HIFuzz: Human Interaction Fuzzing for Small Unmanned Aerial Vehicles

Theodore Chambers
tchambe2@nd.edu
University of Notre Dame
South Bend, USA

Michael Vierhauser
michael.vierhauser@jku.at
University of Innsbruck, Department of Computer Science
Innsbruck, Austria

Ankit Agrawal
ankit.agrawal.1@slu.edu
St. Louis University
St. Louis, Missouri, USA

Michael Murphy
Jason Matthew Brauer
murphym18@gmail.com
jbrauer90@gmail.com
Drone Response
Indiana, USA

Salil Purandare
Myra B. Cohen
salil@iastate.edu
mcohen@iastate.edu
Iowa State University
Ames, Iowa, USA

Jane Cleland-Huang
JaneHuang@nd.edu
University of Notre Dame
South Bend, USA

ABSTRACT

Small Unmanned Aerial Systems (sUAS) must meet rigorous safety standards when deployed in high-stress emergency response scenarios; however many reported accidents have involved humans in the loop. In this paper, we, therefore, present the HiFuzz testing framework, which uses fuzz testing to identify system vulnerabilities associated with human interactions. HiFuzz includes three distinct levels that progress from a low-cost, limited-fidelity, large-scale, no-hazard environment, using fully simulated Proxy Human Agents, via an intermediate level, where proxy humans are replaced with real humans, to a high-stakes, high-cost, real-world environment. Through applying HiFuzz to an autonomous multi-sUAS system-under-test, we show that each test level serves a unique purpose in revealing vulnerabilities and making the system more robust with respect to human mistakes. While HiFuzz is designed for testing sUAS systems, we further discuss its potential for use in other Cyber-Physical Systems.

CCS CONCEPTS

• **Computer systems organization** → *External interfaces for robotics*; • **Human-centered computing** → *Interaction devices*; • **Software and its engineering**;

KEYWORDS

human-interaction, safety, sUAS, Cyber-Physical Systems

ACM Reference Format:
Theodore Chambers, Michael Vierhauser, Ankit Agrawal, Michael Murphy, Jason Matthew Brauer, Salil Purandare, Myra B. Cohen, and Jane Cleland-Huang. 2024. HIFuzz: Human Interaction Fuzzing for Small Unmanned Aerial Vehicles. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3613904.3642958

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642958

1 INTRODUCTION AND MOTIVATION

Small Unmanned Aerial Systems (sUAS) need to meet rigorous safety requirements when deployed in high-stress emergency response scenarios [27, 31]. However, the continual growth in sUAS deployment increases the risk of major incidents. Furthermore, several studies have reported that human "errors" have contributed to 65% to 85% of reported accidents in Cyber-Physical Systems (CPS) such as sUAS [19, 34, 41, 61]. We observed this phenomenon firsthand during a test flight in the Spring of 2023 (cf. Figure 1), when one of our autonomous sUAS breached a geofence, flew off its designated flight path, and ascended to an altitude of 734 feet above ground level (AGL) – far above the legal limit of 400 feet AGL. A post-mortem analysis revealed a series of factors, including human-related missteps, that contributed to the incident. The remote pilot in charge (RPIC), who plays only a supervisory role under normal conditions, failed to set appropriate geofence-breach actions prior to the mission, placed the throttle in an incorrect position, lost situational awareness of the sUAS' trajectory following the geofence breach, and failed to take timely action when the sUAS started to fly off-course. However, blaming the operator for these accidents is very shortsighted.

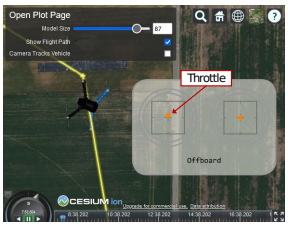
Human-Centered Design (HCD) focuses on creating and validating intuitive interfaces that are tailored to human cognitive capabilities [26, 46] and, therefore, are designed to reduce human error. However, in the emergent area of sUAS, any failure to anticipate and address normal human "mistakes" [17, 18] can eventually lead to potentially dangerous incidents at critical moments of a flight. A more systematic approach is therefore urgently needed to detect and mitigate design weaknesses that make the system vulnerable to human mistakes. In this paper, we propose human interaction testing techniques designed to reveal aspects of the system for which incorrect and unexpected human actions and inputs can result in potentially hazardous system behavior [15, 27, 37].

We present the **H**uman-machine **I**nteraction **Fuzz** testing framework named "*HIFuzz*", where "HI" represents both human interactions and the fact that sUAS fly at height. **Fuzz** is analogous

(a) The SW corner of the mission intersected with the geofence; however no geofence-action was set.

(b) Flight replay showed that the sUAS flew north at increasing altitude after the geofence breach.

(c) The RPIC must position the throttle correctly in case in-flight problems require human control.



(d) Flight replay revealed that the RPIC had incorrectly set the throttle above neutral.

Figure 1: Due to a combination of mistakes, including 'operator error' by the Remote Pilot in Command, the sUAS flew off-route and ascended to 734 feet AGL. Note: All required regulatory reports were filed describing the incident.

to traditional fuzz testing, where inputs are iteratively mutated and tested against the system to cover a large part of the behavior (and/or the code base) of an application, in order to reveal software defects and vulnerabilities [64]. Fuzz testing, also known as fuzzing, has been applied across various domains in software and system testing due to its effectiveness in uncovering vulnerabilities and defects [22, 60]; however, to the best of our knowledge, it has not previously been leveraged to probe for undesirable outcomes associated with human interactions.

Our *HIFuzz* framework includes three distinct levels (L1, L2, L3) progressing from a low-cost, limited-fidelity, large-scale, no-hazard environment, with fully simulated Proxy Human Agents (L1), via an intermediate level, where proxy humans are replaced with real humans (L2), to a high-stakes, high-cost, real-world environment (L3). Replacing the human with a proxy in level L1 allows us to achieve fuzz-testing goals of rapid test coverage which would be impossible if a human were in the loop. At the same time, engaging humans in a small number of carefully selected tests at L2, allows us to investigate the human's situational awareness of the sUAS flight behavior [17]. We can leverage this knowledge to identify appropriate design mitigations in the form of alerts, explanations, and even automated failsafe actions. Finally, level L3 further increases test fidelity by repeating tests that have successfully passed level L2, whilst introducing additional real-life stressors such as physical safety concerns and environmental detractors such as the glare of the sun, that are an inevitable part of field deployments.

The levels are separated by two dedicated gateways. G1 resides between L1 and L2 and is responsible for down-selecting an appropriate set of tests to be executed in L2; while G2 represents a significant safety gateway in which standard safety assurance processes are followed, and hard decisions are made about executing *HIFuzz* tests in the real world. Our tests are supported by tools for generating and executing the Fuzz Tests. For example, in the case of levels L2 and L3, where real humans participate in the tests, we have developed a mobile app to interactively guide users through the actions they need to perform during test execution.

Our *HIFuzz* framework makes three key research contributions. First, it presents a *novel and systematic approach for human-interaction testing*, aimed at detecting, analyzing, and mitigating previously unknown hazards associated with human-sUAS interactions. Second, while *Fuzz Testing* has been commonly used for software and systems tests, to the best of our knowledge it has not previously been used *for human-interaction testing*. HiFuzz, therefore, makes a novel contribution, improving system robustness at the intersection of Human-Computer Interaction and Software Testing. Third, we conduct an in-depth analysis of *HIFuzz* applied to our own multi-sUAS system, and a preliminary analysis of its generalizability across additional CPS. Results reported in this paper show that *HIFuzz* reveals system vulnerabilities associated with human interactions, potentially leading to their mitigation and improved design solutions, and that all three test levels play a unique role in the testing process.

The remainder of the paper is structured as follows. Section 2 describes related work. Section 3 explains how an individual fuzz test is specified, and Section 4 describes the various test levels and gateways. Sections 5 and 6 describe experiments we conducted by applying *HIFuzz* to a multi-sUAS system and provide a comprehensive discussion of the results. Finally, Section 7 discusses limitations of our work, and Section 8 draws conclusions.

2 RELATED WORK

In this section we discuss related work associated with human-centered design of CPS, fuzz-testing, human error and interaction in sUAS operations, and human interaction testing methodologies. Based on this prior work we argue that fuzz-testing can be an effective strategy for uncovering human-interaction vulnerabilities in the complex and dynamic CPS domains.

2.1 Human Error in sUAS Operations

Herdel *et al.* [27] conducted a comprehensive study focusing on over 100 applications across 16 diverse domains including emergency response and surveillance. They identified several research challenges pertaining to human-drone interactions, including one directly related to our work, addressing different ways in which people interact with sUAS to perform complex tasks. We address this issue through systematically testing outcomes of expected and unexpected human inputs for diverse tasks.

Rakotonarivo *et al.* [54] conducted interviews with drone operators, safety consultants, and regulators to identify operational risks and challenges when operating sUAS. One of their key recommendations was to "Support exploration of operational parameters and estimate their impact on mission safety" in order to allow "operators

to explore options that could simplify their procedures". Our multi-level *HIFuzz* process is designed to identify and mitigate potential safety issues before they arise in field testing, or worst-case, during live mission execution. It supports the systematic testing of diverse mission parameters and tasks and generates respective reports and documentation as inputs for subsequent safety analysis.

Balot *et al.* [4] have collected a set of challenges associated with sUAS operations, related to HMIs, command and control, and management of sUAS operations. They argue that sUAS HMIs "should be designed to take best advantage of human performance capabilities", to "[...] promote safety of flight operations". While efforts have been taken to increase safety of sUAS operations [42], complex operational environments require thorough testing. This challenge was further investigated by Mccarley and Wickens [39] who proposed rules guiding levels of automation for different flight phases and operations and investigated different forms of control interfaces. With *HIFuzz*, we focus on this intersection in both simulated and real-world environments, by providing a thorough and structured multi-level testing framework.

2.2 Formal Methods for User Interaction Testing

Several researchers have used formal methods to make mathematical claims about the correctness of the system with respect to user interactions, using a formal language such as temporal logic, a state machine, or process algebra [7, 14]. Diverse aspects of the system are modeled including expected outputs for given inputs, timing constraints, error handling requirements, the sequence of user interactions allowed by the UI, underlying state transitions, data flow and finally expected user behavior, including potential misuse or unexpected interactions [45, 50]. Formal verification techniques, such as model checking or theorem proving are then used to mathematically prove that the UI model satisfies the formal specifications, and meets the initially stated requirements and intended use cases. Formal models can also be used to generate test cases. For example, Bolton *et al.* [8] conducted a review on formal approaches in human-automation interaction. They showed that formal methods help to uncover potential shortcomings in human automation interfaces, and are useful for diagnosing human-related system failures. However, formal methods are only as good as the assumptions made during the specification and modeling process. In particular the models of expected user interactions including misuse cases, in an emergent area, such as sUAS are unlikely to be complete or correct. *HIFuzz* takes a somewhat orthogonal approach to formal methods, in that it assumes that the system is flawed, and probes the system to unearth these flaws.

2.3 Fuzz Testing in Software Engineering

In the more general area of systems engineering, fuzz testing has emerged as an effective approach for testing large search spaces exhibiting high degrees of uncertainty (e.g., environmental factors) [11, 62]. The majority of fuzzing techniques are greybox (using code-guided metrics to diversify coverage of program paths in the code) [5, 6, 20, 47, 49]; however, scenario-based approaches, as adopted by *HIFuzz*, represent an alternative approach for specification-based fuzzing [11, 24, 58]. Fuzzing has been used effectively within the CPS domain. For example, Kim *et al.* [33] developed RVFuzzer to detect input validation bugs in robotic vehicle control programs including sUAS applications. However, they focused on detecting low-level controller malfunctions by monitoring vehicle control states. Similarly, Kim *et al.* [32] created PGFUZZ, a policy-based fuzzing framework for robotic vehicles, and focused on safety and functional policies with respect to user inputs, configuration parameters, and physical sUAS states. While they explicitly included user inputs and commands, they did not provide a comprehensive multi-level testing framework supported by safety analysis as used in *HIFuzz*. Finally, Han *et al.* [25] proposed a grey-box-based fuzzing framework for detecting incorrect configurations in sUAS flight controllers. Their LGDFuzzer combined fuzzing with a genetic algorithm to detect potentially incorrect configurations and to test them in simulation, but did not consider human-related actions or real-world physical testing.

3 DEFINING AN INDIVIDUAL *HIFUZZ* TEST

Each individual *HIFuzz* test focuses upon a human-interaction task that is conducted within a specific context. In this section, we therefore describe the elements and properties used to define an individual test.

3.1 Test Setup: Actors, Props, and Environment

Roles: Each human enacted task is assigned to a specific role such as a *Remote Pilot in Command* (RPIC), *Observer* (OBS), *Mission Commander* (MC), or *Safety Officer* (SO). We define $\mathbf{R}$ as the set of roles represented by $\mathbf{R} = \{r_1, r_2, r_3, \ldots, r_i\}$, and assumed by either a human or proxy-human depending on the current test level.

Interaction Devices: Humans perform a task using an interface device such as the radio control transmitter (RC), a GUI supported by a keyboard, mouse, and/or joystick, or another type of haptic device [10, 38, 43, 63]. We define $\mathbf{UI}$ as the set of all available user interfaces, represented as $\mathbf{UI} = \{ui_1, ui_2, ui_3, \ldots, ui_j\}$.

Drones and their Configurations: Tests can specify a specific drone or set of drones. Note that we utilize the word "drone", to emphasize the actual vehicle and its onboard flight controller, versus the complete software system. Inconsistencies across drones can cause accidents when their behavior fails to meet the human's current mental model [18]. We therefore define $\mathbb{D}$ as the set of drones, represented as $\mathbf{D} = \{d_1, d_2, d_3, \ldots, d_l\}$. Further, each drone in $\mathbf{D}$ can be configured by the user prior to flight -- for example, by setting a geofence around the drone or assigning it a unique RTL (return to launch) flight altitude. We define $\mathbf{P}$ as a set of configurable parameters for an sUAS given by $\mathbf{P} = \{p_1, p_2, p_3, \ldots, p_m\}$; however, low-level parameter configuration, that normally occurs when tuning the flight-controller [25] is out of scope of this paper, and we assume that each drone has been adequately tuned and is flight-worthy. Parameters of interest are therefore limited to those exposed to the operator through interfaces (e.g., GUI screens) and therefore accessible during pre-flight setup.

Simulation Environment: Finally, for Level L1 and L2 tests, depending upon the simulation environment used, we can directly configure elements such as wind. We define $\mathbf{E}$ as a set of configurable environmental parameters given by $\mathbf{E} = \{e_1, e_2, e_3, \ldots, e_n\}$.

3.2 *HIFuzz* Scripts

Humans (serving in a specific role) enact a human-interaction task (HIT) in the context of an sUAS mission. Further, they execute the HIT when the sUAS and/or mission is in a specific state. For example, the RPIC might be asked to perform the action of switching to POSITION mode when the drone is FLYING in OFFBOARD mode. This leads to the following specifications.

Missions: A mission represents the flight plans and other tasks that one or more sUAS will execute to provide context for the test. We define $\mathbb{MSN}$ as the set of available missions, represented as $\mathbf{MSN} = \{m_1, m_2, m_3, \ldots, m_q\}$.

Human Interaction Task (HIT): There are two types of HIT that a human will perform during a test. First, the human could provide input to an individual sUAS through a hardware device such as the RC – for example, by increasing the throttle, holding down the kill switch, or switching between modes. Second, the human could send a command to one or more sUAS via a GUI -- for example, issuing a global RTL command. We define $\mathbb{HIT}$ as the ordered set of interaction tasks performed by a user, represented as $\mathbf{HIT} = \{hit_1, hit_2, hit_3, \ldots, hit_r\}$.

However, CPS behavior is impacted by the current state of the system. Therefore, each HIT has an associated set of preconditions that also need to be defined. These preconditions are based on MODES, FLIGHT LIFE-CYCLE STATES, and CONFIGURATIONS. Modes are used by almost every flight controller to support common flight tasks such as TAKE-OFF and LOITER, and to provide various degrees of flight stability (e.g., STABILIZED and POSITION-HOLD) [3, 52]. We define $\mathbf{M}$ as the set of flight modes, given by $\mathbf{M} = \{m_1, m_2, m_3, \ldots, m_s\}$, where each mode m_i in $\mathbf{M}$ is reachable in the SuT. We also define $\mathbf{S}$ as a set of flight life-cycle states such as taking-off, flying, and landing, given by $\mathbf{S} = \{s_1, s_2, s_3, \ldots, s_t\}$. A drone can only be in one mode and one state at any time. Finally, we define configurations as the value assigned to any underlying parameter defined earlier as $\mathbf{P}$. Each HIT includes a mode and life-cycle precondition, and can optionally define a set of configuration parameters that serve as preconditions. Further, the precondition state must be reachable in at least one of the defined missions in order for any subsequent *HIFuzz* test to be valid.

3.3 Defining the *HIFuzz* Test

Based on these definitions, we can now specify an individual *HIFuzz* test in a way that is sufficiently formal for automating test execution, but also readable to humans who serve as participants in the testing process. We utilize JSON to represent each test as shown in Listing 1. The test definition includes the mission, environmental factors, roles, the locally sequenced HITS, and preconditions performed by each role using a specific interaction device and drone. The *HIFuzz* fuzzing engine ultimately uses these specifications to generate diverse combinations of properties, and the *HIFuzz* Test Runner uses it to deploy the test, monitor its progress, and to generate test prompts that are sent to the mobile app.

```
{
  "Mission": "BASIC-WAYPOINTS",
  "Environment": {
    "Wind": {
      "SPEED": "20KTS",
      "DIRECTION": "NORTH"
    }
  },
  "Roles": [
    {
      "Role": "RPIC",
      "HITS": [
        {
          "ID": "1",
          "Drones": ["GREEN"],
          "Task": "MOVE THROTTLE TO +1",
          "Mode": "OFFBOARD",
          "State": "TAKING-OFF"
        },
        {
          "ID": "2"
          "Drones": ["GREEN"],
          "Task": "SET MODE TO STABILIZED",
          "Mode": "OFFBOARD",
          "State": "FLYING"
        }
      ],
      "Interaction_Device": "RC TRANSMITTER"
    },
    {
      "Role": "MC",
      "HITS": [
        {
          "ID": "1"
          "Drones": ["GREEN"],
          "Task": "PRESS RTL BUTTON",
          "Mode": "STABILIZED",
          "State": "FLYING"
        }
      ],
      "Interaction_Device": "GUI"
    }
  ]
}
```

Listing 1: In this example test, each of two roles is assigned specific actions to perform.

3.4 Test Outcome

Each fuzz test is ultimately executed within the *HIFuzz* platform, and its outcome is evaluated across two different dimensions – first to determine if the test was *valid* or *invalid*, and second to determine if valid tests *passed* or *failed*. An invalid test fails to execute the full sequence of HITS, typically because preconditions for one or more of the HITS are never met. The outcome of valid tests is assessed as *passed* or *failed* based on *mission completion* and *mission adherence* criteria.

4 *HIFUZZ* TEST LEVELS AND GATEWAYS

The *HIFuzz* process involves three testing stages (L1-L3) separated by two gateways (G1, G2), each of which serves a unique purpose (cf. Figure 2). Individual tests are executed at each stage, however,

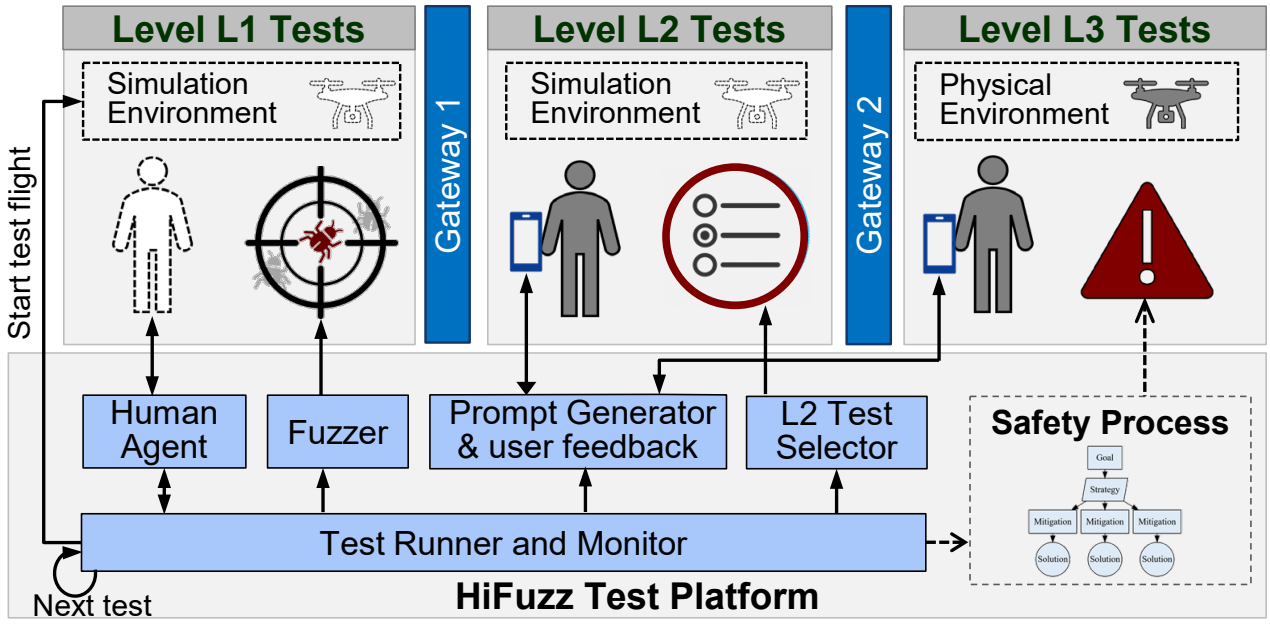


Figure 2: The *HIFuzz* framework supports tests at all three levels. L1 operates fully in a simulated environment with support from a fuzzer and a proxy human agent. L2 operates with real humans in an otherwise simulated environment, and L3 operates in the physical world.

the way they are executed, the role of human stakeholders, and the safety analysis that is performed prior to test execution differ greatly across stages. In this section, we therefore describe each stage and gateway.

4.1 Level L1: Large scale, simulated, fuzzing

The goal of L1 is to execute a large number of tests, as quickly as possible, without any of the risks involved in real-world sUAS flights. Therefore, L1 tests are run in the simulator using proxy human agents instead of humans. In the physical world, humans interact with sUAS via hardware devices, such as RC transmitters, and their inputs are encoded into radio signals transmitted to the flight controller and transformed into flight commands (e.g., throttle, yaw, pitch, and roll adjustments, or mode changes). These inputs can be simulated through software-based, low-level function calls to the flight controller. Humans also interact with sUAS via GUIs, and these interactions can be simulated if the SuT exposes its API function calls. Utilizing these techniques, L1 is able to simulate human interactions (i.e., HITS) entirely in software, enabling thousands of fuzz tests to be run in a low-cost, low-effort, non-hazardous environment.

The L1 process starts with a planning task in which the *HIFuzz tester* specifies the test features that constitute the fuzz space. As described in 3, these include roles, interaction devices, drones, environmental factors, missions, and HITs. The *HIFuzz* fuzzer then uses this specification to automatically generate combinations of the defined properties and input values constrained by specific scenarios of interest. The *Test Runner* iterates through the generated tests, invoking the mission in the simulation environment, monitoring the runtime state of each drone, checking for precondition states, and delegating HITs to the Proxy Human Agent when precondition states have been reached. The proxy mimics human input by replacing radio signals normally sent by the RC Transmitter, with MavROS manual control messages to simulate various switch changes and button presses for mode changes, throttle adjustments, and the kill switch. Results from each individual test are evaluated to determine if the test *passed*, *failed*, or was *untested* if the sUAS completed its mission without the preconditions ever being met. All passed and failed outcomes are passed to Gateway G1.

L1 requires a simulation environment that accepts and executes a mission request – potentially involving multiple drones, reports the progress of each drone throughout the mission, reports error messages, and produces a readable flight log at the end of each flight. Common examples of simulation environments that can be used to meet these requirements are Gazebo [48], jMAVSim [51], and AirSim [56].

4.2 Gateway G1: Downselecting for Human-in-the-Loop Tests

G1 serves as a gateway between levels L1 and L2, and is responsible for selecting tests to be passed to L2. Its inputs are the tests and results from L1. It clusters these tests to identify groupings of similar inputs and outcomes, in order to guide the L2 test selection process. The number of clusters is based on budgeted L2 testing time or based on a standard approach such as the "elbow-approach" which looks for the sweet spot in terms of coupling and cohesion of clusters [59].

Typically, one or two representative tests are selected from each cluster for execution at level L2.

4.3 Level L2: Humans in Simulated Environment

L2 tests are executed in the same simulation environment, however, humans replace the proxy agents, and interact with the sUAS through hardware devices (e.g., RC transmitters) and GUIs used in physical deployments. As explained earlier, Level L2 is designed to provide higher degrees of fidelity than L1, while operating within a completely safe testing environment; however, it introduces higher testing costs with respect to human time and effort. By integrating humans into the testing environment, L2 allows us to issue commands directly from the RC transmitter used in the field, providing increased fidelity of user inputs, and allowing direct observation of the sUAS behavior by human operators. Intuitively, Level L2 is needed to (1) execute a subset of interesting tests in a higher-fidelity environment, (2) to elicit feedback from humans about any failures that occurred in order to better understand their impact upon human operators, and ultimately (3) to evaluate the efficacy of user-facing mitigations, such as warnings or recommendations.

From a practical perspective, humans need help in determining when to perform a HIT, as many of the HIT's precondition states are internal, and not readily visible to human observers. *HIFuzz*, therefore, provides a mobile app responsible for generating timely prompts. In order to minimize unnecessary mental overload of processing and responding to prompts, the Mobile App is designed with a simple GUI which gives the user planning time as well as clear instructions on what actions to perform. We designed and implemented the mobile app following principles of human-centered design, and our two test participants reported that it was intuitive and gave them clear and timely directions. However, a full assessment of the mobile app is outside the scope of this paper, and

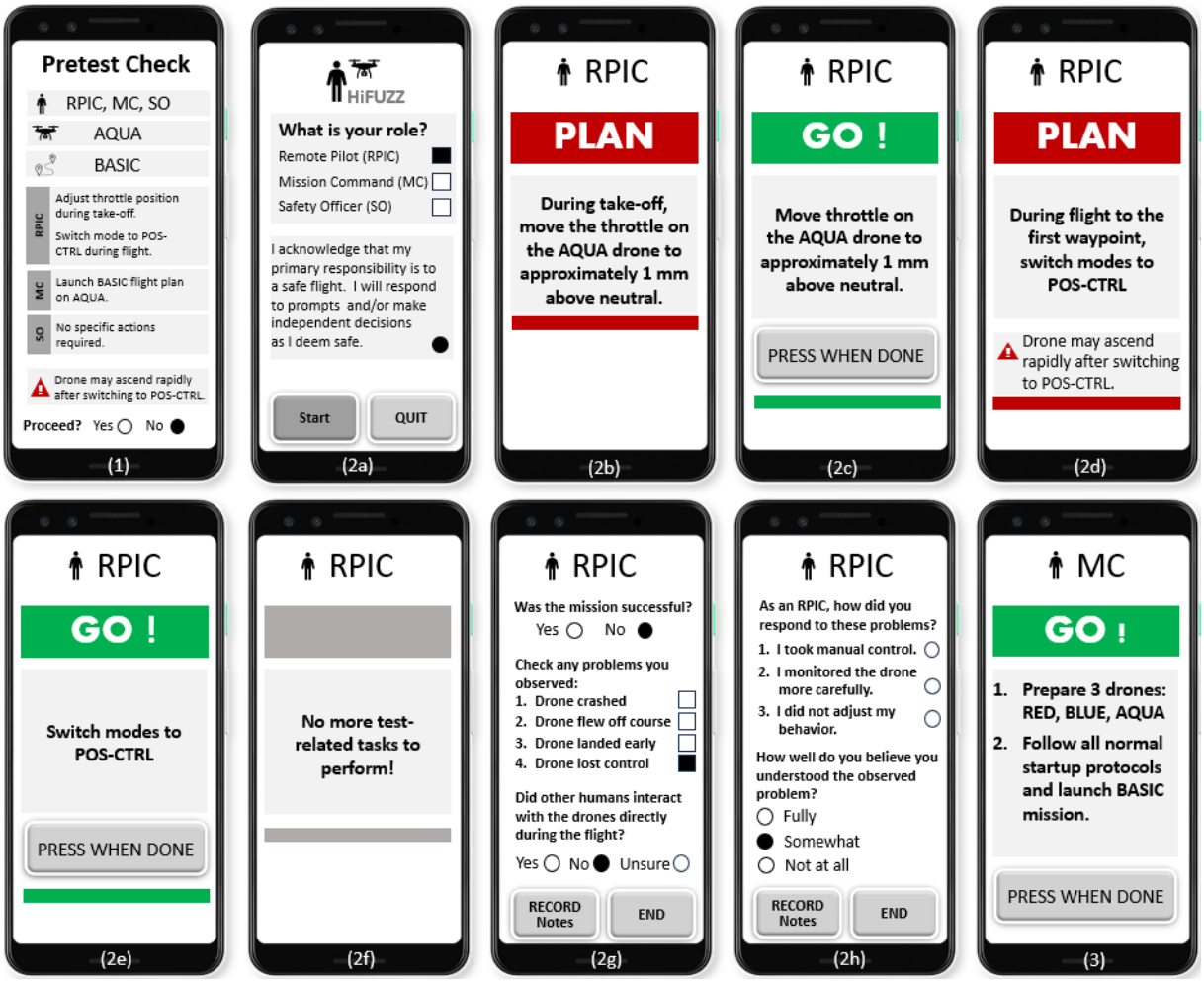


Figure 3: *HIFuzz* Prompts are shared with human test participants via a mobile app. Here we show the design of the tester's precheck screen (1), followed by a series of prompts shared with the RPIC (2a-h), and MC (3) roles respectively. Figures represent the design which was fully implemented and deployed using React-Native.

we therefore present it as a supporting tool rather than a primary contribution of this work.

A set of sample screens are depicted in Figure 3. The screens include preflight instructions and preparation (1, 2a, 3), a sequence of prompts that guide the RPIC (or other tester) through a sequence of tasks (2b-2f), and a series of post-test questions concerning the situational awareness of the operator (2g, 2h). We only engage trained personnel in these tests, with the expectation (as required by regulations) that all participants are fully trained in their roles and know which switches and knobs to manipulate in order to execute the intended task.

4.4 Gateway G2: Safety Assessment and Mitigation

While Test levels L1 and L2 seek to safely explore mission-breaking human-interaction faults that potentially cause erratic sUAS behavior, such as crashes and flight deviations, level L3's real-world deployment means that failures are potentially hazardous and costly. Therefore, Gateway G2 serves as a **safety gateway** that ensures that each failed test from L2 is carefully assessed to determine if mitigations are needed, and that all tests deployed on the field with physical sUAS have undergone a rigorous hazard analysis with all identified hazards sufficiently mitigated. The aim is to (1) assess human-interaction vulnerabilities and flaws identified in levels L1 and L2, (2) mitigate them, (3) repeat level L2 tests to demonstrate that they have been successfully mitigated, and only then (4) proceed to level L3 tests. *HIFuzz* does not dictate how the safety assessment should be performed as long as the process assesses hazards associated with each test case, e.g., using Fault-Tree Analysis (FTA) or Failure Mode Effect Criticality Analysis (FMEA/FMECA) [36, 55, 57], evaluates mitigations to determine whether the risk has been satisfactorily addressed, and when needed, provides a semi-formal safety case, e.g., a Safety Assurance Case (SAC) that includes guidelines targeted at the human participants describing how field tests can be conducted safely.

4.5 Level L3: Field Testing with humans-in-the-loop

The goal at level L3 is to validate that all tests that have previously produced a failed L2 outcome have been demonstrably mitigated. Intuitively, real-world tests are essential for two reasons. First, certain types of failures (especially race conditions) may only occur in the real world, and second, the human experience is different in the physical world than in simulation. For example, our own sUAS system was plagued for several months by a random take-off bug that appeared approximately once in every seven take-offs in the real world, but never in the simulator. Therefore, while simulations reveal many potential failures, real-world testing is essential for demonstrating that tests which executed successfully in simulation will also perform safely and correctly in the physical world.

4.6 Assessing Test Outcomes

HIFuzz utilizes an ensemble of test oracles and techniques to determine whether each flight has been executed correctly. These include analyzing runtime alerts generated by the flight controller and our own software system, reviewing mission logs, and considering human feedback received via the mobile app. For the log analysis, we establish a "blueprint" representing an ideal mission outcome, and then use it as a point of comparison to measure deviation in the flight logs for each test. For each position timestamp in the blueprint we compute the distance to the nearest sUAS position in the current test log across the x, y, and z axes, and record the largest distance as the maximum observed deviation of the current log from the blueprint. We also extract other features from each log, such as the maximum altitude, the duration of the flight, the occurrence of free-falls, the final landing state, and the reported mission status throughout each mission.

5 EXPERIMENTATION: *HIFUZZ* APPLIED TO *DRONERESPONSE* SYSTEM

We evaluated *HIFuzz* using a multi-sUAS system that we have developed and deployed in the real world as the *System-under-Test*. Our evaluation focuses upon the outcomes of *HIFuzz* rather than on the tools we have developed (i.e., the Mobile App), or the safety assessment (i.e., based on standard FMECA). We address three research questions.

RQ1: *What kinds of human-interaction vulnerabilities were identified using the HIFuzz process?*
This question investigates the types of vulnerabilities detected using *HIFuzz*.

RQ2: *Did each of the three test levels play a unique role in identifying human-related systems vulnerabilities?*
This question explores the efficacy of the three test levels versus the additional costs of human-in-the-loop testing.

RQ3: *Is HIFuzz generalizable across other human-intensive CPS applications?*
This question takes a preliminary look at the generalizability of *HIFuzz* to other domains.

The experiments described in this section were all executed in our *HIFuzz* platform.

5.1 System under Test: *DroneResponse*

DroneResponse is a distributed multi-user, multi-sUAS system, designed to support search-and-rescue, aerial data collection, and surveillance activities [2, 12, 28]. Each sUAS is equipped with an *Onboard Autonomous Pilot* (OAP) organized around a state machine which is dynamically configured for each mission. States support specific sUAS tasks such as *takeoff*, *search*, or *fly-to-waypoint* and vary greatly in complexity. For example, in the *takeoff state* the sUAS ascends to a predefined altitude and then transitions to a subsequent state such as *fly-to-waypoint*; while a *search* state utilizes AI-based computer vision capabilities to detect objects and make intelligent decisions, such as to *track* a person. A Ground Control Station (GCS) utilizes the MQTT message broker [40] to coordinate system-level communication between sUAS, humans, and microservices by publishing messages over a mesh radio. Status data (e.g., GPS location, battery, health) and task progress updates (e.g., current task, potential adaptations), are continually published by sUAS to support monitoring, analysis, and planning. Under normal operating conditions, humans set goals and send mission plans via GUI-based front-end clients; however, they can also directly

issue commands via RC Transmitters. A video of the *DroneResponse* system is available online.[1]

5.2 Scenario-Based Fuzz Tests

We adopted a scenario-based approach to test specific parts of the system. To select appropriate scenarios, we browsed through 272 issues (dated from 07/24/21 to 08/31/23) reported in the *DroneResponse* GitHub repository to identify incident reports associated with human-related incidents at the field (e.g., see Figure 4). We selected two incidents as depicted in Figure 5.

> **Github Issue #271:** Posted by: murphym18, 08/03/2023
>
> On August 3, during a flight test at peppermint field, a near-crash incident occurred that exposed a safety issue in our current mission format. Right now we specify the altitude as meters above sea level. But due to several unlucky coincidences, on Aug 3rd the flight controller ended up with an incorrect altitude reading.
>
> When the drone was flying home, it ended up flying to a waypoint that was alarmingly close to the ground. This required immediate intervention by the Remote Pilot in Command (Jane). The incident was not only intense and risky but could have resulted in significant damage.

Figure 4: An issue posted to Github describing a human-interaction incident, where the RPIC was forced to take control due to an altitude anomaly on the drone.

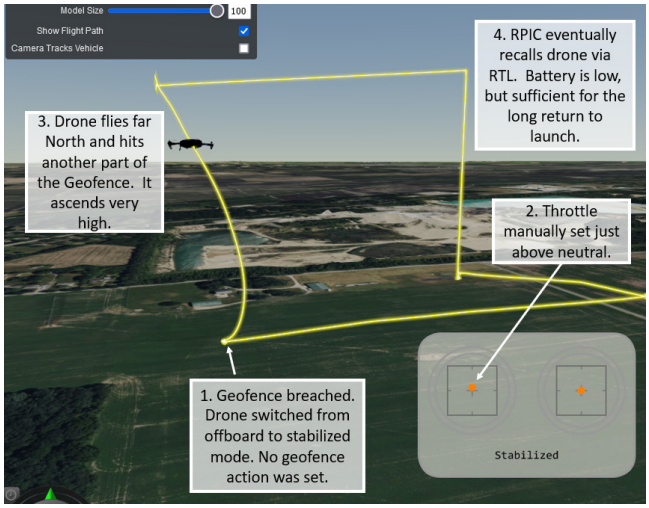
5.3 Modeling the test space

We defined relevant properties as described in Section 3. For example, to test Scenario 1, we created a flight route that intersected a geofence. We defined the search space as all reachable modes and states, one drone (BLUE), one human role (RPIC), two types of wind, several properties associated with geofence settings, and several throttle settings. For all additional flight controller parameters, we accepted values defined during the drone's prior configuration process. Finally, we included three human actions (HITs) to (a) change mode, (b) adjust the throttle position, and (3) kill the motors (essential in case of dire emergencies or for failed takeoffs). This resulted in a test space of approximately 160,524 test configurations. We then systematically generated combinations of these properties and human actions (as explained in 3) and fuzzed the exact timing at which each action was to be executed once all test properties were satisfied. Finally, we created a simple flight test involving one drone taking off, flying to two waypoints, and returning home.

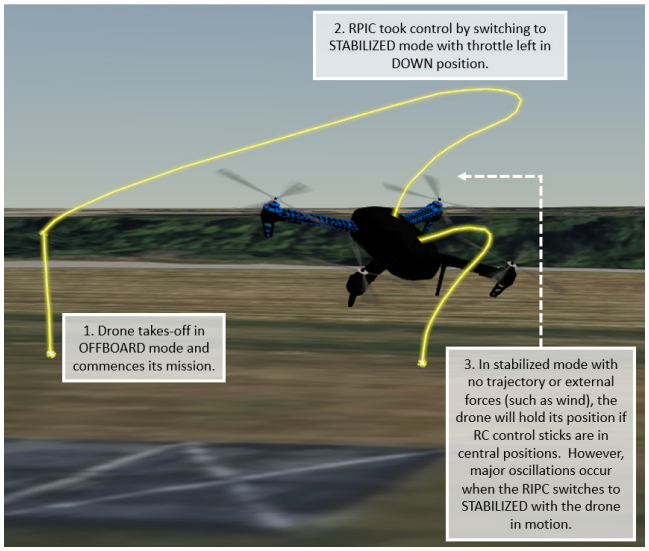
5.4 Applying *HIFuzz* to *DroneResponse*

We executed all levels (L1 - L3) and gateways (G1, G2) for the planned scenario-based fuzzing of the *DroneResponse* system with the following setup.

[1]DroneResponse demo: https://youtu.be/DyKqxkesgg0?si=2fVD1PNFpavYDI2y



(a) The drone hit the geofence with no geofence actions set and switched to stabilized mode. It then ascended rapidly and flew North. The RPIC had accidentally set the throttle just above neutral at the start of the flight. Fuzz tests explored various geofence breaches with diverse geofence actions, sUAS modes, and throttle positions.



(b) The RPIC took control of the drone by switching to STABILIZED mode with the throttle down. The drone oscillated as it attempted to stabilize and had a hard landing. Fuzz tests explored scenarios in which control was ceded to the RPIC whilst the sUAS was in various states and diverse throttle positions.

Figure 5: Two scenarios were selected in which human interactions were associated with flight failures. These scenarios were used in our experiments to drive scenario-based Fuzz Testing.

5.4.1 L1 Tests: We ran 700 L1 tests based on various combinations of properties from Table 1. Each test result was flagged with outcomes including the maximum altitude reached, flight duration, landed state, and mission completion. Any test exhibiting excessive

Table 1: Actual specification of the *HIFuzz* fuzzing space used for experimentation purposes. Legend: blue=initial states and modes, yellow=configuration settings, orange=drones, green=human tasks. For Level L1 we only utilize the RPIC role and BLUE drone. Further Geofence_Pred = 'On' ⇒ Geofence_stat='On' AND Geofence_ACT ⇒ Geofence_Stat='On'. This combination of features produced a test space of approximately 160,524 tests assuming no additional fuzzing around the precise timing of each test.

Modes
ALTCTRL
POSCTRL
OFFBOARD
STABILIZED
AUTO.LOITER
AUTO.RTL
AUTO.LAND

States
Pre-arm
Arm
Takeoff
Fly
Hover
Land

Throttle POS
Maximum HIGH
Medium HIGH
Just above neutral
Neutral
Just below neutral
Medium LOW
Maximum LOW

Wind
Medium Northerly
High Northerly

Geofence Stat.
On/Off

Geofence Pred.
On/Off

Geofence Act.
0: None
1: Warning
2: Hold mode
3: Return mode
4: Terminate
5: Land mode

Roles
RPIC
MC
SO

Drones
BLUE
ORANGE
PURPLE

Human Tasks
CHANGE-MODE
MOVE-THROTTLE
KILL-MOTORS

altitudes, duration, excessively fast landing, or failure to complete the mission with final disarm, was labeled as "Abnormal".

5.4.2 G1 Gateway: All tests in the profile were clustered using Within-Cluster Sum of Squares (WCSS), using the elbow method to determine the number of clusters to be generated [35]. This ultimately resulted in nine unique clusters which were used as a guide to search for interesting test cases to pass to L2. For clusters containing at least one *abnormal* test outcome, we selected the *abnormal* test case that was closest to the centroid. We then inspected the profiles of tests close to the boundaries of each cluster in order to identify interesting edge cases. This task took approximately one hour and resulted in the selection of 29 tests to pass to L2.

5.4.3 L2 Tests: Two researchers from our team executed all of the selected tests in the L2 simulation environment using a FrSky XD9 Plus Taranis Radio Handheld Controller [23]. The *tester* was responsible for the test setup, including launching the test runner, while the *RPIC* followed instructions displayed on the Mobile App, to conduct the planned human task at the correct stage of the mission. For each executed test, we preserved the flight logs, uploaded them into the PX4 flight log evaluation platform [53], then inspected the replayed flight log, logged messages, and graphs extracted from flight log data to further evaluate the flight outcomes. Figure 6 shows (a) the intended flight path of each test, (b) an actual flight path from one of the tests, and (c) one of the flight log data plots used to analyze the outcomes of a specific test. In this case, the RPIC switched modes to STABILIZED (as directed by the test runner) whilst the sUAS was flying in OFFBOARD mode. Due to the current trajectory and momentum of the sUAS, the sUAS continued its upward trajectory, ultimately reaching a height of 377 meters and a distance of over 550 meters. The tester ultimately issued a LAND command to force an end to the mission.

5.4.4 G2 Gateway: Two flight tests entered the G2 gateway during the course of our study. We leveraged our existing safety analysis process to assess safety risks associated with executing them in the physical world, and constructed a safety case using the Goal Structuring Notation (GSN) [30]. Once the tests were deemed safe to deploy we placed them into the field-test backlog. Due to space constraints, and the fact that the safety analysis process follows standard assurance practices, a deeper discussion on this gateway is out of scope of the paper. When necessary, additional tests were written to validate specific mitigations.

5.4.5 L3 Tests: So far, we have only executed one L3 test in the field, which successfully validated that a previously revealed vulnerability from L1 and L2 had been successfully mitigated. We discuss this particular L3 test in Section 6. Other identified mitigations are currently backlogged in our development pipeline.

6 ANALYSIS OF RESULTS

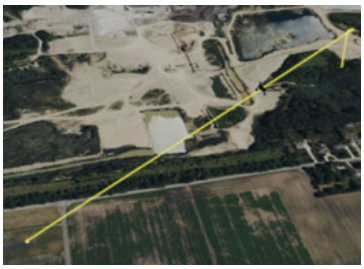
We now discuss the results from our experiment with respect to each of the research questions.

6.1 RQ1: What kinds of human-interaction vulnerabilities were identified using the *HIFuzz* process?

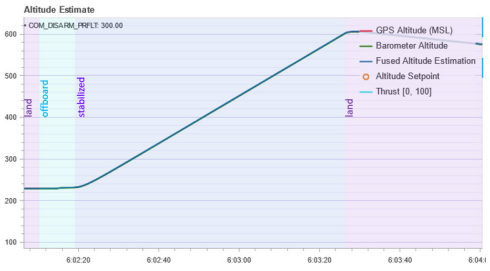
To address this question we conducted a systematic inductive analysis of the L2 test results. As a first step, the four reviewers carefully analyzed each test case outcome, and marked the test as *acceptable* or *problematic*, where an acceptable test outcome was deemed to be one in which no problems were observed, and a problematic one included at least one undesirable outcome. All four reviewers agreed that nine cases were problematic and eight were acceptable; however, they held differing opinions on the remaining 12 and therefore engaged in discussions in order to reach consensus. For example, there were three tests in which the RPIC pressed the kill switch to kill motors, but all three had different outcomes. In one case, the sUAS landed immediately (desired behavior), in one case it performed an RTL (return to launch), and in a final case, it entered a tug-of-war with the sUAS' autonomous pilot and had a rather spectacular crash landing. Only the third test's outcome might be considered 'bad', but in fact, the second case also was problematic as the observed behavior differed from expected. It was therefore also labeled as problematic. These kinds of nuanced analyses are a known issue in Fuzz Testing – where initial flags (passed/failed) tend to be rather coarsely applied. Based on discussion between the four researchers, 10 tests were ultimately classified as acceptable (i.e., false positives selected at gateway G2), and 19 as problematic.

Each assessor also assigned a tag describing the problem from the human-interaction perspective. One researcher performed an

(a) The basic flight path of the sUAS when GEOFENCE=INACTIVE.



(b) The actual flight path when Geofence=INACTIVE, and the RPIC executes a Mode Change to STABILIZED with DIR-Toggle="BACK".



(c) By inspecting plots and log outputs we assess the outcome of the flight and identify root cause of errors.

Figure 6: In this case the RPIC switched modes to STABILIZED whilst the sUAS was flying in OFFBOARD mode. Due to the current trajectory and momentum of the sUAS, it continued its upward trajectory, ultimately reaching a height of 377 meters and a distance of over 550 meters. Ultimately, the TESTER issued a LAND command to force an end to the mission. To minimize human errors caused by untimely mode-switches to STABILIZED, we can move the stabilized switch to a less prominent position, and add monitors to recognize if the drone is in 'free flight' due to a sudden switch to STABILIZE mode.

initial card-sorting exercise on these tags to create named clusters, producing eight candidate groupings of human-interaction vulnerability types. All four researchers then reviewed these groupings and discussed them in an online meeting. Following the discussion, six of the candidate groupings were retained (labeled 1-6 in Table 2), two groupings (*fly-away* and *failure to land*) were removed as they represented flight observations rather than human-interaction behaviors, and two additional categories were added (labeled 7-8 in Table 2). Table 2 lists the number of failed tests by vulnerability types.

Some of the most common user interface design problems in CPS are related to poor Situational Awareness (SA), impacting the ability of users to perceive, understand, and to make effective decisions [18]. These problems are documented as SA demons by Endsley [17] with three additional ones identified by Agrawal et al. [2], as listed in Table 2. To gain deeper insights into the underlying design flaws we mapped each vulnerability to one or more relevant SA demon, and then leveraged these mappings as a useful resource for identifying meaningful mitigations.

Here we describe one type of human-interaction vulnerability associated with *incorrect stick positioning* (See Case #1 from Table 2) as observed in five of the 29 test outcomes. Two of these cases involved incorrect throttle positions which is problematic if and when a human operator assumes manual control of the drone during flight. The problem originated from the default behavior of PX4 flight controllers, which requires the throttle to be fully down for arming. This behavior conflicts with the need for the throttle to be in the neutral position when the operator takes control so that the drone doesn't immediately crash land. We originally compensated for this problem by requiring the RPIC to move the throttle to the neutral position during takeoff in preparation for any later emergency. However, this created a stressful burden on the RPIC during a multi-sUAS takeoff. Our mappings to SA Demons associated the vulnerability with WAFOS (Workload, Anxiety, Fatigue, and Other Stressors) and MUI (transition failures across multiple interfaces) design demons. After gaining an understanding of the problem, we reprogrammed the takeoff routine to allow take-offs with the throttle in the neutral position thereby eliminating the previously required, error-prone human task. We also designed new alerts to warn the RPIC when the throttle was placed or left in a non-neutral position following takeoff.

Table 2 depicts several other types of vulnerabilities that we identified through the inductive analysis. HIE-1 and HIE-2 represented cases in which failures repeatedly occurred due to expectations placed upon the human operators at high-pressure points in the timeline. Both were mitigated through automation thereby relieving humans from these high-stress, error-prone activities. HIE-3 and HIE-4 both revealed *previously unknown vulnerabilities*. In HIE-3, the onboard autonomous pilot failed to recognize human interventions, thereby creating a tug-of-war between the human and the drone, leading to bizarre and unsafe flights; while in HIE-4, tests showed that the RC transmitter mappings included the ability for the operator to manually switch to offboard mode, meaning that the vehicle would no longer respond to commands from the RC transmitter. The remaining issues were all associated with loss of situational awareness related to a mode change. Brief descriptions are provided in Table 2.

6.1.1 Types of Vulnerabilities. Based on this analysis we can answer RQ1. The types of human-interaction vulnerabilities identified by *HIFuzz* covered diverse areas of the system design. They included unrealistic expectations placed on operators to perform tasks under time pressure, affordances that allowed human operators to perform actions that they should not be able to do, and missing alerts that meant that operators often lost situational awareness. Furthermore, we found two cases (HIE-3 and HIE-4), which were entirely unanticipated vulnerabilities associated with human actions. In the case of HIE-3, the tug-of-war detected by *HIFuzz* was very similar to the root cause of Lion Air Flight 610 and Ethiopian Airlines Flight 302 in which the MCAS (Maneuvering Characteristics Augmentation System) incorrectly perceived the angle of attack to exceed predefined limits and therefore pushed the nose of the plane down,

whilst pilots struggled to push it back up [21, 44]. This demonstrates that the *HIFuzz* process is capable of identifying highly critical and entirely unanticipated vulnerabilities. Furthermore, in other cases, such as HIE-1, we had already observed related incidents in the field but had previously not fully understood the behavior. *HIFuzz* tests provided new insights into the problem, leading to meaningful mitigations associated with automating prearming configurations and understanding when and where to issue warnings.

6.2 RQ2: Did each of the three test levels play a unique role in identifying human-related systems vulnerabilities?

To answer this question we take a retrospective look at whether *HiFuzz's* three test levels all served a unique role. Level L1 tests were fully automated, not requiring human intervention, and answered questions such as "did the flight complete successfully?", and "were there unexpected divergences from the planned route?". However, we had to *imagine* how an actual user would have observed and responded to the flight events that occurred. Therefore, even though **significant insights about potential human-interaction failures were gleaned from Level L1**, the results were insufficient for understanding users' perceptions and reactions to the problems as they occurred. Drawing upon our previous example of the incorrect throttle position during takeoff, field tests showed that (1) the RPICs almost always adjusted the throttle, but (2) frequently placed the throttle in a slightly incorrect position, with large consequences. Feedback from RPICs clearly showed that these 'mistakes' were due to stress and workload of supervising multiple sUAS during takeoff. A simple reminder would therefore be insufficient, and so we mitigated the problem through a complete redesign of the arming and takeoff routines, thereby removing this responsibility entirely from the operator. This type of insight is not obtainable with level L1 testing alone. Further, while we have not yet conducted a full user study with the Mobile App we developed, in future work we will ask deeper questions of test participants concerning the current system and the efficacy of mitigations such as the use of specific alerts and recommendations.

So far, this is one of only two tests that have been mitigated at L3. However, based on these two data points we observed that gateway G3 allowed us to take a deep dive into analyzing the safety concerns associated with executing tests in the field. It provided a safety net that helped us ensure that tests could be executed safely at Level L3. Demonstrating that the problem had been fixed and successfully deployed in the field built confidence that the system had satisfactorily addressed this particular system vulnerability. We conclude therefore that all three *HIFuzz* levels provide critical support for human-interaction testing.

6.3 RQ3: Is *HIFuzz* generalizable across other human-intensive CPS applications?

While our *HIFuzz* framework has been designed to identify risks related to human interactions in sUAS operations, its underlying concepts are applicable to a much broader range of CPS including other types of autonomous vehicles and ground-based robots. *HIFuzz* operates by fuzzing key system properties including (a) various modes in which a vehicle or robot operates, (b) different states it might transition into during the execution of a task or mission, and (c) potential human interactions with the system or robot. These core properties are found in other CPS, allowing *HIFuzz* to be applied in other domains and for other types of system applications. To investigate the potential use of *HIFuzz* across diverse CPS, we conducted a preliminary exercise of mapping the modes, states, and human interactions for systems from three different domains into *HIFuzz*. These included a centrally controlled sUAS system named Dronology, that used the Ardupilot Flight Controller [13, 16], a small robotic system developed by students to control a robot using a mobile phone, and a self-driving vehicle platform which we discuss in further detail.

The open-source, self-driving vehicle platform Autoware [1, 29] controls car operations and supports developers in creating autonomous car software systems. Similar to the modes available for our sUAS, Autoware manages different vehicle modes including Stop, Autonomous, Local, and Remote. Each of these modes represents a distinct operational setting for the vehicle. The *Stop* mode halts all autonomous functions, while the *Autonomous* mode enables full self-driving capabilities. *Local* and *Remote* modes refer to how humans interact with the car either with a steering wheel or over a network using a web application. An Autoware system can transition through multiple operational states such as *Idle*, where the vehicle is not actively navigating; *Active Navigation*, where the vehicle autonomously maneuvers through traffic or environments; and *Emergency*, a state triggered during critical situations requiring immediate action or human intervention. Other states include *Lane Following*, *Lane Changing*, and *Parking*. Further, the Autoware system also supports human intervention during vehicle operations, such as steering adjustments or mode switching. Additionally, self-driving vehicles operate in different environmental conditions, such as rain, snow, and bad lighting, and hence require rigorous testing. The concept of a *HIFuzz* test (as defined in Section 3) is therefore not unique to sUAS applications and potentially could be extended to other CPS that interact with humans and operate in a safety-critical, real-world environment. While individual aspects of a system are domain-specific (e.g., a role might be the backup driver instead of an RPIC), its key elements (Roles, Interaction Devices, Tasks, Modes, etc.) are applicable across very diverse contexts. For example, CARLA [9] provides a high-fidelity simulation environment for executing driving simulations with a multitude of configuration options. Scenario-based tests, such as driving an autonomous car on the road, under controlled conditions, can provide the context for the *HIFuzz* fuzzing.

Having defined properties for each of these three systems according to the types of properties used to define and execute *HIFuzz* tests, we draw the preliminary conclusion that *HIFuzz* is well suited to probing for human-interaction vulnerabilities across diverse CPS systems. Further, many parts of the *HIFuzz* infrastructure are entirely reusable including the test-runner, the mobile app, and the G1 clustering analysis. However, other parts of the infrastructure will need to be customized to each application and/or domain. These include adapters for interfacing with the simulation environments and metrics for evaluating acceptable versus problematic test outcomes. Primary adopters of *HIFuzz* are therefore likely to be domain experts with the technical skills needed to test a complex safety-critical system.

Table 2: Mapping to Situational Awareness Demons

HIE	Human Error Category	#	Outcome	SA Demon										
				AT	MS	IOL	OLS	EMM	RMT	WAFOS	CC	MUI	STC	EAU
1	RC transmitter sticks set incorrectly	5	Unexpected flight behavior (e.g., ascends, descends, or flies off course after control is ceded to user).				○			●		●		
2	Missing failsafe configurations	3	Operator fails to configure failsafes for each drone in the fleet in a consistent & standard way.			●		●		●				
3	Human input ignored by autonomous pilot	7	The autonomous system ignores a human-issued command, creating a "tug-of-war".				●				●			●
4	Inappropriate RC Switch options	1	The RC transmitter switches are mapped to modes that the operator should not use.									●		
5	Autonomous mode changes without notification	3	Human is unaware that the sUAS has switched mode and does not understand flight behavior.					●						●
6	Inappropriately timed mode change by operator	4	Human changed to a mode that was inappropriate for current phase and state of the flight.					●			●			
7	Failure to operate drone according to its current mode	4	Operator lacked or failed to apply appropriate piloting skills for current mode.					●						
8	Human loses situational awareness of sUAS behavior	6	Complex series of events led to loss of situational awareness and inability to recover from a failure.				●			●			●	●

Legend: *AT*=Attention tunneling, *MS*=Misplaced Salience, *IOL*=Information Overload, *OLS*=Out of the loop syndrome, *EMM*=Errant Mental Models, *RMT*=Requisite Memory Trap, *WAFOS*=Workload, Anxiety, Fatigue, & other Stressors, *CC*=Complexity Creep, *MUI*=transition failures across Graphical & Physical UIs, *STC*=Socio-Technical CPS Communication Failure, *EAU*=Enigmatic Autonomy. SG=Human Skill Gap. ●=Caused by, ○=Leads to.

7 LIMITATIONS AND FUTURE WORK

The research described in this paper is empirical in nature and is subject to three primary threats to validity.

First, our tests were limited to the RPIC, which is potentially the most challenging human role for operating sUAS; however, we need to extend the study to include other roles such as the MC (Mission Commander) and SO (Safety Officer), assign a more extensive set of human-interaction tasks, and study the perception of our stakeholders to identify further points of perceived vulnerabilities. In addition, we plan to allow humans to interact more freely with the L2 simulation environment, and deal with a far broader set of emergency tasks including deviant flight behaviors. Their success at intervening could serve as an indicator of the robustness of the design with respect to human interactions.

Second, while we conducted a preliminary investigation into the generalizability of *HIFuzz*, due to time constraints, we have not yet implemented *HIFuzz* in these systems. Instead, the experiments reported here focused on our own multi-sUAS system as the system-under-test. In future work, we plan to run experiments in the application of *HIFuzz* to other sUAS and CPS systems.

Third, we claimed that human-in-the-loop tests are essential for understanding how humans perceive problems and potential mitigations. We built the mobile app to not only guide users through the testing process but also to collect data from them describing their experiences during the test. Future work is needed to conduct user studies with the mobile app to evaluate its effectiveness.

Finally, as previously mentioned, the L2 level, while fully functional, had less fidelity to the field than we had intended, primarily because libraries used to interface the radio signals with software-based PX4 simulations had some limitations. In future work, we plan to augment, or ultimately entirely replace the L2 layer with a Hardware-In-The-Loop layer in which a physical flight controller is integrated closely into the simulated environment. This would further increase test fidelity and allow the RC transmitter to communicate over radio signals directly with the PX4 controller. Overall, increasing fidelity would allow more robust human-interaction testing, and improve the overall fidelity of our *HIFuzz* pipeline.

8 CONCLUSIONS

In this paper, we have presented the *HIFuzz* testing framework for probing a system for human interaction vulnerabilities. The multi-level approach progresses from a low-cost, limited-fidelity, large-scale, no-hazard environment, with fully simulated Proxy Human Agents (L1), through an intermediate level, where proxy humans are replaced with real humans (L2), to a high-stakes, high-cost, real-world environment (L3). In this paper we have focused on the systematic application of each part of the *HIFuzz* process, to identify human-interaction hazards so that we can design, implement, and validate mitigations. The end goal is to increase the robustness of the system so that it is fault-tolerant to normal human errors.

HIFuzz can be beneficial in two different ways. First, for testing individual systems, *HiFuzz's* multi-level approach provides a safe pathway for detecting vulnerabilities associated with human interactions in the system under test. While deploying *HIFuzz* for a new system is non-trivial, the return on investment in terms of human-interaction safety can make it worthwhile. Second, the lessons learned within a specific project can be documented and reused across other projects from similar domains, in order to help designers to avoid vulnerabilities in the first place. We therefore

plan to extend the scope of our *HIFuzz* tests, and document results in the form of a catalog.

In conclusion, results from applying *HIFuzz* to our own system under test have shown it to be effective in identifying critical human-interaction vulnerabilities, thereby directly addressing the need for improved system safety and robustness.

ACKNOWLEDGMENTS

The work in this paper conducted by US researchers was partially funded under USA National Aeronautics and Space Administration (NASA) Grant Number: 80NSSC21M0185 and the National Science Foundation (NSF) CNS-1931962 and CCF-1909688.

REFERENCES

[1] 2015. Autoware - the world's leading open-source software project for autonomous driving. https://github.com/autowarefoundation/autoware. (Accessed on 12/01/2023).

[2] Ankit Agrawal, Sophia J. Abraham, Benjamin Burger, Chichi Christine, Luke Fraser, John M. Hoeksema, Sarah Hwang, Elizabeth Travnik, Shreya Kumar, Walter J. Scheirer, Jane Cleland-Huang, Michael Vierhauser, Ryan Bauer, and Steve Cox. 2020. The Next Generation of Human-Drone Partnerships: Co-Designing an Emergency Response System. In *Proc. of CHI Conference on Human Factors in Computing Systems*. ACM, New York, 1–13. https://doi.org/10.1145/3313831.3376825

[3] Ardupilot. 2023. Flight Controller Modes. https://ardupilot.org/plane/docs/flight-modes.html. [Online; Accessed 01-07-2023].

[4] Clint R Balog, Brent A Terwilliger, Dennis A Vincenzi, and David C Ison. 2017. Examining human factors challenges of sustainable small unmanned aircraft system (sUAS) operations. In *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA*. Springer, 61–73.

[5] Marcel Böhme, Van-Thuan Pham, Manh-Dung Nguyen, and Abhik Roychoudhury. 2017. Directed Greybox Fuzzing. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 2329–2344. https://doi.org/10.1145/3133956.3134020

[6] Marcel Böhme, Van-Thuan Pham, and Abhik Roychoudhury. 2016. Coverage-Based Greybox Fuzzing as Markov Chain. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 1032–1043. https://doi.org/10.1145/2976749.2978428

[7] Matthew L Bolton and Ellen J Bass. 2009. A method for the formal verification of human-interactive systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. SAGE Publications Sage CA: Los Angeles, CA, 764–768.

[8] Matthew L Bolton, Ellen J Bass, and Radu I Siminiceanu. 2013. Using formal verification to evaluate human-automation interaction: A review. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43, 3 (2013), 488–503.

[9] CARLA. 2023. Open-source simulator for autonomous driving research. https://carla.org. [Online: accessed 8-14-2023].

[10] Linfeng Chen, Kazuki Takashima, Kazuyuki Fujita, and Yoshifumi Kitamura. 2021. Pinpointfly: An egocentric position-control drone interface using mobile ar. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[11] Yuqi Chen, Bohan Xuan, Christopher M Poskitt, Jun Sun, and Fan Zhang. 2020. Active fuzzing for testing and securing cyber-physical systems. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 14–26.

[12] Jane Cleland-Huang, Theodore Chambers, Sebastian Zudaire, Muhammed Tawfiq Chowdhury, Ankit Agrawal, and Michael Vierhauser. 2024. Human–machine Teaming with Small Unmanned Aerial Systems in a MAPE-K Environment. *ACM Trans. Auton. Adapt. Syst.* 19, 1, Article 3 (feb 2024), 35 pages. https://doi.org/10.1145/3618001

[13] Jane Cleland-Huang, Michael Vierhauser, and Sean Bayley. 2018. Dronology: an incubator for cyber-physical systems research. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*. 109–112. https://doi.org/10.1145/3183399.3183408

[14] Paul Curzon, Rimvydas Rukšėnas, and Ann Blandford. 2007. An approach to formal verification of human–computer interaction. *Formal Aspects of Computing* 19 (2007), 513–550.

[15] Byron DeVries and Betty HC Cheng. 2018. Run-time monitoring of self-adaptive systems to detect n-way feature interactions and their causes. In *Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems*. 94–100.

[16] Dronology. 2020. Research Incubator and Dataset. https://dronology.info. [Last accessed 01-01-2022].

[17] Mica R. Endsley. 2011. *Designing for Situation Awareness: An Approach to User-Centered Design, Second Edition* (2nd ed.). CRC Press, Inc., Boca Raton, FL, USA.

[18] Mica R Endsley. 2017. Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making* 11, 3 (2017), 225–238.

[19] Chin-Feng Fan, Ching-Chieh Chan, Hsiang-Yu Yu, and Swu Yih. 2018. A simulation platform for human-machine interaction safety analysis of cyber-physical systems. *International journal of industrial ergonomics* 68 (2018), 89–100.

[20] Andrea Fioraldi, Alessandro Mantovani, Dominik Maier, and Davide Balzarotti. 2023. Dissecting American Fuzzy Lop: A FuzzBench Evaluation. *ACM Trans. Softw. Eng. Methodol.* 32, 2, Article 52 (mar 2023), 26 pages. https://doi.org/10.1145/3580596

[21] Flight Safety Foundation. 2019. Preliminary Report B737-800MAX. https://flightsafety.org/preliminary-report-b737-800max-et-avj. [Last accessed 01-01-2022].

[22] Daniel S Fowler, Jeremy Bryans, Siraj Ahmed Shaikh, and Paul Wooderson. 2018. Fuzz testing for automotive cyber-security. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 239–246.

[23] FrySky. 2023. Taranis Series Handheld RC. https://www.frsky-rc.com/product-category/transmitters/taranis-series. [Online: accessed 8-14-2023].

[24] Jia Cheng Han and Zhi Quan Zhou. 2020. Metamorphic Fuzz Testing of Autonomous Vehicles. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops* (Seoul, Republic of Korea) *(ICSEW'20)*. Association for Computing Machinery, New York, NY, USA, 380–385. https://doi.org/10.1145/3387940.3392252

[25] Ruidong Han, Chao Yang, Siqi Ma, JiangFeng Ma, Cong Sun, Juanru Li, and Elisa Bertino. 2022. Control parameters considered harmful: Detecting range specification bugs in drone configuration modules via learning-guided search. In *Proceedings of the 44th International Conference on Software Engineering*. 462–473.

[26] Chenxu Hao, Anany Dwivedi, and Philipp Beckerle. 2022. A Literature-Based Perspective on Human-Centered Design and Evaluation of Interfaces for Virtual Reality in Robotics. In *Human-Friendly Robotics 2022 - HFR: 15th International Workshop on Human-Friendly Robotics, Delft, The Netherlands, 22-23 September 2022 (Springer Proceedings in Advanced Robotics, Vol. 26)*, Pablo Borja, Cosimo Della Santina, Luka Peternel, and Elena Torta (Eds.). Springer, 1–13. https://doi.org/10.1007/978-3-031-22731-8_1

[27] Viviane Herdel, Lee J Yamin, and Jessica R Cauchard. 2022. Above and beyond: A scoping review of domains and applications for human-drone interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[28] Md Nafee Al Islam, Muhammed Tawfiq Chowdhury, Ankit Agrawal, Michael Murphy, Raj Mehta, Daria Kudriavtseva, Jane Cleland-Huang, Michael Vierhauser, and Marsha Chechik. 2023. Configuring mission-specific behavior in a product line of collaborating Small Unmanned Aerial Systems. *J. Syst. Softw.* 197 (2023), 111543. https://doi.org/10.1016/J.JSS.2022.111543

[29] Shinpei Kato, Eijiro Takeuchi, Yoshio Ishiguro, Yoshiki Ninomiya, Kazuya Takeda, and Tsuyoshi Hamada. 2015. An open approach to autonomous vehicles. *IEEE Micro* 35, 6 (2015), 60–68.

[30] Tim Kelly and Rob Weaver. 2004. The Goal Structuring Notation – A Safety Argument Notation. In *Proc. Dependable Syst. Networks 2004 Work. Assur. Cases*.

[31] Md Nafiz Hasan Khan and Carman Neustaedter. 2019. An exploratory study of the use of drones for assisting firefighters during emergency situations. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[32] Hyungsub Kim, Muslum Ozgur Ozmen, Antonio Bianchi, Z Berkay Celik, and Dongyan Xu. 2021. PGFUZZ: Policy-Guided Fuzzing for Robotic Vehicles.. In *NDSS*.

[33] Taegyu Kim, Chung Hwan Kim, Junghwan Rhee, Fan Fei, Zhan Tu, Gregory Walkup, Xiangyu Zhang, Xinyan Deng, and Dongyan Xu. 2019. {RVFuzzer}: Finding input validation bugs in robotic vehicles through {Control-Guided} testing. In *28th USENIX Security Symposium (USENIX Security 19)*. 425–442.

[34] L.T. Kohn, J.M. Corrigan, and M.s. Donaldson. 1999. To err is human, Building a safety health system. *Washington, DC: National Academy Press* (1999).

[35] Wojtek J Krzanowski and YT Lai. 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* (1988), 23–34.

[36] Nancy G Leveson and Peter R Harvey. 1983. Software fault tree analysis. *Journal of Systems and Software* 3, 2 (1983), 173–181.

[37] Christoph Luckeneder, Michael Rathmair, and Hermann Kaindl. 2017. Investigating and coordinating safety-critical feature interactions in automotive systems using simulation. (2017).

[38] Vasudev S Mallan, Syam Gopi, Alexander Muir, and Rao R Bhavani. 2017. Comparative empirical usability assessment of two HRI input devices for a mobile

robot. In *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 331–337.
[39] Jason S. Mccarley and Christopher D. Wickens. [n. d.]. *Human factors concerns in UAV flight.* Technical Report.
[40] Henry Muccini and Mahyar Tourchi Moghaddam. 2018. IOT Architectural Styles. In *Proc. of 2018 European Conference on Software Architecture*. Springer, 68–85.
[41] D.C. Nagel. 1998. Human error in aviation Operations. *Human factors in Aviation, E.L.Weiner and E.C.Nagel (Eds)* 19890047069, 34 (1998), 263–303. https://doi.org/10.1109/2.910904
[42] NASA. 2023. NASA-UTM: Unmanned Aircraft Systems Traffic Management. https://www.nasa.gov/centers-and-facilities/ames/what-is-unmanned-aircraft-systems-traffic-management. [Online: accessed 8-14-2023].
[43] Pedro Neto, J Norberto Pires, and A Paulo Moreira. 2010. High-level programming and control for industrial robotics: using a hand-held accelerometer-based input device for gesture and posture recognition. *Industrial Robot: An International Journal* 37, 2 (2010), 137–147.
[44] Jack Nicas, Natalie Kitroeff, David Gelles, and James Glanz. 2019. Boeing Built Deadly Assumptions Into 737 Max, Blind to a Late Design Change. *The New York Times, https://www.nytimes.com/2019/06/01/business/boeing-737-maxcrash html [accessed: 23.01.2020]* (2019).
[45] Sara Nikula, Célia Martinie, Philippe A. Palanque, Julius Hekkala, Outi-Marja Latvala, and Kimmo Halunen. 2022. Models-Based Analysis of Both User and Attacker Tasks: Application to EEVEHAC. In *Human-Centered Software Engineering - 9th IFIP WG 13.2 International Working Conference, HCSE 2022, Eindhoven, The Netherlands, August 24-26, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13482)*, Regina Bernhaupt, Carmelo Ardito, and Stefan Sauer (Eds.). Springer, 70–89. https://doi.org/10.1007/978-3-031-14785-2_5
[46] Donald A. Norman and Stephen W. Draper (Eds.). 1986. *User centered system design: New perspectives on human-computer interaction.* Lawrence Erlbaum Associates, Hillsdale, NJ.
[47] Mitchell Olsthoorn, Arie van Deursen, and Annibale Panichella. 2021. Generating Highly-Structured Input Data by Combining Search-Based Testing and Grammar-Based Fuzzing. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering* (Virtual Event, Australia) *(ASE '20)*. Association for Computing Machinery, New York, NY, USA, 1224–1228. https://doi.org/10.1145/3324884.3418930
[48] Open Robotics. 2023. Gazebo. https://gazebosim.org. [Online: accessed 8-14-2023].
[49] Rohan Padhye, Caroline Lemieux, and Koushik Sen. 2019. JQF: Coverage-Guided Property-Based Testing in Java. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis* (Beijing, China) *(ISSTA 2019)*. Association for Computing Machinery, New York, NY, USA, 398–401. https://doi.org/10.1145/3293882.3339002
[50] Philippe Palanque and Célia Martinie. [n. d.]. Designing and Assessing Interactive Systems Using Task Models. 2016. In *ACM CHI Extended Abstracts*. 976–979.
[51] PX4. 2022. jMAVSim. https://docs.px4.io/master/en/simulation/jmavsim.html. [Last accessed 01-01-2022].
[52] PX4. 2023. Flight Controller Modes. https://docs.px4.io/main/en/flight_modes. [Online; Accessed 01-07-2023].
[53] PX4. 2023. Flight Review Platform. https://logs.px4.io/. [Online: accessed 8-14-2023].
[54] Balita Heriniaina Rakotonarivo, Nicolas Drougard, Stéphane Conversy, and Jérémie Garcia. 2023. Cleared for Safe Take-off? Improving the Usability of Mission Preparation to Mitigate the Safety Risks of Drone Operations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[55] Donald J. Reifer. 1979. Software Failure Modes and Effects Analysis. *IEEE Trans. Reliability* R-28,3 (1979), 247–249.
[56] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*. Springer, 621–635.
[57] Kevin J Sullivan, Joanne Bechta Dugan, and David Coppit. 1999. The Galileo fault tree analysis tool. In *Digest of Papers. Twenty-Ninth Annual International Symposium on Fault-Tolerant Computing (Cat. No. 99CB36352)*. IEEE, 232–235.
[58] Yang Sun, Christopher M. Poskitt, Jun Sun, Yuqi Chen, and Zijiang Yang. 2023. LawBreaker: An Approach for Specifying Traffic Laws and Fuzzing Autonomous Vehicles. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (Rochester, MI, USA) *(ASE '22)*. Association for Computing Machinery, New York, NY, USA, Article 62, 12 pages. https://doi.org/10.1145/3551349.3556897
[59] MA Syakur, BK Khotimah, EMS Rochman, and Budi Dwi Satoto. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, Vol. 336. IOP Publishing, 012017.
[60] Ari Takanen, Jared D Demott, Charles Miller, and Atte Kettunen. 2018. *Fuzzing for software security testing and quality assurance.* Artech House.
[61] Michael Vierhauser, Md Nafee Al Islam, Ankit Agrawal, Jane Cleland-Huang, and James Mason. 2021. Hazard analysis for human-on-the-loop interactions in sUAS systems. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 8–19.
[62] Herman Wijaya, Maurício Aniche, and Aditya Mathur. 2020. Domain-based fuzzing for supervised learning of anomaly detection in cyber-physical systems. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*. 237–244.
[63] Mingxin Yu, Yingzi Lin, David Schmidt, Xiangzhou Wang, and Yu Wang. 2014. Human-robot interaction based on gaze gestures for the drone teleoperation. *Journal of Eye Movement Research* 7, 4 (2014), 1–14.
[64] Xiaogang Zhu, Sheng Wen, Seyit Camtepe, and Yang Xiang. 2022. Fuzzing: A Survey for Roadmap. *ACM Comput. Surv.* 54, 11s, Article 230 (sep 2022), 36 pages. https://doi.org/10.1145/3512345