Governance by Glass- box: Implementing transparent moral bounds for AI behaviour*

Andrea Aler Tubella[1], Andreas Theodorou[1],
Frank Dignum[1], and Virginia Dignum[1]

Department of Computer Science, Umeå University, Sweden
{andrea.aler,andreas.theodorou,frank.dignum,virginia.dignum}@umu.se

1 Introduction

Trust in Artificial Intelligence (AI) is often linked to algorithmic transparency [9]. This concept includes more than just ensuring algorithm visibility: the different factors that influence the decisions made by algorithms should be visible to the people who use, regulate, and are impacted by systems that employ those algorithms [7]. However, decisions made by predictive algorithms can be opaque because of many factors, for instance IP protection, which may not always be possible or desirable to eliminate [2]. Yet, accidents, misuse, disuse, and malicious use are all bound to happen. Since human decisions can also be quite opaque, as are the decisions made by corporations and organisations, mechanisms such as audits, contracts, and monitoring are in place to regulate and ensure attribution of accountability. The goal of transparency should not be complete comprehension, but rather to provide sufficient information to ensure at least safe usage and human accountability [3]. Where AI is applied to make decisions that affect people and society, the most important issue to consider is perhaps the need to rethink responsibility [4].

Many organisations and nations have produced, or are in the process of announcing, statements on the values or principles that should guide the development and deployment of AI in society. The current emphasis on the delivery of high-level statements on AI ethics may also bring with it the risk of implicitly setting the 'moral background' for conversation about ethics and technology as being about abstract principles [5]. The high-level values and principles are dependent on the socio-cultural context [10]; they are often only implicit in deliberation processes. The shift from abstract to concrete necessarily involves careful consideration of the context. In this sense, the implementation of each value will vary from context to context the same way it can vary from system to system.

For example, consider the development of an intelligent recruitment application. A value that can be assumed for this system is *fairness*. However, fairness can have different normative interpretations, e.g. *equal access to resources* or *equal opportunities*, which can lead to different actions. It is, therefore, necessary to make explicit which interpretation is taken into the design. This decision may be informed by domain requirements and regulations, e.g.national law.

* This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

2 Glass Box Approach

The Glass Box approach, as depicted in Figure 1, consists of two phases: interpretation and observation. It takes into account the contextual interpretations of abstract principles by taking a *Design for Values* perspective [8].

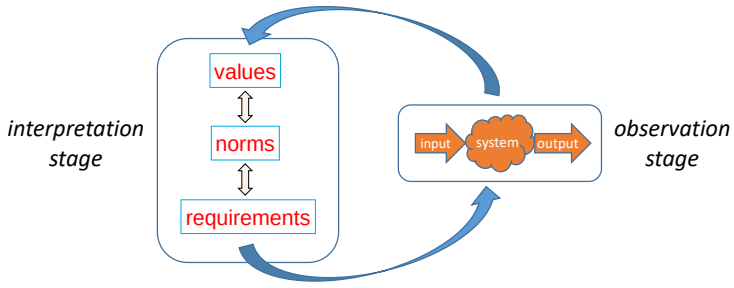

Fig. 1. The two stages of the Glass Box Approach: an Interpretation stage, where values are translated into design requirements, and an Observation stage, where we can observe and qualify the behaviour of the system.

The interpretation stage is the explicit and structured process of translating values into design requirements. It entails a translation from abstract values into concrete norms comprehensive enough so that fulfilling the norm will be considered as adhering to the value. Following a Design for Values approach, the shift from abstract to concrete necessarily involves careful consideration of the context. Normative systems are often described in deontic-based languages, which allow for the representation of obligations, permissions and prohibitions. In the Glass Box Approach we aim to not only describe the norms themselves, but also the exact contextual connection between abstract and concrete concepts.

We use the *counts-as* statements to formalise our interpretation [1]. *Counts-as* is a contextual subsumption relation that describes the translations of higher level elements into lower level, contextualised, concepts. The *counts-as* operator admits formalisations based in modal logic and description logic and thus it lends itself to implementations. Furthermore, the interactions of *counts-as* relations operating in different contexts may be formalised as well [6]. With this approach we can formally represent the explicit relations between abstract and concrete concepts, and, given that the relations between concepts are dependent on the context in which that relation is evaluated, the definition of the context of those relations is made explicit as well. In other words, we can build logical statements of the form: "X counts as Y in context C". In this formalisation, the conjunctions and disjunctions of different norms will therefore stand for the explicit interpretation of a value in a specific, explicit context.

The second step in the interpretation stage is the concretisation of norms into specific system requirements. In the Glass Box Approach these requirements

will be given in terms of the inputs and outputs of the intelligent system. The connection of these lower level requirements to higher level elements can be described in terms of a *for-the-sake-of* relation [8]. This relation would allow for the formalisation of statements of the form "X is done for the sake of Y".

At the end of the interpretation stage we will therefore have built 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, and the connections between nodes on each level are contextual. The concrete requirements inform the observation stage of our approach, as they indicate what must be verified and checked. On the other hand, this hierarchy can be used after the observation stage to provide high-level transparency for a deployed system: depending on which requirements are being fulfilled, we can provide explanations for how and exactly in which context the system adheres to a value.

In the observation stage, the behaviour of the system is evaluated with respect to the values by studying its compliance with these requirements. In [11] two requirements for norms to be enforceable are identified: verifiability i.e., the low-level norms must allow for being machine-verified given the time and resources needed, and computational tractability, i.e. whether the functionalities comply with the norms can be checked on any moment in a fast, low cost way. Note that this is a requirement for the observation stage and not necessarily for the design stage! Hence, some of the norms chosen for the design stage might be easily implementable, but hard to monitor. E.g. a neural network approach to implement a mortgage decision, in which the neural net is trained on all decisions of the last years can provide an implementation not to deviate from decisions in similar cases. However, it is not easy to monitor or govern that the decisions never deviate more than a certain percentage from similar cases.

In order to ensure that the glass box approach is enforceable, governance mechanisms that include the specification of quality of service levels. Observing these levels poses different constraints to the glass box framework: whereas for the former only the behaviour with respect to one given applicant is needed, the latter is dependent of data about many applicants within a given time frame and region.

The mechanisms to observe this behaviour can be implemented without knowledge about the internal workings of the system under observation, by monitoring input and output streams. We insist on this feature as we do not always have access to the internals of the system, neither do we always have access to the designs of a system.

3 Discussion and Conclusions

The focus on inputs and outputs allows for the verification and comparison of vastly different intelligent systems, from neural networks to agent based systems. Moreover, the versatility of our approach allows us to check the compliance of

the system against different interpretations of the same value, e.g. the American interpretation of 'fair use' for data handling is different compare to the one set by the EU. Furthermore, the system can include its adherence to a value as an explanation for its actions, providing a high-level transparency necessary to ensure the due diligence of a system.

The Glass Box Approach opens interesting avenues for follow-up research. The development of a formalism to express concrete input/output requirements would be an interesting first step. Furthermore, such a formalism could be linked to formal verification procedures. In addition, although the mechanisms for describing the connections between requirements and norms and between norms and values are outlined in this paper, further insight may be gained through the development of an all encompassing requirement-to-value formalism that would allow for the description of the full hierarchy of concepts. On the other hand, given a glass box, it would be interesting to study whether the systems that fulfil its requirements can be characterised. Last, technical implementations of concrete glass boxes for a system may be developed.

References

1. Aldewereld, H., Álvarez-Napagao, S., Dignum, F., Vázquez-Salceda, J.: Making Norms Concrete. In: Hoek, v.d., Kaminka, Lespérance, Luck, Sen (eds.) Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010). pp. 807–814. Toronto, Canada (2010)
2. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645
3. Bryson, J.J., Theodorou, A.: How society can maintain human-centric artificial intelligence. In: Toivonen-Noro, M., Saari, E., Melkas, H., Hasu, M. (eds.) Human-centered digitalization and services. Springer (2019)
4. Dignum, V.: Responsible autonomy. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'2017) (2017)
5. Greene, D., Hoffmann, A., Stark, L.: Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In: Proceedings of the 52nd International Conference on System Sciences (2019)
6. Grossi, D.: Designing Invisible Handcuffs: Formal Investigations in Institutions and Organizations for Multi-agent Systems. Ph.D. thesis, Universiteit Utrecht (2007)
7. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., Vinck, P.: Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology **31**(4), 611–627 (2018)
8. Van de Poel, I.: Translating values into design requirements. In: Philosophy and engineering: Reflections on practice, principles and process. Springer (2013)
9. Theodorou, A., Wortham, R.H., Bryson, J.J.: Designing and implementing transparency for real time inspection of autonomous robots. Connection Science **29**(3), 230–241 (7 2017)
10. Turiel, E.: The culture of morality: Social development, context, and conflict. Cambridge University Press (2002)
11. Vázquez-Salceda, J., Aldewereld, H., Grossi, D., Dignum, F.: From human regulations to regulated software agents' behavior. Artificial Intelligence and Law **16**(1), 73–87 (3 2007)