enrichment for captive animals: A review. 118(1):1–11. 5.10
- [56] Leland Wilkinson. The Grammar of Graphics. In James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori, editors, Handbook of Computational Statistics: Concepts and Methods, Springer Handbooks of Computational Statistics, pages 375–414. Springer Berlin Heidelberg. 3.9

- [57] Alexander B. Wiltschko, Matthew J. Johnson, Giuliano Iurilli, Ralph E. Peterson, Jesse M. Katon, Stan L. Pashkovski, Victoria E. Abraira, Ryan P. Adams, and Sandeep Robert Datta. Mapping Sub-Second Structure in Mouse Behavior. 88(6):1121–1135. 5.3