Towards a Unified Model of Accountability Infrastructures

Severin Kacianka
Technische Universität München
Garching b. München, Germany
kacianka@in.tum.de

Florian Kelbert
Technische Universität München
Garching b. München, Germany
kelbert@in.tum.de

Alexander Pretschner
Technische Universität München
Garching b. München, Germany
pretschn@in.tum.de

Accountability aims to provide explanations for why unwanted situations occurred, thus providing means to assign responsibility and liability. As such, accountability has slightly different meanings across the sciences. In computer science, our focus is on providing explanations for technical systems, in particular if they interact with their physical environment using sensors and actuators and may do serious harm. Accountability is relevant when considering safety, security and privacy properties and we realize that all these incarnations are facets of the same core idea. Hence, in this paper we motivate and propose a model for accountability infrastructures that is expressive enough to capture all of these domains. At its core, this model leverages formal causality models from the literature in order to provide a solid reasoning framework. We show how this model can be instantiated for several real-world use cases.

1 Introduction

We want to illustrate our understanding of accountability with a simple scenario: Imagine a vase being placed in the middle of an otherwise empty room. You place a cleaning robot (like the popular Roomba model) in the same room and set it to clean the room. You leave the room for one hour, and upon return you find the vase shattered; Figure 1 sketches the scenario.

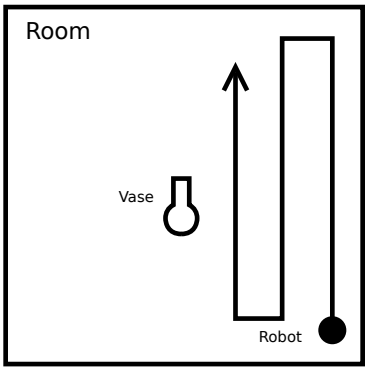

Figure 1: The vase room.

Given this, you would like to answer the following questions:

1. What caused the destruction of the vase?
2. Who is responsible for the destruction of the vase?
3. Who is liable for the destruction of the vase?

At first glance, it might seem that the answer to these questions is obvious: the robot. After careful consideration, however, the answer might not be so clear. We have no clear *evidence* of the robot's action. An alternative explanation might be that a window was left open and that a cat snuck into the room and tipped the vase. Another explanation could be an earthquake in the area.

Thus, in order to answer the questions above, we need some *evidence* whether it was indeed the robot that caused the vase to be shattered. *Accountability mechanisms* are able to provide such evidence. While such accountability mechanisms require technical components, they necessarily also need non-technical components. To answer the question of liability, for instance, lawyers need to be part of the overall system. Evidence cannot only be provided by technical components, but also by humans (e.g., Susy saw that the cat tipped the vase over) or knowledge of the world (e.g., there was an earthquake).

Submitted to:
CREST 2016

Generalizing from this example, we realize that accountability is relevant in many domains. As the word suggests, it is relevant in financial accounting. It is relevant when building road vehicles according to ISO 26262, or when building high security systems according to the common criteria. It is relevant when running systems that are governed by standards like ITIL [2], COBIT [10], or HIPAA [15]. In practice, accountability is implemented in airplanes with voice recorders, or black boxes, according to JAR-OPS. While this list is not meant to be comprehensive, it shows that accountability is comprehensive. It is concerned with safety, security, and privacy, as well as the adherence to laws, regulations, and standards. Understanding accountability may hence appear to be a daunting task. We believe, however, that the conceptualization presented in this work is sufficiently rich to capture all these domains. At the same time, however, we are fully aware that in such complex environments "We do not know." may sometimes be the best answer that an accountability mechanism can provide.

In our considerations we focus on accountability for cyber-physical systems that operate in the real world, and we pay attention to both technical and social components. Indeed, from our point of view accountability is especially relevant for systems that interact with humans.

2 Related Work

In computer science, the term *accountability* was popularized by the seminal paper by Weitzner et al. [16]. They introduced accountability as a way to ensure privacy and contrasted it with the traditional approach of achieving privacy through information hiding. In Section 5.2 we show how their work is an instance of our proposed model. Following Weitzner et al., research focused on accountability as a privacy mechanism and as a property of cryptographic protocols. Küster et al. [11] provide a definition of accountability that links accountability to verifiability. With the rise of cloud computing, privacy concerns were naturally applied to the cloud and accountability was seen as a way to ensure data security, e.g. by Pearson et al. [14]. Suggesting accountability as a mechanism to ensure privacy was also the result of the PATS EU project and is detailed in a book edited by Guagnin et al. [6]. Xiao et al. [17] provide a survey of accountability in computer networks and distributed systems. Finally, Papanikolaou and Pearson [12] summarize the understanding of the term accountability in different fields and contexts.

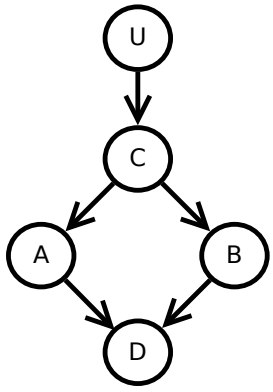

Figure 2: Firing squad example [13].

In contrast to aforementioned literature, we see accountability not merely as a mechanism to protect privacy or as an extension to non-repudiation. We believe that accountability aids us in (i) understanding a system, (ii) improving a system, and, most importantly, (iii) attributing actions of a system to responsible entities (people or organizations). We further believe that accountability mechanisms can be generalized to be applicable to multiple domains.

Causality was given a mathematical foundation by the works of Pearl [13] and his collaboration with Halpern [8, 9], which resulted in the so-called "Halpern-Pearl Definition of Causality". Their definition originates in their earlier work on Bayes networks (which can be also be used for inference, learning and error diagnosis). Halpern published a modified version of their definition recently [7]. They base their model of causality on *counterfactuals*. The idea of counterfactuals is that an event A is the cause of another event B, if B would not have happened without A. While this definition is sufficient in many cases, there is a host of examples in the literature where it fails; see Halpern [7] for an overview. A comprehensive criticism of using counterfactuals for causal reasoning was published by Dawid [3].

Figure 2 is the graphical representation of an example given by Pearl [13]. In this example, the court (U) orders the execution of a prisoner. The order to shoot is given by a captain (C) and carried out by two rifleman (A and B). The final result is that the prisoner dies (D). Pearl and Halpern show how such a model can be constructed and used to reason about causes and possible alternative worlds. Their model of causality allows us to reason about sentences like "If the prisoner is dead, then the prisoner would be dead even if rifleman A had not shot."

Recently, Gössler with Le Métayer and Astefanoaei respectively [5, 4] applied the idea of causal reasoning with counterfactuals to the fault analysis of real-time and embedded systems. They model system components as timed automata and use execution traces to blame individual components.

Concepts similar to accountability can be found under different names, e.g. fault localization or error tracing, and are often summarized under the term *dependable computing*. An overview and a taxonomy of these terms was published by Avivzienis et al. [1].

3 Preliminaries

Before providing our model of accountability in Section 4, this section introduces key concepts that are central in the accountability context.

3.1 Means to Provide Accountability

Getting back to our example (Figure 1), we want to understand which means, both technical and non-technical, exist to find out the cause of the vase's destruction. Later, the findings of any such means will be considered in the process of assigning responsibility.

3.1.1 External Observation Systems

In general, an external observation system is a system that is passive regarding the events under investigation (i.e., does not influence them). It merely records events in a log. We realize that in practice a definition of "external" necessitates a clear description of system boundaries.

In our example, an external observation system providing accountability could be a person watching the room and telling us what has happened; or a video surveillance system that records everything that happens. However, both approaches might be insufficient: a person may lie (on their own accord or acting under duress), while the video system's recordings might be tampered with. None of these two means might be able to sense the light earthquake leading to the vase's destruction. Voice recorders in airplanes are another example for such an external observation system.

Another problem with external observation systems is that in many cases they are simply inexistent or hard to build. Hence, we can not just assume their existence. While software running on a computer, for example, can be monitored by the operating system, a cyber-physical system, such as a diving robot or a long-distance truck, will often act autonomously and without any possibility of direct observation.

3.1.2 Internal System Logs

To provide accountability, technical systems may keep some form of log. In such a log (or log book), all observable and (potentially) relevant events are recorded. Historically, logs for computer systems are collected in unstructured text files that seldom follow a common standard or convention. The log entries

are usually driven by the needs of developers and administrators and as such they are not necessarily useful to understand the cause of a system's behavior.

In the vase room example (c.f. Section 1), the cleaning robot might have recorded a collision at a certain time, the reason being that one of the robot's bumpers (c.f. Section 6) was triggered and consequently a corresponding event was logged. However, with only this knowledge we could not tell if the robot collided with a wall (which is intended behavior) or with the vase. An expert might be able to analyze the logs and tell that collisions with walls are usually either around 30 seconds apart (the time it takes the robot to cross the room) or 5 seconds apart (when the robot navigates a turn). If, on the other hand, a collision is recorded after 15 seconds, then this observation could suggest that the robot bumped into something unexpected—likely the vase. This, however, may not be the only explanation. The robot could also have bumped into the cat, which in turn got scared and jumped into the vase. As a result, we realize that while recording the triggering of the robot's bumpers does provide some essential knowledge, this information might not be sufficient to reason comprehensively about the cause for the vase's destruction. Recording further information, such as the robot's position every five seconds, might have provided sufficient information.

To improve the utility of logs, we suggest that logging mechanisms used as part of an accountability mechanism should be both *standardized* and *extendable*. Standardization (e.g., using the same date format and the same taxonomy of events) allows for logs from different technical systems to be correlated and analyzed using different (standardized) tools. Extendability of logging mechanisms is another important requirement, because it will never be possible to enumerate all (potentially) relevant events at design time. E.g., in our robot example the relevance of the robot's position might not have been considered beforehand. Hence, every cyber-physical system's logging facility should be designed with extendability in mind. This is particularly important because the system's actual software might not be updated frequently and hence updating its logging mechanism out-of-band seems to be an adequate alternative approach. In the example at hand we could possibly further log the resistance that the bumper met. Colliding with a solid wall will likely cause a harder impact than colliding with a fragile vase (which will quickly give way) or the light touch required to startle a cat. Although we might not be able to find the cause (in an automated manner) the first time, an extendable logging facility can make sure that we can determine the cause in similar situations in the future.

Note that by definition it is challenging to imagine how unanticipated events can be handled at the design time of the system. Interestingly, combining internal system logs with external observation systems as described in Section 3.1.1 might be able to address this problem, since the latter will in many cases also record certain types of unanticipated events.

The recordings of both, external observation systems and internal system logs, will be considered in the process of assigning responsibility. Thus, it seems natural to require these recordings to be sound (recorded facts have actually happened), complete (facts that have happened and that should be recorded, have actually been recorded), and tamper-proof (recordings should not be modifiable in hindsight, be it by fire, water, or malicious adversaries). In addition, we require the process of event recording to be efficient, meaning that the act of logging does not impair system usability.

3.2 Assigning Responsibility

The answer to the second question —Who is responsible for breaking the vase?— should name a person or entity that started the chain of events that ended with the vase breaking. Unfortunately, this answer cannot always be found and will often be ambiguous. If the vase was shattered by an earthquake, no one

is directly responsible. If the cat entered the room through an open window, the person who opened the window might be considered responsible. However, when considering the technical system consisting of the robot, then there are multiple entities that might be responsible.

First, the owner of the robot might have used the robot against its specifications, e.g., if it might only be used in rooms with no fragile goods around. If, however, the robot was meant to be used in rooms with vases in them, then the manufacturer may be responsible for not adding a "vase detection mechanism". In such a case it might be of further importance *which* parts of the robot caused the destruction, i.e., which (combination of) software or hardware module(s), eventually assigning responsibility to the provider of the corresponding module(s) or their integrator. Should the robot have been compromised by an attacker, then the attacker is the root cause of the destruction. However, in many real world cases it is impossible to trace the attacker and assign blame to him. Thus, the question remains whether the manufacturer developed the robot in correspondence with state-of-the-art security practices. If this is not the case, then the manufacturer might be held responsible.

In order to simplify the process of finding causes and assigning blame, an accountability system should be augmented with a causal reasoning system that helps an investigator parse the evidence created by logs. It should offer explanations and highlight possible causes and suspects.

3.3 Liability

These causal explanations can then be used to determine the legal liability for an event. Legal proceedings cannot be mechanized and will always require people to judge their peers. Aiding the legal process is the ultimate goal of any accountability system. Legal liability will not always be straightforward to determine. In the case of an malicious adversary hijacking the robot and breaking the vase, the owner of the robot might be to blame for not patching the robot; alternatively, the manufacturer might be to blame for not securing the robot better. In extremis, these questions will be answered in a legal trail. In such a trail the accountability mechanism should provide evidence that can be interpreted by the judges themselves or at least aid expert witnesses in their role as interpreters.

Based on these considerations, Section 4 presents our understanding of an accountability model.

4 An Accountability Model

As motivated in Sections 1 and 3, accountability mechanisms ought to explain why a certain behavior was (not) observed in the real world. Hence, our model aims to infer such explanations from real world observations. At the core of our model of accountability lies the *causal model* in the tradition of Halpern and Pearl as described in Section 2. Figure 3 depicts our high level understanding of such an accountability mechanism. The *causal model*, which is detailed in Section 4.2, will use knowledge about reality in the form of *laws* and *facts* to provide *explanations*. The knowledge originates from four different sources: (i) the *system model* represents (technical) knowledge of our system, (ii) the *world model* represents independent laws of the real world, (iii) the *log* represents facts (or events) that have happened, and finally (iv) the *background* represents everything we do not know. While modeling the background might seem dubious at first, it makes explicit that there are events and causes that have not been thought of. Most importantly, this background will affect the *causal model* by causing spurious relationships.

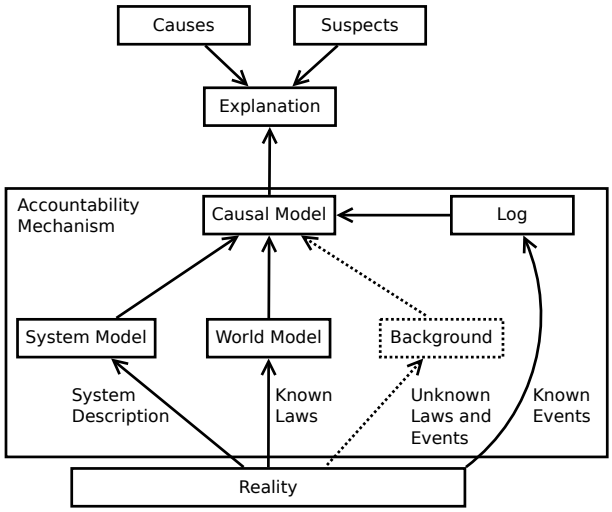

Figure 3: Our high level overview of accountability: Arrows read as "is represented in".

4.1 World Model and System Model

The world model captures "universal rules" that hold regardless of the system being modeled. Such rules include gravity, the knowledge that objects cannot move faster than the speed of light, or the idea that time moves forward. In contrast, a (technical) system model captures rules that (are expected to) hold within the observed (technical) system. Such rules include knowledge about the technical system altogether, its single hardware or software components, or the environment it acts in. In our example, such knowledge might include the robot's maximum speed, its battery's capacity, or the knowledge about the presence of a window in the room to be cleaned. Admittedly, distinguishing between the world model and the system model will not in all cases be trivial. Ideally, both of these models should be "complete", meaning that in no case any knowledge required for reasoning is missing.

Being aware that such completeness is likely to be unrealistic in practice, these models should be made as detailed as possible. The gap between reality and the corresponding models, i.e., everything we do not know or have not considered upfront, is called *background* (c.f. Figure 3) or *exogenous variables* and may still affect the causal model, e.g., via confounding (i.e., cause spurious relationships). Intuitively, the more knowledge is considered in the world model and the system model, the better will the quality of the reasoning provided by the causal model be. The fact that initial world models and system models are likely to miss relevant laws, requires them to be extendable/updatable.

4.2 Causal Model and Reasoning

The causal model is an "instance" of the world. It incorporates the ("always-true") rules from the world and system model (c.f. Section 4.1) and logs, i.e., recordings of events of actual system runs (c.f. Section 3.1.2). As described, one limitation of such a log-based approach is that only expected and observable (which is subject to the logging mechanism being used) events can be logged. Unexpected events will not be logged and enter the causal model as background variables. To improve the logs over time, we require that the logging mechanism can be updated during the lifetime of the system.

We base the construction of causal diagrams on Pearl [13] and use the following procedure to construct the causal diagrams for our instantiations in Sections 5 and 6:

1. Draw a node for every fact in the log.
2. Connect all nodes that have an observed causal relationship with a solid directed edge. These are the causal relationships we have observed according to the logs.
3. Connect all nodes whose causal relationship follows from the laws in the world and system model with a dashed directed edge. These are the causal relations that should have occurred.

Once the causal model has been constructed, we can use it to reason about what has happened in reality. Unsurprisingly, this *reasoning* is then based on the world model, the system model, and the provided system logs. Eventually, such reasoning will generate *explanations* consisting of *causes* and possible *suspects* (c.f. Figure 3). Causes are reasons for an event to happen, while suspects are concrete entities, i.e., persons or organizations, that are to blame. To date, we are unsure whether, and, if so, in which cases, such reasoning and blaming can be fully automated; the possible degree of automation depends to a large degree on the available knowledge base. Hence, we also do not necessarily expect our system to find unique causes. Instead, our primary goal is to create a list of "candidate" causes and suspects, aiding humans in understanding the system's behavior. For example, we envision the user of this information to be a judge trying to settle a case or a developer tracking down a bug. We provide concrete examples for such reasoning in Sections 5.1.4 and 5.1.5.

5 Instantiations

This and the following section show how our accountability model is able to capture security, privacy, and safety requirements. An instantiation for a security case focusing on an Unmanned Aerial Vehicle (UAV) is provided in Section 5.1, while Section 5.2 gives an intuition on how to initialize a privacy case based on the idea of "Information Accountability" [16]. Section 6 shows how a safety case for our real-world research platform could look like.

5.1 Security

Our security use case features an Unmanned Aerial Vehicle (UAV) which enters restricted airspace, e.g., if it gets too close to an airport or military site. In such an event we want to know *why* the UAV crossed into forbidden airspace. It could either have been steered there by the pilot, by a malicious third party or due to a technical failure, such as wrong location information. The accountability mechanism ought to help us discern the true cause.

5.1.1 System Model

When modeling UAVs, usually two intertwined systems are considered: the Unmanned Aerial System (UAS), containing the pilot and any communication equipment, as well as the UAV (Figure 4). The UAV, as well as parts of the UAS, are technical systems and can be modeled accordingly. In our scenario we consider a human operator who interacts with a control system. We do not model the control system in detail, but imagine a remote control that sends commands to the UAV and receives telemetry data (e.g., height) from the UAV. The control system could also contain a receiver for a live video feed from the UAV to enable "first person flying". A real-world example would be the Futaba controls used by AscTec[1] to control their UAVs. Figure 5 depicts a simplified state machine that describes the UAV's behavior.

[1] `http://www.asctec.de/en/uav-uas-drones-rpas-roav/asctec-falcon-8/mobile-ground-station/`

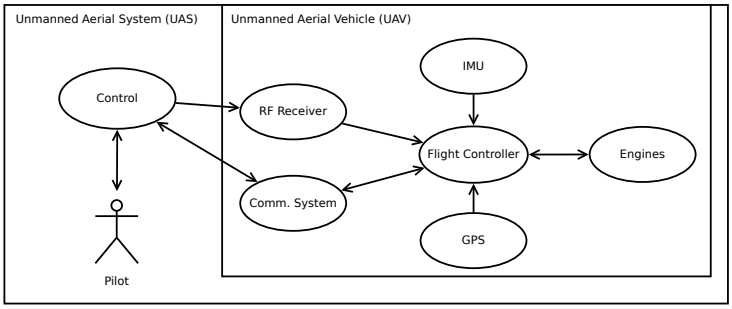

Figure 4: A high-level system model for an Unmanned Aerial System. Arrows represent information and command flows.

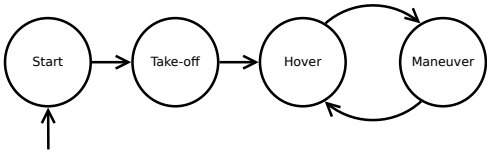

Figure 5: A state machine modeling the UAV's behavior.

In the model of the UAV, the RF receiver will simply receive commands from the remote control and pass them on to the flight controller. The communication system can be a Wi-Fi connection to stream a live video feed, or a low bandwidth XBee link to transmit minimal telemetry data (e.g., height and GPS position). UAVs usually have at least two main sensors: a GPS receiver to determine their position in 3D space (often missing in cheap "toy UAVs") and an Inertial Measurement Unit (IMU) that determines the aircraft's bearing and can be used to estimate a craft position and orientation (dead reckoning). Additionally, a UAV might be fitted with video cameras, barometers or laser scanners to independently measure their height above ground level, and a host of other sensors. All this data is usually processed at the central flight controller which then in turn controls the engines. In case of a quadrocopter, for example, the flight controller will keep the UAV hovering at a fixed position and height when it receives no commands from the remote control.

From this model we can then infer basic causal connections within the system. We may, for example, want to figure out whether the pilot was responsible for steering the UAV into the forbidden airspace. If the pilot has not given any such steering command, our next "most likely" cause would be a malfunctioning remote control. If we can also rule out the remote control, we need to continue our investigation down the causality path, investigating whether any of the UAVs components malfunctioned. Lastly, also an attacker might have injected corresponding steering commands.

5.1.2 World Model

Among other things, the world model incorporates theories of gravity and motion, as well as a model for meteorological laws. Further, it might incorporate a database of GPS coordinates of restricted airspaces.

5.1.3 Log

The recorded event log should offer as many details as possible, without being too costly in terms of runtime performance overhead and storage space. In this scenario, a comprehensive log could be structured

Timestamp	Component	Parent	Log Message
1	Pilot (P)	-	"I started the drone to take a video of the nearby wood"
2	Control (C)	-	Control started
3	Flight Controller(FC)	-	Startup and System self check
4	RF Receiver (RFR)	FC	online
5	IMU	FC	online
6	Comm. System (CS)	FC	online
7	GPS	FC	online
8	Engines (E)	FC	online
9	FC	FC	all systems green
10	P	-	"starts UAV - full throttle"
11	C	-	send control code START to UAV
12	RFR	-	receive message START
13	FC	RFR	START engines
14	E	FC	Startup - Engines to 5V
15	C	-	FULL THROTTLE (FT)
16	RFR	-	received FT
17	FC	RFR	FT to engines
18	E	FC	Engines at 5V
19	GPS	-	Current position is: 0/0
20	IMU	-	Speed is: 5m/s Height:1m ...
...	...	...	...
600	P	-	"I want to go left"
601	C	-	send command GO LEFT to UAV
602	RFR	-	received GO LEFT
603	FC	RFR	received GO LEFT
604	FC	FC	calculated new Engine speeds
605	E	FC	set left engine 2.5V, others: 5V
606	GPS	-	Current position is: 0/5
607	IMU	-	Speed is:5m/s Height:25m...

Figure 6: Example log for the UAS. The parent field should be filled when messages can carry sender information and indicates which component issued the command.

Timestamp	Component	Log Message
1	RC	Take-off command
2	UAV	All engines at full throttle
3	UAV	no RC commands; hover
4	RC	Go left command
5	UAV	Increase speed in left engine, keep others stable
6	RC	Go left command
7	UAV	already going left, keep going left
(repeat 6 and 7)		

Figure 7: The pilot steering the UAV into the restricted airspace.

similar to Figure 6. This log shows in some detail how a UAV takes off and goes to an altitude of 25m. After reaching that altitude, the pilot steers the UAV to the left. Note that this log contains intentions of the pilot at timestamps 1, 10 and 600. These are non-technical events that are impossible to log using a technical accountability infrastructure. Hence, in this example a technical log was enriched with additional knowledge which was, e.g., acquired by asking the pilot (accepting that she might lie), or by inferring it from other witnesses. As even this simple example shows, creating "sufficiently complete" logs is a non-trivial task.

To demonstrate the need to support different levels of abstraction, in addition we show two highly abstracted logs as they might have been created by a log parser after the UAV was found in restricted airspace. First, Figure 7 shows a log in which the pilot gives a clear command to go left, the direction of the supposed restricted airspace, at timestamps 4 to 7. The log depicted in Figure 8, in contrast, shows how the UAV goes left without any action from the pilot.

5.1.4 Causal Model

The causal model of the UAV scenario, as depicted in Figure 9, is obtained by combining the system model (Section 5.1.1), the world model (Section 5.1.2), and the obtained logs (Section 5.1.3). As de-

```
Timestamp   Component   Log Message
1           RC          Take-off command
2           UAV         All engines at full throttle
3           UAV         no RC commands; hover
4           RC          Go left command
5           UAV         Increase speed in left engine, keep others stable
6           RC          no command
7           UAV         already going left, keep going left
(repeat 6 and 7)
```

Figure 8: The pilot gives no command. The UAV should hover, yet still flies into the restricted airspace.

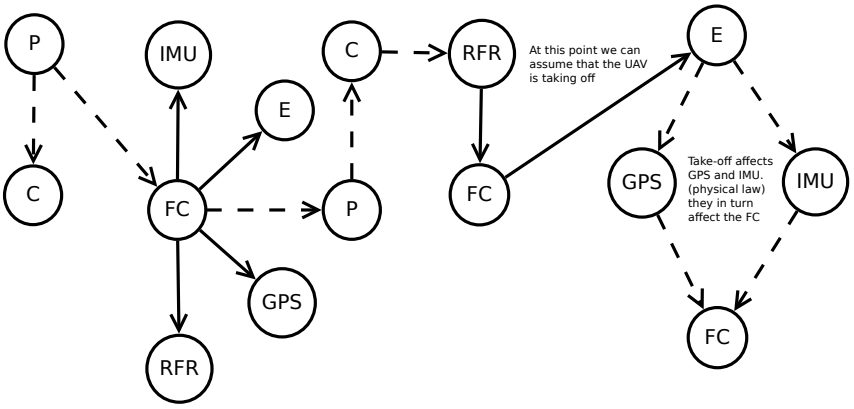

Figure 9: A causal model for the UAS example, encompassing the system model (Figure 4), the obtained logs (Figure 6), and the world model (Section 5.1.2).

scribed in Section 4.2, log entries are represented as *nodes*, while the system and world models' laws constitute the *directed edges*.

Examining the model, we realize that it "ends" after the GPS and the IMU sent messages to the flight controller (Timestamp 20 in Figure 6; right hand side of Figure 9). If we correspond the nodes with the log (Figure 6), we observe that the last message of the IMU indicated a height of 1m and a speed of 5m/s. This allows us to conclude that the UAV is indeed airborne. Following the arrows back to the origin, we can conclude that the pilot is the sole cause for take-off. Put as a counterfactual: had the pilot not started the UAV, it would never have taken off. If Figure 9 would lack an arrow from the pilot to either the control or the flight controller, the causal chain would end at the flight controller. We would then need to contemplate ways in which the flight controller could have started itself, e.g. due to a faulty autopilot.

Figure 10 depicts the (abstracted) causal model in which the pilot steers the UAV into the restricted airspace (the log is shown in Figure 7). Figure 11 is based upon the logs shown in Figure 8 and visualizes the fact that the UAV flew into the restricted airspace on its own accord and that the pilot did not give

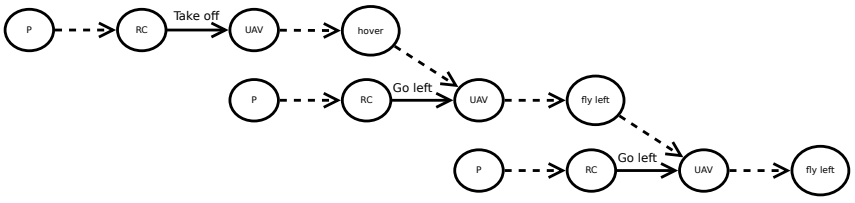

Figure 10: A causal model in which the pilot deliberately steers the UAV into an restricted airspace.

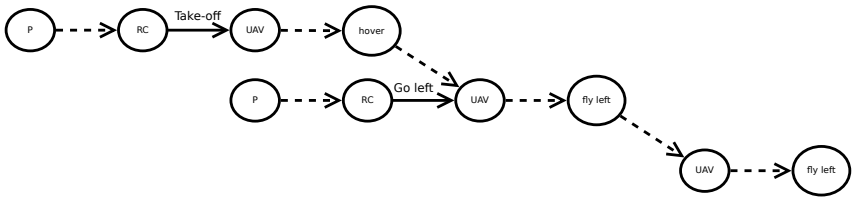

Figure 11: A causal model in which the pilot is not to blame; the fault lies with the UAV.

the respective commands. In the first case, Figure 10, we can see that the pilot repeatedly gives the "go left" command. If we consult our system model, we can see that without this command the UAV would just hover in the same spot. Thus we can assume that, had the pilot not given the command, the UAV would not have ended up in the restricted airspace. In the second example, depicted in Figure 11, we can also see that the pilot started the UAV. So she is at least partly to blame. However, we can also see that she only gives a single "go left" command. Drawing on our understanding of the system, the UAV should have then hovered in place and not crossed into the restricted airspace. We would now need to investigate and find out why the UAV still kept on going left.

5.1.5 Explanation

These causal models can now be used to reason about causes for events and thus create *explanations*. We intend to create causal models based on the definitions by Halpern and Pearl [8, 9], but have not yet formalized our approach on how to (semi-)automatically extract causes from such causal models. Intuitively, we just need to follow the "arrows" back to their origin to find a set of root causes. In the first example (Figure 9) we can see that it was the pilot who started the UAV and thus she is on some level responsible for all events that follow. This can also be seen in Figures 10 and 11: In both cases the pilot starts the UAV. The only difference is that in the latter example the pilot no longer gives commands to go left, the direction of the restricted airspace.

5.2 Privacy

In this section we provide the intuition of how to map the ideas for "Information Accountability" proposed by Weitzner et al. [16] on our model. In their work they discuss ideas to move from the traditional "information hiding" approach for privacy to an "information accountability" approach. Instead of trying to keep information secret (which hardly ever works in the long term), they envision a world in which information is freely shared. However, an accountability system ensures that if information is accessed or used against the owner's intend, such abuse can be detected and traced back to the original perpetrator. To realize their idea, they propose an architecture that is comprised of three components: (i) "Policy Aware Transaction Logs", (ii) a "Policy Language Framework", and (iii) "Policy Reasoning Tools".

Policy Aware Transaction Logs will "initially resemble traditional network and database transaction logs, but also include data provenance, annotations about how the information was used, and what rules are known to be associated with that information" [16]. They can mostly be mapped to our notion of *logs*. As they also include a notion of rules, however, some parts of these logs should naturally be mapped to either our system or world model.

The *Policy Language Framework* will be used to "describ[e] policy rules and restrictions with respect to the information being used" [16]. Just like we do for our world model, they require these policies to

be shareable and extendable. They propose to use ontologies and semantic web technologies to merge these rules. This framework can easily be mapped to our notion of world and system model.

Policy Reasoning Tools are then used to "assist users in seeking answers to questions [regarding their data]" [16]. This concept can be mapped to our notion of a causal model.

6 Research Platform

To study how useful and capable the proposed accountability models and mechanism are in a real world setting, we plan to evaluate our results using a simple cyber-physical system. In line with our example, we chose a vacuum cleaning robot, thus being able to investigate the core problems of accountability for cyber-physical systems in a single use case. Additionally, we consider such a system a good proxy for more complex real-world cyber-physical systems since it (i) moves in the physical world, (ii) interacts with people and objects, (iii) can operate autonomously, and (iv) can be remotely controlled. In the future we plan to build an additional research platform based on a UAV.

6.1 Technical Robot Description

We base our platform on an iRobot Create (Figure 12), the development version of the popular Roomba robotic vacuum cleaner[2]. This robot features several bumper sensors that detect if the robot collides with an object, as well as cliff sensors that detect if the robot is about to fall off a cliff. Further, we equipped the robot with a Full-HD webcam (including a H.264 video encoder) as well as a microphone. To be able to remote control the robot, we further added a battery-powered RaspberriPi and a USB WiFi adapter. Using these components, control commands as well as sensor data can be exchanged with any other computer. This setup allows us to simulate existing industrial service robots like the Aethon TUG[3].

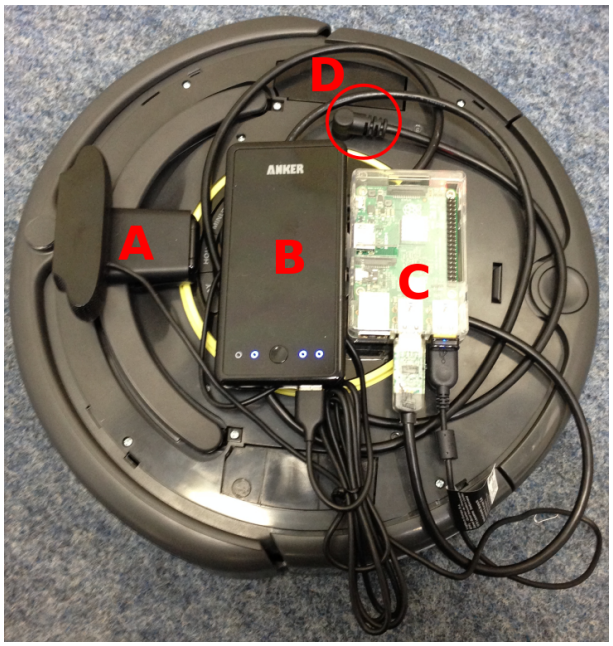

Figure 12: The Create robot: (A) Webcam, (B) Battery Pack, (C) RaspberryPi, (D) Serial-to-USB connector between the robot and the Pi.

Figure 13 provides a high-level technical system diagram: The Logitech C920 camera (1) is used to record audio and video. The camera provides the video as H.264 stream. The microphone is recognized as a normal Linux audio device. Video and audio are transmitted to the RaspberryPi (4) via USB (2 and 3). To navigate the robot (11), the video data is streamed to the command laptop (6) via an RTP UDP Stream over WiFi (5). The robot (11) is controlled via serial commands sent over an USB-to-serial cable (10) connected to the

[2] http://wiki.ros.org/roomba_500_series
[3] http://www.aethon.com/tug/

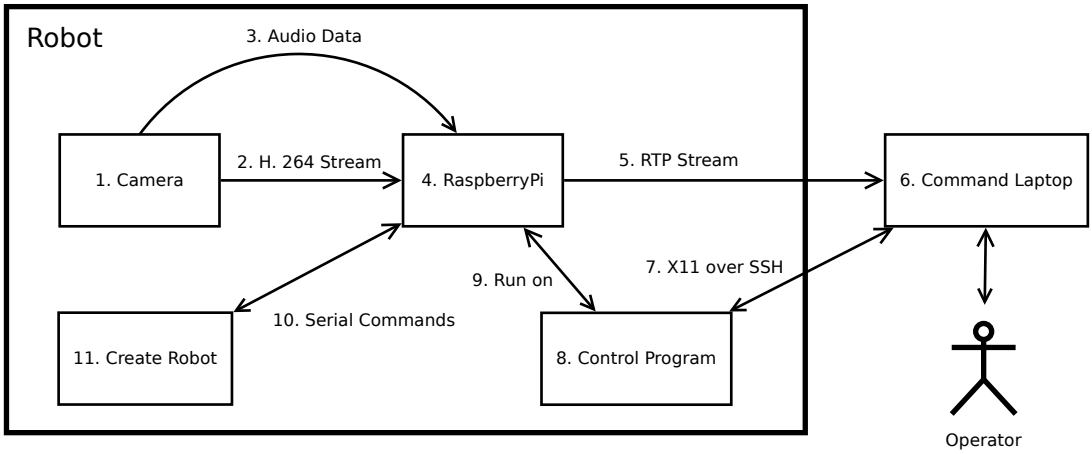

Figure 13: The high-level system model of our iRobot Create research platform.

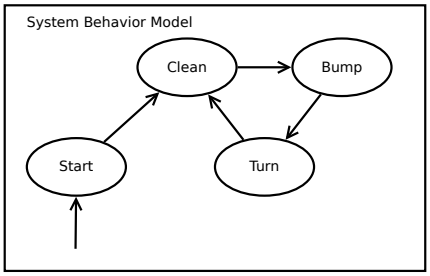

Figure 14: State machine that models the robot's behavior.

Time	Component	Message
0	Operator	"I started my Roomba"
3	Laptop	send start command to robot
5	Robot (R)	START
35	R	BUMP
70	R	BUMP
90	R	BUMP

Figure 15: Example log for the cleaning robot.

RaspberryPi (4). The serial commands are sent by a Python script (8) that runs on the RaspberryPi (9) and is displayed on the Command Laptop via X11-forwarding over SSH (7).

6.2 Model Instantiation

In order to reason about the vase room (Section 1), we extend our system model (Figure 13) with a simple system behavior model (Figure 14). These models are then combined with the log (Figure 15) to yield the causal model depicted in Figure 16. This causal model allows us to infer, that the operator (O) started the robot and that it could complete two "lanes" of cleaning without problems. When going down the third "lane", it bumped against something after 15 seconds. By adding knowledge from the world model (each lane takes 30 seconds plus 5 seconds to turn), we can infer that this unexpected bump is an anomaly. If we now add the (supposed) fact that the windows where closed and no cat could enter the room, then we might conclude that it is highly likely that the robot did indeed break the vase. This conclusion can then be used to decide if the operator is to blame (because she did not remove the vase beforehand) or if the manufacturer is to blame (because they did not add a "vase detection mechanism").

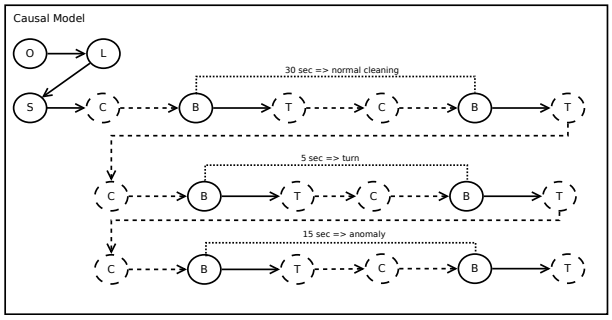

Figure 16: Causal model combining the system models (Figures 13 and 14) with a log (Figure 15).

7 Conclusions and Future Work

We believe that such accountability infrastructures will be useful in many fields and applications. They might even become a (legal) requirement in sensitive or potentially dangerous domains (e.g., data processing or autonomous vehicles). Despite their obvious appeal, we have not seen a wide application of accountability mechanisms. In our opinion the reason for this is the complexity of the task: Causality has only recently been given a clear mathematical definition and reasoning about complex models is a hard problem (state explosion). We hope that designing and developing an accountability mechanism for a real world system will allow us to gain a deeper understanding of the problems and allow us to develop heuristics to keep the complexity of the model checking manageable.

Our next steps will be to "fill the boxes" in our model of accountability with actual functionality, conduct the first real world experiments and gather real data. To this end we are currently in the process of building a system model that contains different layers of abstraction and join it with a basic model of our world (fortunately the world model for a cleaning robot can be kept simple). We are also evaluating different log styles and levels of granularity. As with the state explosion for models, too detailed logs supposedly cost too much performance. We are also not yet sure how to best incorporate external observations (e.g., the observations of a user) into our logs.

We expect efficiency, both in terms of space and time, to be a core problem in any real-world accountability mechanism. If we save (storage) space by not logging some events, these events might be missing in the causal model. If, on the other hand, a logging mechanism slows down a system too much, its usability might be compromised.

Finally, we are also investigating non-technical aspects of accountability: (i) How can accountability be incorporated into the juridical process, and (ii) how can we ensure the accountability of an accountability system (second order accountability)? These questions are especially relevant in the context of autonomous vehicles, the Internet-of-Things, and the increasing autonomy of cyber-physical systems. We need to ensure that such systems confirm with current laws and regulations and, if that is not possible, suggest changes to these laws. To argue our case we need to ensure that accountability systems are safe and that, should something go wrong, we can explain why it went wrong and who should be held responsible. To solve these problems, we are confident that accountability infrastructures provide a promising approach.

Acknowledgments. We want to thank Eric Hilgendorf for first suggesting the vase example and our colleagues Prachi Kumari and Kristian Beckers for valuable discussions on accountability. This research

has been partially funded by the German Federal Ministry of Education and Research (BMBF) with grant number TUM: 01IS12057.

References

[1] A. Avizienis, J.-C. Laprie, B. Randell & C. Landwehr (2004): *Basic Concepts and Taxonomy of Dependable and Secure Computing*. *IEEE Transactions on Dependable and Secure Computing* 1(1), pp. 11–33, doi:10.1109/TDSC.2004.2.

[2] AXELOS (2016): *ITIL — AXELOS*. Available at `https://www.axelos.com/best-practice-solutions/itil`.

[3] Philip Dawid (2000): *Causality Without Counterfactuals (With Discussion)*. *J. Amer. Statist. Assoc* 95.

[4] Gregor Gössler & Lăcrămioara Aştefănoaei (2014): *Blaming in Component-based Real-time Systems*. In: *Proceedings of the 14th International Conference on Embedded Software*, ACM, pp. 7:1–7:10, doi:10.1145/2656045.2656048.

[5] Gregor Gössler & Daniel Le Métayer (2014): *A general trace-based framework of logical causality*. In: *Formal Aspects of Component Software*, Springer International Publishing, pp. 157–173, doi:10.1007/978-3-319-07602-7_11.

[6] Daniel Guagnin (2012): *Managing privacy through accountability*. Palgrave Macmillan.

[7] Joseph Y. Halpern (2015): *A Modification of the Halpern-Pearl Definition of Causality*. AAAI Press, pp. 3022–3033. Available at `http://arxiv.org/pdf/1505.00162`.

[8] Joseph Y. Halpern & Judea Pearl (2005): *Causes and Explanations: A Structural-Model Approach. Part I: Causes*. *The British Journal for the Philosophy of Science* 56(4), pp. 843–887, doi:10.1093/bjps/axi147.

[9] Joseph Y. Halpern & Judea Pearl (2005): *Causes and Explanations: A Structural-Model Approach. Part II: Explanations*. *The British Journal for the Philosophy of Science* 56(4), pp. 889–911, doi:10.1093/bjps/axi148.

[10] ISACA (2016): *COBIT 5: A Business Framework for the Governance and Management of Enterprise IT*. Available at `http://www.isaca.org/COBIT/Pages/default.aspx`.

[11] Ralf Küsters, Tomasz Truderung & Andreas Vogt (2010): *Accountability: Definition and Relationship to Verifiability*. In: *Proceedings of the 17th ACM Conference on Computer and Communications Security*, ACM, New York, NY, USA, pp. 526–535, doi:10.1145/1866307.1866366.

[12] Nick Papanikolaou & Siani Pearson (2013): *A Cross-Disciplinary Review of the Concept of Accountability A Survey of the Literature*.

[13] Judea Pearl (2009): *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press.

[14] Siani Pearson & Andrew Charlesworth (2009): *Accountability as a way forward for privacy protection in the cloud*. In: *Cloud computing*, 5931, Springer Berlin Heidelberg, pp. 131–144.

[15] U.S. Department of Health & Human Services (2016): *Health Information Privacy — HHS.gov*. Available at `http://www.hhs.gov/hipaa/`.

[16] Daniel J Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler & Gerald Jay Sussman (2008): *Information Accountability*. *Communications of the ACM* 51(6), pp. 82–87.

[17] Zhifeng Xiao, Nandhakumar Kathiresshan & Yang Xiao (2012): *A survey of accountability in computer networks and distributed systems*. *Security and Communication Networks*.