Comprehensive Human Oversight Framework to Ensure Accountability over Autonomous Weapon Systems

Verdiesen, E.P.

DOI
10.1145/3514094.3539523

Publication date
2022

Document Version
Final published version

Published in
AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society

Comprehensive Human Oversight Framework to Ensure Accountability over Autonomous Weapon Systems

Ilse Verdiesen
Technology, Policy and Management
Delft University of Technology
Delft, The Netherlands
e.p.verdiesen@tudelft.nl

In former research [1], human oversight over an autonomous system was operationalised by proposing a socio-technical framework projecting the Glass Box approach [2] on the Comprehensive Human Oversight Framework (CHOF) (see figure 1). This showed that it is possible to rely on observable elements during the Interpretation stage before deployment and Observation stage during deployment, without having to make assumptions made on the internal workings of the autonomous system nor the technical fluency of the operator. This approach allows for a transparent human oversight process which ensures accountability when deploying an autonomous (weapon) system.

During the Interpretation stage of the Glass Box framework, values in the governance layer of the CHOF are turned into concrete norms before deployment of the autonomous system, constraining the observable elements and actions in the socio-technical layer of the CHOF, which in turn are translated into requirements in the technical layer of the CHOF. 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. The block in the socio-technical layer during deployment is treated as a black box. A Review stage is required after deployment as an accountability process [3] 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. 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 [1].

The first implementation concept to operationalize the socio-technical framework by projecting the Glass Box framework over the CHOF was based on existing operational norms (e.g. rules of engagement in the military domain) and left the value elicitation - which is the first step of the Interpretation stage - out-of-scope for the implementation concept [1]. In the next step of the research this gap was filled by conducting value elicitation by means of the Value Deliberation process [4] for the deployment of an autonomous weapon system (AWS).

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08 https://doi.org/10.1145/3514094.353952

The value elicitation conducted using the Value Deliberation process is a form of participative deliberation and gives insight into which values are deemed important. As a next step in the Interpretation stage of the Glass Box framework, norms and requirements can be derived based on this value elicitation. These requirements will feed in the Observation stage as observable elements to monitor and verify. The Review stage is required after deployment as an accountability process of which findings should feed back into the Interpretation stage for a next deployment of an AWS and thereby close the loop between the stages.

In future work, I will close the feedback loop from the Review stage as accountability process back to the Interpretation stage of the Glass Box framework. This ensures that the lessons and recommendations from the review stage will be incorporated in the Interpretation stage in a next iteration.

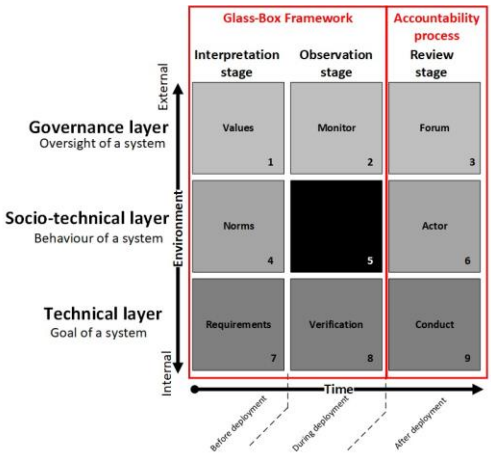

Figure 1: Glass Box framework projected on CHOF (in: [1])

REFERENCES

[1] Verdiesen, I.; Aler Tubella, A.; and Dignum, V. 2021. Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones. Information 12(9): 385.

[2] Aler Tubella, A., & Dignum, V. (2019). The glass box approach: Verifying contextual adherence to values. In AISafety 2019, Macao, China, August 11-12, 2019. CEUR-WS.

[3] Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. European Law Journal, 13(4), 447–468.

[4] Pigmans, K. 2020. Value Deliberation: Towards mutual understanding of stakeholder perspectives in policymaking. Ph.D. thesis, Delft University of Technology.