

**Integrating Comprehensive Human Oversight in Drone Deployment
A Conceptual Framework Applied to the Case of Military Surveillance Drones**

Verdiesen, E.P.; Aler Tubella, Andrea; Dignum, M.V.

DOI

[10.3390/info12090385](https://doi.org/10.3390/info12090385)

Publication date

2021

Document Version

Final published version

Published in

Information (Switzerland)

Citation (APA)

Verdiesen, E. P., Aler Tubella, A., & Dignum, M. V. (2021). Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones. *Information (Switzerland)*, 12(9), 1-13. Article 385. <https://doi.org/10.3390/info12090385>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright


Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Article

Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones

Ilse Verdiesen ^{1,*} , Andrea Aler Tubella ²  and Virginia Dignum ^{1,2} 

¹ Faculty of Technology, Policy and Management, Delft University of Technology, 2600 AA Delft, The Netherlands; virginia.dignum@umu.se

² Department of Computing Science, Umeå University, 907 36 Umeå, Sweden; andrea.aler@umu.se

* Correspondence: e.p.verdiesen@tudelft.nl

Abstract: Accountability is a value often mentioned in the debate on intelligent systems and their increased pervasiveness in our society. When focusing specifically on autonomous systems, a critical gap emerges: although there is much work on governance and attribution of accountability, there is a significant lack of methods for the operationalisation of accountability within the socio-technical layer of autonomous systems. In the case of autonomous unmanned aerial vehicles- or drones—the critical question of how to maintain accountability as they undertake fully autonomous flights becomes increasingly important as their uses multiply in both the commercial and military fields. In this paper, we aim to fill the operationalisation gap by proposing a socio-technical framework to guarantee human oversight and accountability in drone deployments, showing its enforceability in the real case of military surveillance drones. By keeping a focus on accountability and human oversight as values, we align with the emphasis placed on human responsibility, while requiring a concretisation of what these principles mean for each specific application, connecting them with concrete socio-technical requirements. In addition, by constraining the framework to observable elements of pre- and post-deployment, we do not rely on assumptions made on the internal workings of the drone nor the technical fluency of the operator.

Keywords: comprehensive human oversight; surveillance drones; responsible AI; accountability; autonomous systems



Citation: Verdiesen, I.; Aler Tubella, A.; Dignum, V. Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones.

Information **2021**, *12*, 385. <https://doi.org/10.3390/info12090385>

Academic Editor: Steven Umbrello

Received: 21 July 2021

Accepted: 16 September 2021

Published: 21 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Accountability is a value often mentioned in the debate on autonomous systems and their increased pervasiveness in our society (see Verdiesen, de Sio, and Dignum [1] for an overview). In a narrow sense, it is regarded as a mechanism for corporate and public governance to impart responsibility into agents and organisations. Bovens [2] (p. 450) focuses on this narrow sense of accountability and defines it as follows: “*Accountability is a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences*”. The relationship between an actor and a forum is a key notion in the concept of accountability. In a broad sense, accountability is seen as a virtue and used to criticise or praise the performance of organisations or states regarding policy and decision and their willingness to give information and explanations about their actions [3]. Throughout the literature, the notion of accountability is often framed as a form of backward-looking responsibility [4] and there is much public administration literature on accountability procedures and sanctions that can be imposed when ex-post explanations are inadequate [2,5,6]. However, accountability should not be limited only to scrutiny after an event has occurred: it has also an anticipatory and preventive use to (re)produce, internalise, and adjust norms [1]. Broadly construed, the ability to hold an actor accountable hinges on having

control mechanisms [7] to oversee, discuss, and verify the behaviour of the system to check its alignment with determined values and norms.

When focusing specifically on autonomous systems, a critical gap emerges: although there is much work on governance and attribution of accountability [8,9], there is a significant lack of methods for the *operationalisation* of accountability within the socio-technical layer of autonomous systems [1]. This is particularly salient where autonomous systems are concerned: as executive autonomy is delegated to the system, guaranteeing deployment accountability is a challenge, both in terms of specifications (what does it mean operationally for accountability to be ensured during an autonomous deployment?) and processes (which verifiable behaviours of the autonomous system and the socio-technical system around it guarantee accountability?). In the case of autonomous unmanned aerial vehicles—or drones, as we shall refer to them in the remainder of the text—the critical question of how to maintain accountability as they undertake *fully autonomous* flights becomes increasingly important as their uses multiply in both the commercial and military fields. Although the level of autonomy that should be granted to drones—particularly in the military context—is the subject of debate [10], applications in, e.g., emergency response [11,12] already consider autonomous flights a necessity due to the possibility of failing communication infrastructure or operator unpreparedness. Therefore, we assume in this paper that in-flight communication is not possible and that it is important to implement a monitoring process before and after the flight to ensure human oversight. To the best of our knowledge, there are no other accountability frameworks for human oversight when there is no in-flight contact. We believe that the lack of accountability frameworks when there is an absence of in-flight communication is a gap that needs to be filled.

In this paper, we aim to fill the operationalisation gap by proposing a socio-technical framework to guarantee human oversight and accountability in drone deployments, showing its enforceability in the real case of military surveillance drones. For this purpose, we adapt the Glass Box method of Aler Tubella et al. [13] to provide a monitoring framework for the socio-technical system composed of drone and operator, focusing solely on *observable* constraints on pre- and post-flight processes. By keeping a focus on accountability and human oversight as values, we align with the emphasis placed on human responsibility [14], while requiring a concretisation of what these principles mean for each specific application, connecting them with concrete socio-technical requirements. In addition, by constraining the framework to observable elements of pre- and post-deployment, we do not rely on assumptions on the internal workings of the drone nor the technical fluency of the operator. This paper has a conceptual focus and provides an implementation concept of the pre- and post-deployment observable elements as an illustration of the Glass Box method to ensure human oversight, which is a novel approach.

In the remainder of this paper we first describe related work on accountability and human oversight before describing the Glass Box framework with its interpretation and observation stages. In the following section, we describe our proposed two-stage accountability framework for drone deployment. To illustrate it, we then showcase an initial implementation concept as an example for the real case of military surveillance drones formalised in the discrete-event modelling language given by Coloured Petri Nets (CPNs). Finally, in the conclusion we discuss our findings, limitations of our work, and directions for future work.

2. Background

2.1. Accountability and Human Oversight

To hold an actor (e.g., person, organisation, or institution) accountable an oversight mechanism is required [7,15,16]. This mechanism can be implemented as either an ex-post or ex-ante supervision or as an ex-post review process [17]. Accountability requires strong mechanisms to oversee, discuss, and verify the behaviour of an actor to check if its behaviour is aligned with values and norms.

To ensure accountability over autonomous systems, human oversight is needed and to achieve this. Verdiesen et al. [1] created a comprehensive human oversight framework (CHOF) which connects an engineering, socio-technical, and governance perspective of control to three different temporal phases—before, during, and after deployment of an autonomous weapon system (Figure 1). The *engineering* perspective on control can be described as a mechanism that compares the input and goal function of a system or device to the output by means of a feedback loop to take action to minimise the difference between outcome and goal [18,19]. It holds a very mechanical or cybernetic view on the notion of control which is not well suited to making sense of the interaction between a human agent and an intelligent system for which the human is to remain accountable. The *socio-technical* perspective on control describes which agent has the power to influence the behaviour of another agent [20]. There is a distinction between ex-ante and ex-post control [16]. Ex-ante involvement in decision making is related to managerial control, and accountability-based control is linked to ex-post oversight. Control from a socio-technical perspective is power-oriented and aimed to influence behaviour of agents making use of ex-ante, ongoing, or ex-post instruments [21]. However, it does not explicitly include mechanisms of power over nonhuman intelligent systems, like autonomous systems. The *governance* perspective of control describes which institutions or forums supervise the behaviour of agents to govern their activities. Pesch [22] argues that there is no institutional structure for engineers which calls on them to recognise, reflect upon, and actively integrate values into the designs on a structural basis. Engineers rely on engineering ethics and codes of conduct and the use of these proxies for engineering practices reveals that a governance perspective on responsibility and control lacks robust institutionalised frameworks [1]. The CHOF consists of three horizontal layers that represent the engineering, socio-technical, and governance perspective on control. The vertical columns of the CHOF (x-axis) depict the three temporal phases: (1) before deployment of a system, (2) during deployment of a system, and (3) after deployment of a system. On the y-axis the environment is plotted which ranges from more internal to more external to the technical system.

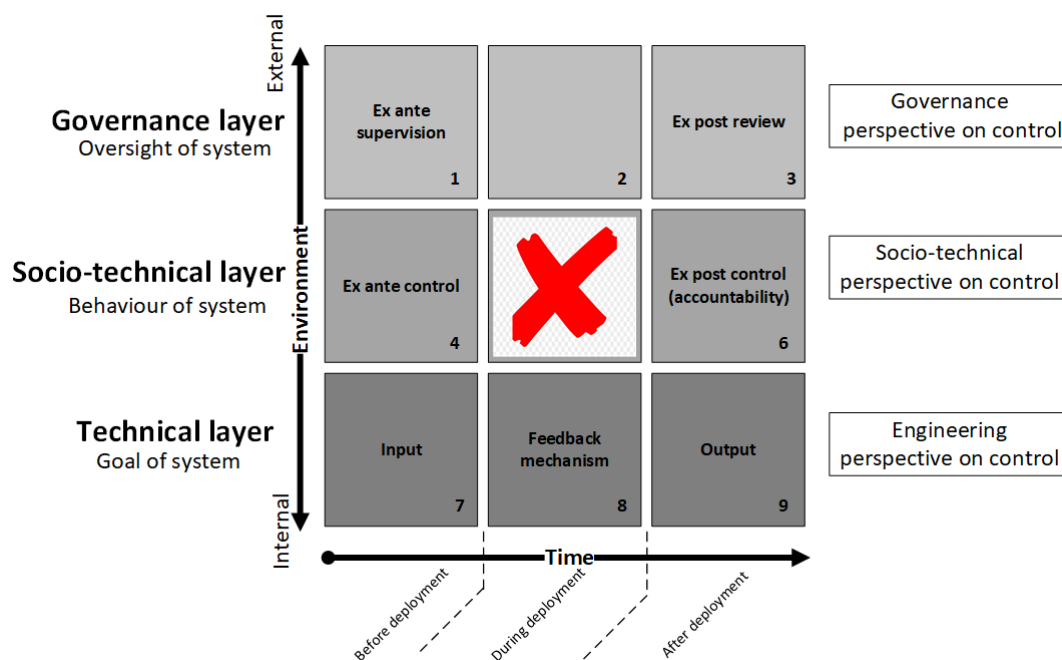


Figure 1. Comprehensive human oversight framework (CHOF).

The CHOF is a comprehensive approach on human oversight that goes beyond a singular engineering, socio-technical, or governance perspective on control. When the CHOF is applied to autonomous (weapon) systems, two gaps in control emerge; one

in the governance layer (block 2 of Figure 1) and one in the socio-technical layer (block 5 of Figure 1). The gap in the governance layer is present as it seems that there is no process, as far as the literature study found, to oversee the system during deployment. It appears that the oversight of the system in the governance layer is conducted before and after deployment by the ex-ante supervision and ex-post review processes, but an oversight mechanism during deployment is lacking. The gap in the socio-technical layer occurs when autonomy is introduced in a autonomous system. Normally an “ongoing control” instrument would occupy block 5, but an autonomous systems has executive autonomy [23] and sets its own means to independently reach the goal a human set for it and autonomously executes its tasks. To fill these gaps a mechanism is needed to monitor the compliance of norms to ensure accountability over autonomous systems. The Glass Box framework could solve these gaps.

2.2. Glass Box Framework

The Glass Box approach [13] is a framework (see Figure 2) for monitoring adherence to the contextual interpretations of abstract values which focuses uniquely on the *observable* inputs and outputs of an intelligent system. Its focus on the observable aspects of the system’s behaviour makes it particularly apt for monitoring autonomous and generally opaque systems.

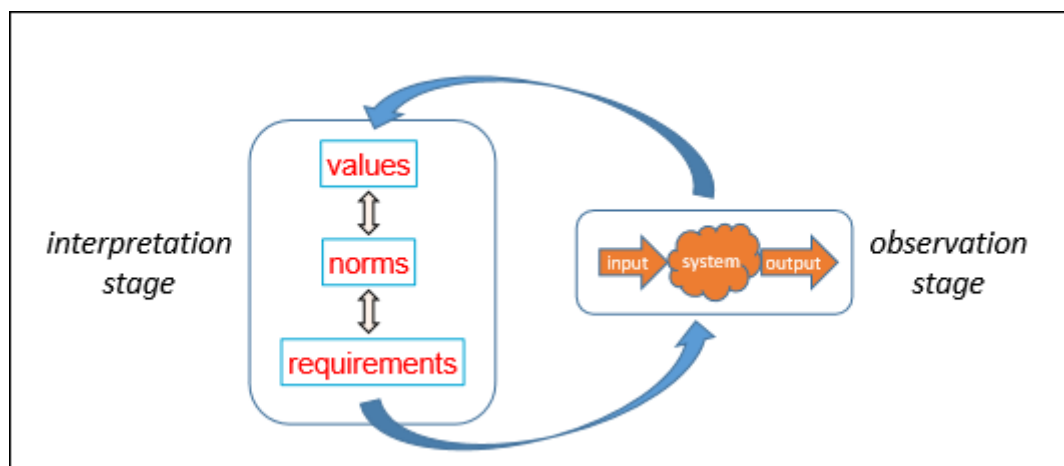


Figure 2. Glass Box framework (as in: Aler Tubella et al. [13]).

The Glass Box approach consists of two phases which inform each other: interpretation and observation. The interpretation stage consists of a progressive process of concretising abstract values into specific design requirements. Following a *Design for Values* perspective [4], the translation from values to requirements is done by considering the different stakeholder interpretations and contexts. The output from the interpretation stage is an abstract-to-concrete hierarchy of norms where the highest level is made up of values and the lowest level is composed of fine-grained concrete requirements for the intelligent system only related to its inputs and outputs. The intermediate levels are composed of progressively more abstract norms, where fulfilling a concrete norm “counts as” fulfilling the more abstract one in a certain context. This hierarchy of norms transparently displays how values are operationalised, together with which contexts have been considered.

The second phase of the approach is given by the observation stage. This stage is informed by the requirements on inputs and outputs identified in the interpretation stage, as they determine what must be verified and checked. In the observation stage, the system is evaluated by studying its compliance with the requirements identified in the previous stage: for each requirement, we assign one or several *tests* to verify whether it is being fulfilled. The difficulty of these tests can range from an extremely simple yes/no check on

whether an action has been performed, to sophisticated statistical analysis depending on the type of norms identified.

Feedback between interpretation and observation stage throughout the lifespan of the system is necessary: continuous observation informs us on which requirements are consistently unfulfilled, which may prompt changes in the implementation or in the chosen requirements. This approach therefore transparently monitors and exposes possible malfunctions or misuse of the system.

3. Framework

Ensuring accountability and adherence to values in the context of drone deployment is inextricably tied to the notion of human oversight and human accountability. For this reason, we propose to consider drone deployment a “process within a socio-technical system”, the monitoring of which includes not only examining the behaviour of the drone itself but also examining human-led procedures in pre- and post-deployment. A specific adaptation of the Glass Box approach to this context is therefore the explicit inclusion of the operator(s) as an entity to which norms can apply.

A significant choice in this framework is the decision to consider the drone a “black box”, the internal logic of which is not accessible. This responds to two motivations. Firstly, relying on access and monitoring capabilities on the internal workings of drones would be a strong assumption, since the proprietary nature of this technology often precludes observation of its software. Second, for auditability purposes, the users of this framework should be able to transparently follow the monitoring process. However, such users, who will respond to the monitoring process, do not necessarily possess the technical background required to understand or check constraints on the internal logic of a drone. Thus, our framework is based on monitoring adherence to norms constraining purely *observable* elements of pre-, and post-deployment. Another choice is that we purposely designed a technology-agnostic approach so that it can be used on many different systems independent from the AI techniques and algorithms that are used as internal workings of the drone. We consider these as part of the black box.

A final adaptation is the explicit call to restrict the specifications and monitoring to pre- and post-flight processes. This choice is due to our focus on autonomous drones: after landing, we can check what has happened during the flight, but during it we assume no contact between the drone and its operator, for example, due to a failing communication structure, an electronic warfare threat, or operator unpreparedness. Of course, if the possibility of in-flight communication exists, expanding the norms to include in-flight behaviour is a possibility.

In what follows, we present an adaptation of the Glass Box approach for the inclusion of human oversight in autonomous drone deployment. The proposed framework includes an interpretation and an observation stage, each discussed in detail.

3.1. Interpretation

The interpretation stage entails turning values into concrete norms constraining observable elements and actions within the socio-technical system. As high-level concepts, values are abstract, whereas norms are prescriptive and impose or forbid courses of action. Such a translation is done by constructing norms progressively, subsuming each norm into several more concrete ones, until the level of norms containing concrete testable requirements is reached. This concretisation of norms will be carried out by all stakeholders involved in the deployment, ideally with legal advisory as well as with participation from operators themselves (whose processes will be subject to the norms identified).

Through a *Design for Values* perspective [4,24–27], concretising values requires carefully adapting to the specific context, as values may take different meanings in different contexts. In the case of drone deployment, the context is made up of two main factors: the context of deployment itself, and the organisation doing the deployment. Thus, some norms may generally apply to any deployment (such as organisational rules), whereas others may

be highly specific (such as regulations governing specific areas or purposes). For this reason, the interpretation stage does not produce a one-size-fits-all normative framework, but rather it needs to be updated in any change of context. The specific tying of norms to a context enforces human oversight in this stage: new human-designed norms are needed for any new context of deployment, thus necessarily implicating the deploying organisation in the process of considering each situation's specificity and risk.

Even though values and their interpretations vary by culture, purpose, organisation, and context, some values are fundamentally tied to the context of drone deployment. As with any technology deployed into society, a fundamental value is that of *lawfulness*. A requirement for any drone deployment is, for example, to respect flight rules (e.g., maximum height of flight and avoidance of airport surroundings). Thus, the identification of requirements for the trajectory taken by the drone is a fundamental aspect of this stage. Given the different capabilities that drones may be equipped with, aspects of the law related to flying over public spaces, commercial liability, or privacy [28], as well as surveillance [29] or warfare, must be considered. The purpose of deployment itself (e.g., humanitarian aid, commercial delivery, or bird observation) will determine the relevant values that guide the process, such as privacy [30], safety [31], humanity [32], or ecological sustainability [33].

Requirements need to refer to the observable behaviour of drone and operator, and are considered in the context of pre- and post-flight procedures. They may apply to checkable behaviours of the drone (flying over a certain altitude or flying over certain areas), to pre-flight processes (getting approval or checking weather conditions), or to post-flight processes (evaluation of route followed or treatment of the data obtained). Crucially, they are not limited to the drones' behaviour, but must include the system around it for human oversight: procedures such as pre-flight safety checks, acquiring authorisations or human review of the data obtained should all be mandated and constrained, so that we can guarantee that the entire flight process has been subject to human oversight.

The norms and observable requirements identified at this stage form the basis for the next stage, indicating what should be monitored and checked, and which actions constitute norm violations.

3.2. Observation

In this stage, the behaviour of the system is evaluated with respect to the values by studying its compliance with the requirements identified in the interpretation stage. As these requirements focus on observable behaviours, in this stage observations are made, and it is reported whether norms are being adhered to or not (and, by extension, whether values are being fulfilled).

Observations can be automated (e.g., automatically trigger a flag if the drone has deviated from its planned path), or manually performed by an operator, depending on the requirement. A specific trade-off to consider is the observation time versus the reliability of the observations: extensive, lengthy manual or computationally expensive checks may take a long time to perform, delaying operations, but may be the only way to check a certain requirement. Depending on how crucial such a requirement is, observations may be relaxed (e.g., performed at random intervals), or the requirement modified for a better fit.

From these observations, we can compute whether norms have been adhered to. Such a computation can be done through a formal representation of the norms and requirements. For example, a formalisation of the Glass Box can be found in Aler Tubella and Dignum [34], using a "counts-as" operator to relate more concrete norms to their more abstract counterparts. Within that formalisation, by assigning ground truth values to a set of propositional atoms through the observations, we can compute which norms have been adhered to, and escalate up the hierarchy of norms to determine which values have been followed in each context. Alternatively, norms can, for example, be expressed in a deontological language [35] and similarly relate to the observations by representing them as ground truths. A different, complementary approach that we showcase in the next section is the use of Coloured Petri Nets (CPNs) as modelling language for the requirements. By adding

tokens to different states depending on the observations (roughly, adding a token if the observation is positive, and not if it is negative), we can simulate the pre- and post-flight processes and determine whether it proceeds correctly or whether norm violations have occurred.

The outcome of the observation stage is either a confirmation that all specifications have been followed, or evidence of norm violations given by the observations that trigger the violation. Human oversight requires that such violations entail accountability processes and a review of the process culminating in the “failed” flight. By providing concrete evidence of where such failures to follow the specifications occurred, this framework therefore explicitly enables oversight without requiring access to the internal logic of the machine, ensuring accountability.

3.3. Glass Box Framework Projected on CHOF

When the two stages of the Glass Box framework are projected on the CHOF, Figure 3 is generated. The Interpretation stage of the Glass Box framework, in which *values* in the governance layer are turned into concrete *norms*, constraining observable elements and actions in the socio-technical layer, which in turn are translated into *requirements* in the technical layer, is done before deployment—visible in the first column of Figure 3. During deployment the behaviour and actions of an autonomous system are monitored in the governance layer and verified in the technical layer in the Observation stage of the Glass Box framework that treats the block in the socio-technical layer as a black box visible in the middle column of Figure 3. After deployment a Review stage is required as an accountability process in which a *forum* in the governance layer can hold an *actor* in the socio-technical layer accountable for its *conduct* in the technical layer—visible in the third column of Figure 3. The outcome of the Review stage should feed back into the Interpretation stage for a next deployment of an autonomous system and thereby close the loop between the stages.

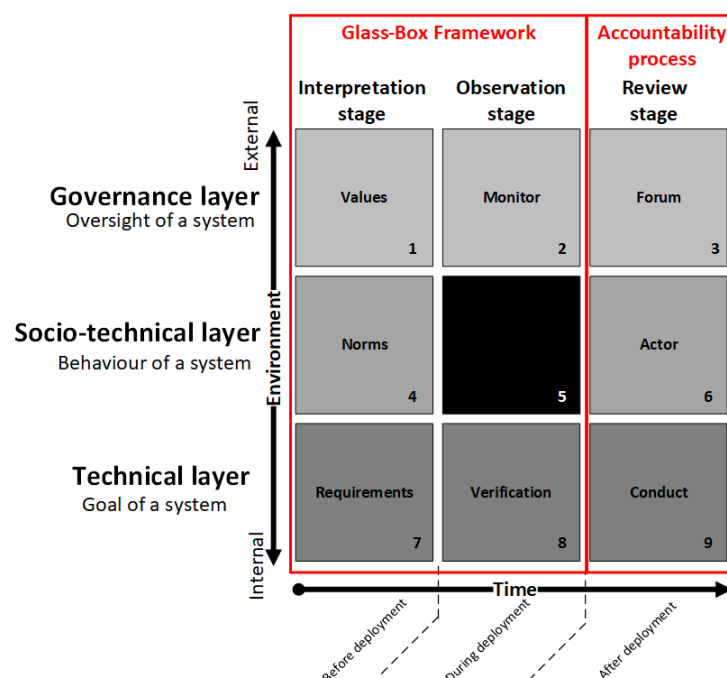


Figure 3. Glass Box framework projected on CHOF.

4. Implementation Concept

To operationalise the Glass Box framework we created an implementation concept as an example to prove that the framework is actionable. The implementation concept is applied to the case of an autonomous military drone. We chose this application area,

because the military domain amplifies the values and norms involved in the decision making in the deployment of an autonomous drone due to the nature of the operating environment, but our choice for this application area should not be seen as an endorsement of autonomous military surveillance drones. In this scenario, the autonomous drone is not weaponised and it flies a surveillance mission over a deployment area to gather intelligence (see Figure 4). To conduct its mission, the weather conditions should be favourable otherwise the camera will not record images. In addition, the drone should have a map in order to calculate its flight path. In this particular scenario it should remain within its Area of Operations and avoid certain areas, such as restricted operating zones and an electronic warfare threat.

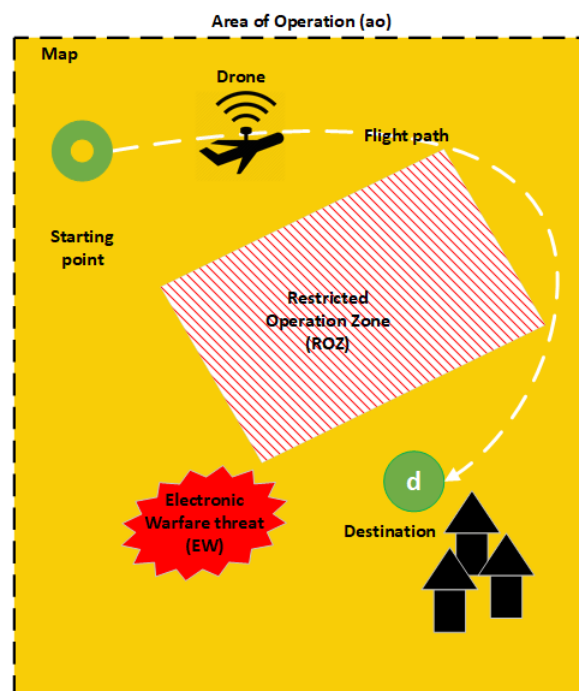


Figure 4. Visualisation scenario.

In the first stage of the Glass Box framework (see Figure 3) the norms are derived from values before drafting (technical) requirements. Our implementation concept is based on existing operational norms within the Dutch Ministry of Defense, for example, rules of engagement, which are already discussed before the deployment of a mission. Therefore, value elicitation is out-of-scope for our implementation concept for now. One of the norms that is identified in the interpretations stage of our scenario is that the flight path should not cross a Restricted Operating Zone. Another norm is that the electronic warfare threat should be avoided. The third norm is that the surveillance drone should remain within the Area of Operation. These norms are input for the requirement of the drone's flight path. After the mission, the norms are observed by manually evaluating the flight path to check if the autonomous drone stayed within the Area of Operation and did not cross the Restricted Operating Zone and electronic warfare threat. Violation of the norms is reported in the debrief report which is part of the review stage of the accountability process.

4.1. Coloured Petri Nets

We created an implementation of a pre- and post-flight procedure as an example using Coloured Petri Nets (CPNs) as modelling language. CPNs is a discrete-event modelling language for modelling synchronisation concurrency and communication processes. The language consists of *states* and *events* and a system can change a state. We used CPN Tools to create a simulation that allows us to check the model and run a simulation-based performance analysis [36]. We created a model that shows several steps of a pre-flight

mission planning and post-flight mission evaluation process for autonomous drones which is not too complex as an example. We based the processes on the scenario described in the previous subsection. As reference we used information obtained in several conversations with domain experts in the Dutch Ministry of Defense and the JFCOM-UAS-PocketGuide-the US Army Unmanned Aerial Systems manual [37]. The CPNs are uploaded as Supplementary Materials (<https://github.com/responsible-ai/DroneCPN>, accessed on 15 September 2021).

4.2. Pre-Flight Mission Planning Process

In the pre-flight mission planning process first the steps are modelled to check the prerequisites for a mission; i.e., the availability of a *map* and the status of the *weather conditions* (Figure 5). Next the compliance criteria *Area of Operation*, *Restricted Operating Zone*, and *Electronic Warfare Threat* are checked and if these are complied with, the *flight path* is calculated. If, for example, the boundaries of the Area of Operation are not known and this criteria is not complied with, then the process enters a feedback loop in which the boundaries of the Area of Operation are requested. When all criteria are met the *approval process* is triggered and sequentially a drone is requested. In the case that the mission is not approved, the reason for disapproval needs to be solved first in order to continue the process. The pre-flight mission planning process is modelled with several feedback loops. For example, if there is no map available then a map is requested or if the weather conditions are adverse than the mission is replanned (Figure 5). In the final step the *mission* is flown and, upon completion of all the steps, the pre-flight mission planning process ends and the drone is returned to the pool of drones and can be deployed for a next mission.

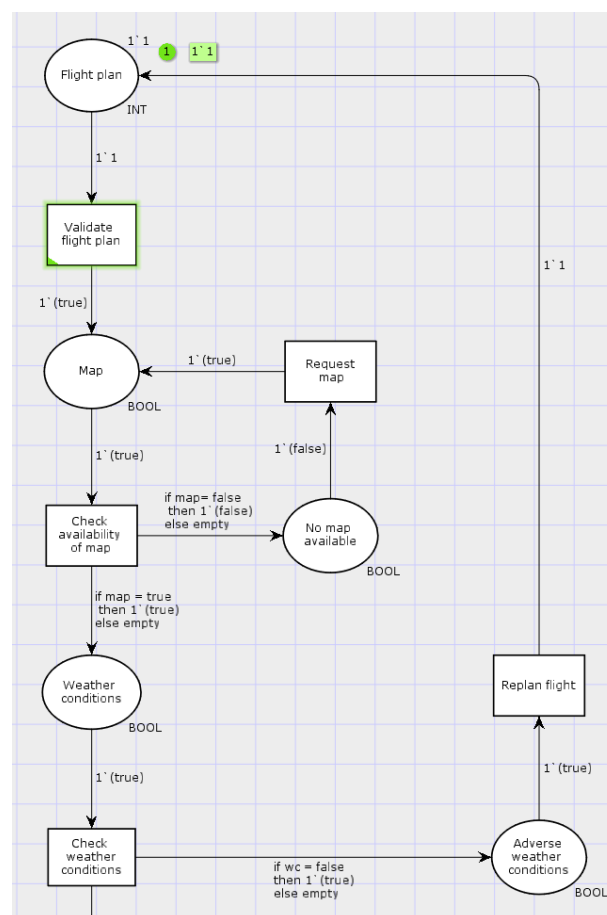


Figure 5. Top part of mission planning process.

4.3. Post-Flight Evaluation Process

The evaluation of the mission will be done manually for now and starts with two concurrent steps. The check of (1) the *compliance criteria* and (2) the *flight path*. The same compliance criteria as in the pre-flight mission planning process are checked (Figure 6); *Area of Operation*, *Restricted Operating Zone* and *Electronic Warfare Threat*. If the criteria, for example, “avoid *Restricted Operating Zone*”, is met, the process passes to the next stage. If the *Restricted Operating Zone* is crossed, the criteria is not met and this norm violation will be noted in the debrief report. Concurrent to this step, the compliance with the *flight path*, or deviation of it, will be checked. Both compliance with the criteria and the flight path as noncompliance will end up in the debrief report. Noncompliance comments can be used as lessons learned for the next mission. The draft of the debrief report is the final step of the post-flight evaluation process and this evaluation can be used in the review stage of the accountability process.

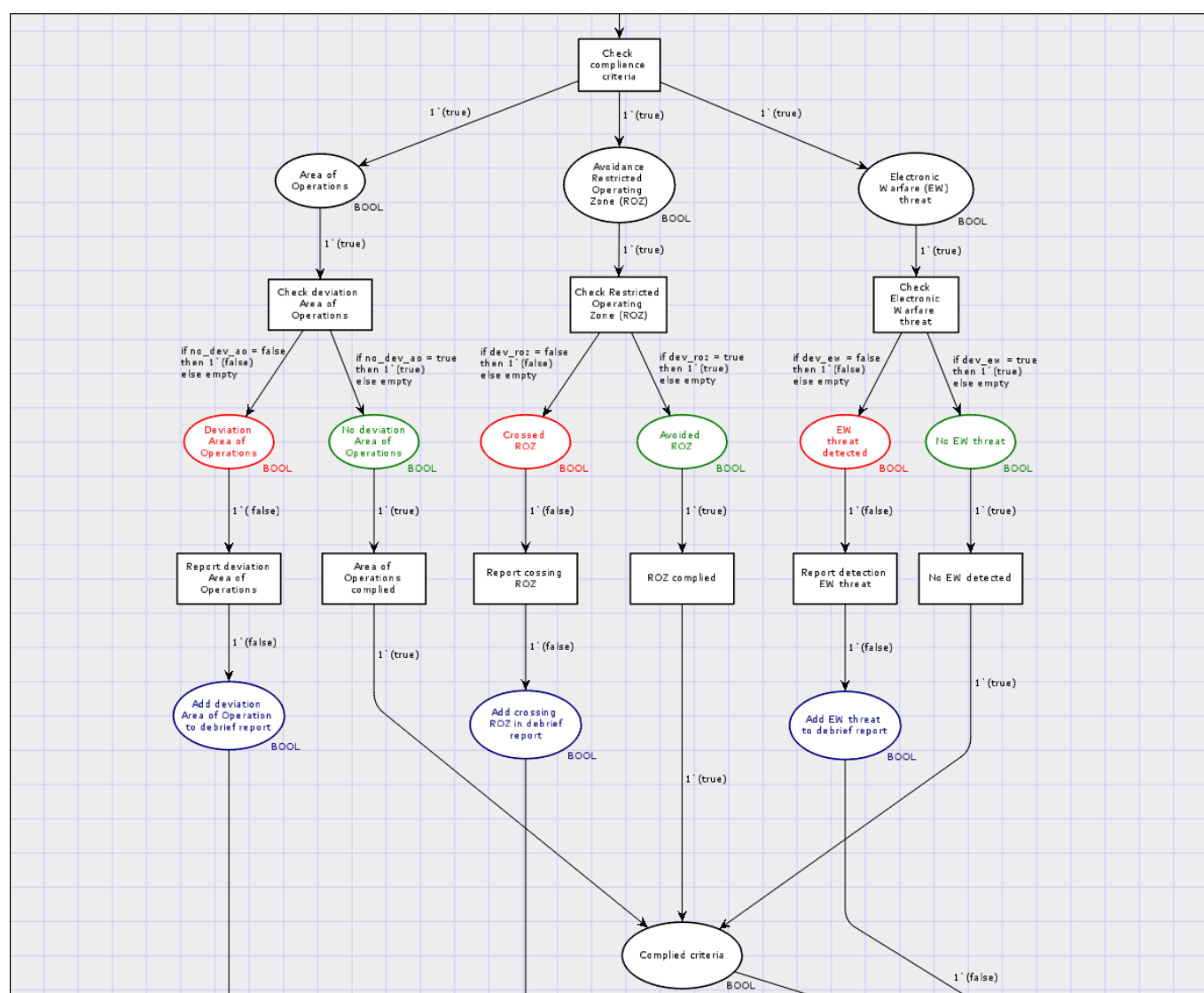


Figure 6. Criteria compliance check post-flight in evaluation process.

5. Conclusions and Future Work

The implementation concept in the previous section shows that it is possible to set criteria in the pre-flight process and to evaluate these criteria post-flight. During flight, the drone itself is treated as a black box of which the internal logic is not accessible. Although being a toy example, it demonstrates that a monitoring process can be designed to

guarantee human oversight where the users set norms—or criteria, in this example—for input and observe and evaluate the output against the input to check for noncompliance of the norms. Deviations of the norms will be reported and can be used to update the norms in a new scenario. This way the users do not need technical skills to understand the workings of drone, but still can monitor and oversee the use of the autonomous system based on observable norms. The Glass Box framework that is built around the black box (the autonomous drone) with the *Interpretation*—or pre-flight mission planning process—and the *Observation* stage—or post-flight evaluation process—allows for a transparent human oversight process which ensures accountability for the deployment of an autonomous system.

However, this is a first attempt to implement the Glass Box framework in a practical manner. Further research is needed to evaluate the implementation concept by experts and further consideration is necessary to assess the suitability of this approach. There is much room to extend the model, especially in cases where in-flight contact is possible. Also, we applied it to a case in which there are existing operational norms within the Dutch Ministry of Defense which are already discussed so no value elicitation is done. We are currently setting up an expert panel to conduct the value elicitation for the implementation concept to get qualitative results.

A limitation of our approach is that the model of the pre-flight and post-flight process does not encompass all steps of mission planning and evaluation for autonomous military drones, but it can be extended and/or adjusted if needed or when the context changes. Additionally, the implementation concept is applied to a very specific use case—that of military surveillance drones—and it is not clear if it will also be applicable to other areas, such as that of other autonomous systems in the military domain, autonomous vehicles, or the medical domain. Another limitation of our approach is that we do not monitor the in-flight actions that the autonomous drone takes, because we assume that in-flight communication is not possible, for example, due to a failing communication structure, an electronic warfare threat, or operator unpreparedness. Therefore, it is not possible to oversee norm violations nor is it possible to intervene during the flight.

Future work will explore the possibility of monitoring the behaviour of autonomous systems during operations, for example, decisions of a drone during its flight, but we will not limit ourselves to drones alone and will include other autonomous systems as well. If norm violations occur during its operation this will impact the safety of the system and its decisions should be monitored and documented in order to account for its behaviour. Another direction for future work is extending the implementation concept to other values such as privacy (for example, during information gathering nearby a village) and other human rights including fuzzier norms. This will be addressed in the next implementation to generalise our approach and to align the implementation concept more to the Design for Values approach. Finally, we hope to apply the Glass Box approach to cases from other domains to see if the implementation concept will be applicable. For example, in humanitarian disaster relief with autonomous drones or in the case of autonomous vehicles. It would be interesting to verify if, based on the Glass Box framework, the user can provide norms and evaluate these to monitor the behaviour of these systems to increase safety and ensure accountability.

Supplementary Materials: The CPNs are available online at <https://github.com/responsible-ai/DroneCPN>.

Author Contributions: Conceptualization, I.V., A.A.T., and V.D.; methodology, I.V.; formal analysis, A.A.T.; domain knowledge, I.V.; writing—original draft preparation, I.V. and A.A.T.; writing—review and editing, V.D.; visualization, I.V.; supervision, V.D. All authors have read and agreed to the published version of the manuscript.

Funding: Aler Tubella and Dignum are supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation.

Data Availability Statement: All data and models are made available as Supplementary Material.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

References

- Verdiesen, I.; de Sio, F.S.; Dignum, V. Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. *Minds Mach.* **2020**, *31*, 137–163. [\[CrossRef\]](#)
- Bovens, M. Analysing and assessing accountability: A conceptual framework 1. *Eur. Law J.* **2007**, *13*, 447–468. [\[CrossRef\]](#)
- Bovens, M.; Goodin, R.E.; Schillemans, T. *The Oxford Handbook Public Accountability*; Oxford University Press: Oxford, UK, 2014.
- Van de Poel, I. Translating values into design requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 253–266.
- Keohane, R.O. *Global Governance and Democratic Accountability*; Citeseer: Princeton, NJ, USA, 2003.
- Greer, S.L.; Wismar, M.; Figueras, J.; McKee, C. Governance: A framework. *Strength. Health Syst. Gov.* **2016**, *22*, 27–56.
- Schedler, A. Conceptualizing accountability. *Self-Restraining State Power Account. New Democr.* **1999**, *13*, 17.
- Pagallo, U. From automation to autonomous systems: A legal phenomenology with problems of accountability. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 17–23. [\[CrossRef\]](#)
- De Sio, F.S.; van den Hoven, J. Meaningful human control over autonomous systems: A philosophical account. *Front. Robot. AI* **2018**, *5*, 28. [\[CrossRef\]](#)
- Horowitz, M.C. The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus* **2016**, *145*, 25–36. [\[CrossRef\]](#)
- López, L.B.; van Manen, N.; van der Zee, E.; Bos, S. DroneAlert: Autonomous Drones for Emergency Response. In *Multi-Technology Positioning*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 303–321. [\[CrossRef\]](#)
- Waharte, S.; Trigoni, N. Supporting search and rescue operations with UAVs. In Proceedings of the In 2010 International Conference on Emerging Security Technologies, Canterbury, UK, 6–7 September 2010; pp. 142–147. [\[CrossRef\]](#)
- Aler Tubella, A.; Theodorou, A.; Dignum, F.; Dignum, V. Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'2019), Macao, China, 10–16 August 2019.
- GGE. *Emerging Commonalities, Conclusions and Recommendations*; Possible Guiding Principles; United Nations: Geneva, Switzerland, 2018.
- Caparini, M. *Media and the Security Sector: Oversight and Accountability*; Geneva Centre for the Democratic Control of Armed Forces (DCAF) Publication: Addis Ababa, Ethiopia, 2004; pp. 1–49.
- Scott, C. Accountability in the regulatory state. *J. Law Soc.* **2000**, *27*, 38–60. [\[CrossRef\]](#)
- Pelizzo, R.; Stapenhurst, R.; Olson, D. Parliamentary Oversight for Government Accountability. 2006. Available online: https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1136&context=soss_research (accessed on 15 September 2021).
- Åström, K.J.; Kumar, P.R. Control: A perspective. *Autom.* **2014**, *50*, 3–43. [\[CrossRef\]](#)
- Pigeau, R.; McCann, C. *Re-Conceptualizing Command and Control*; Technical Report; Defence and Civil Institute of Environmental Medicine: Toronto, ON, Canada, 2002.
- Koppell, J.G. Pathologies of accountability: ICANN and the challenge of “multiple accountabilities disorder”. *Public Adm. Rev.* **2005**, *65*, 94–108. [\[CrossRef\]](#)
- Busuioc, M. Autonomy, Accountability and Control. The Case of European Agencies. In Proceedings of the 4th ECPR General Conference, Pisa, Italy, 6–8 September 2007; pp. 5–8.
- Pesch, U. Engineers and active responsibility. *Sci. Eng. Ethics* **2015**, *21*, 925–939. [\[CrossRef\]](#)
- Castelfranchi, C.; Falcone, R. From automaticity to autonomy: The frontier of artificial agents. In *Agent Autonomy*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 103–136.
- Friedman, B.; Kahn, P.H.; Borning, A.; Hultgren, A. Value sensitive design and information systems. In *Early Engagement and New Technologies: Opening Up the Laboratory*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 55–95.
- Cummings, M.L. Integrating ethics in design through the value-sensitive design approach. *Sci. Eng. Ethics* **2006**, *12*, 701–715. [\[CrossRef\]](#) [\[PubMed\]](#)
- Davis, J.; Nathan, L.P. Value sensitive design: Applications, adaptations, and critiques. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 11–40.
- Van den Hoven, J.; Vermaas, P.; Van de Poel, I. Design for values: An introduction. In *Handbook of Ethics, Values, and Technological Design*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 1–7.
- Rao, B.; Gopi, A.G.; Maione, R. The societal impact of commercial drones. *Technol. Soc.* **2016**, *45*, 83–90. [\[CrossRef\]](#)
- Rosen, F. Extremely Stealthy and Incredibly Close: Drones, Control and Legal Responsibility. *J. Confl. Secur. Law* **2014**, *19*, 113–131. [\[CrossRef\]](#)
- Luppici, R.; So, A. A technoethical review of commercial drone use in the context of governance, ethics, and privacy. *Technol. Soc.* **2016**, *46*, 109–119. [\[CrossRef\]](#)
- Clarke, R.; Bennett Moses, L. The regulation of civilian drones’ impacts on public safety. *Comput. Law Secur. Rev.* **2014**, *30*, 263–285. [\[CrossRef\]](#)

-
32. van Wynsberghe, A.; Comes, T. Drones in humanitarian contexts, robot ethics, and the human–robot interaction. *Ethics Inf. Technol.* **2020**, *22*, 43–53. [[CrossRef](#)]
 33. Vas, E.; Lescroël, A.; Duriez, O.; Boguszewski, G.; Grémillet, D. Approaching birds with drones: First experiments and ethical guidelines. *Biol. Lett.* **2015**, *11*, 20140754. [[CrossRef](#)] [[PubMed](#)]
 34. Aler Tubella, A.; Dignum, V. The Glass Box Approach: Verifying Contextual Adherence to Values. In Proceedings of the AISafety 2019, Macao, China, 11–12 August 2019; CEUR-WS.
 35. Von Wright, G.H. On the Logic of Norms and Actions. In *New Studies in Deontic Logic*; Springer: Dordrecht, The Netherlands, 1981; pp. 3–35. [[CrossRef](#)]
 36. Jensen, K.; Kristensen, L.M.; Wells, L. Coloured Petri Nets and CPN Tools for modelling and validation of concurrent systems. *Int. J. Softw. Tools Technol. Transf.* **2007**, *9*, 213–254. [[CrossRef](#)]
 37. JUAS-COE. *JFCOM-UAS-PocketGuide*; JUAS-COE: Arlington County, VA, USA, 2010.