# Synthetic Consequential Reasoning: Facilitating the Design of Synthetic Morality in Highly Automated Systems via a Multidimensional-scalar Framework

Fernando Galdon(✉) and Ashley Hall

School of Design, Royal College of Art, London, UK
fernando.galdon@network.rca.ac.uk

**Abstract.** This paper reviews the four fundamental frameworks available in normative ethics to underpin the most suitable strategy to facilitate the design of synthetic morality in the context of Highly Automated Systems (HAS). Based on research findings, it will present an updated multidimensional-scalar system of levels of automation specifically adapted to Highly Automated Systems (HAS) in the context of Human-Human-Interaction (HHI). This framework integrates the variables of autonomy, accountability, reparation, actions, contexts, access and inferences to build and facilitate the design of synthetic morality on highly automated unsupervised systems form a consequential perspective. As part of this process, a form of calculation emerges to facilitate the calibration of moral computational reasoning in the context of HAS.

AQ1

**Keywords:** Human factors · Human-systems integration · Systems design · Synthetic morality · Consequential reasoning · Highly Automated Systems · Complex dynamic systems

## 1 Introduction

Due to the ever-evolving complexity of highly automated systems (HAS), "no amount of testing can formally guarantee that a system will behave as we want" [1]. With this statement, DeepMind, the most advanced AI company in the world, summaries the current state of the art in the field of Artificial Intelligence. In large-scale models enumerating all possible outputs for a given set of inputs, remains intractable due to the incredible number of choices for the input perturbation. In this context, users and the general public are becoming increasingly concerned with issues of unexpected outcomes leading to a lack of trust in these systems. This problematic demands the design of systems and tools that can integrate moral reasoning as part of their reasoning capabilities.

In order to resolve this conundrum, this paper will critically analyse the four most relevant frameworks in normative ethics to underpin the most reliable approach to design moral computational reasoning; Socrates's virtue, Jeremy Bentham's Consequentialism, Emmanuel Kant's Deontology and John Dewey's Pragmatism.


T. Ahram et al. (Eds.): IHIET 2020, AISC 1152, pp. 1–6, 2020.
https://doi.org/10.1007/978-3-030-44267-5_76

They incorporate the three main existing frameworks in normative ethics plus Dewey's Pragmatism. The latest allows the authors to address ethics at a systemic level.

- Virtue refers to being. In this paradigm, morality emerges from the identity of the individual rather than their actions or consequences. Practical reason results in action or decisions chosen by a suitably 'virtuous' agent.
- Consequentialism states that the consequences of somebody actions are the ultimate basis for any kind of judgment regarding that action. This perspective focuses on the outcome of conduct.
- In deontology, actions are conditioned by a set of rules, may they be natural, religious or social. This perspective focuses on the intentionality of conduct.
- In Pragmatism, actions and consequences are possible because the context or system allows for them. This paradigm aims for social reform as a strategy to address morality. In this perspective social reform is prioritised over intentionality, consequences, individual virtue or duty.

The fundamental problem with Dewey's perspective is that in order to change the system we need to generate global consensus. In this context, some initiatives such as GDPR have been taking place, but they felt short [2] and rapidly become redundant. The GDPR was introduced on 25 May 2018 and on 24 October 2019, the German government was introducing a new framework [3]. The limitations of access and rapid technological exponential development prevent Dewey's framework from addressing the design of a system.

In Socrates's virtue, the fundamental problem is the limited capability of humans to assess what is happening. The acceleration and volume of information delivered by social interactions and algorithmic updates is fragmenting reflection and cognition by disconnecting the pre-frontal cortex by saturation. Our attention span has been reduced from 12" to 8" in four years by multitasking [4]. After 21 min comparing information our pre-frontal cortex shuts down [5] and only information with a big emotional impact is retained. These processes are transforming society from reflexive to reactive and it is questioning the idea of truth and reality by repositioning the decision centre from reason to emotional experience. Thus invalidating the model proposed by Socrates based on reason.

In Emmanuel Kant's deontology the ethical intervention is placed on the intentionality of the system. Its fundamental problems are interpretability and interruptibility. The system does not know what is doing, therefore, it cannot stop. According to researchers from the most advanced AI company in the world DeepMind, this is currently impossible [6]. Insofar as we are not capable of designing them, it is not a suitable strategy. Consequently, the only paradigm remaining is Consequentialism. In this framework the fundamental elements are the consequences of an action, therefore, the system will be judged by the consequences of its actions.

### 1.1 Synthetic Consequential Reasoning

Building from this ethical perspective, the lead author developed a multi-dimensional scale system to build consequential reasoning in computational systems [7]. The enquiry started by building a foundational scale of levels of autonomy. This technique

has been widely used in the human factors field over the last 40 years. However, as a consequence of the impossibility of monitoring complex dynamic systems due to their complexity, two fundamental questions emerge; if something goes wrong who is responsible? and is it possible to repair the trust of the user in the system? These question led to the articulation of two complementary consequential scales. In this context, two workshops were conducted to map highly sensitive areas and from the knowledge generated eight case studies (two cases with high and low intensities for each sensitive area) were built to understand whether this multi-dimensional system would be able to address them. The results were positive, however context and actions emerged as fundamental variables to incorporate in the framework, as they determined the right combinations of levels. Then, building from the work of Sandra Wachter in the area of law and algorithms, we integrated two more variables; access (data points) and inferences (predicting capabilities) [2]. Finally, we designed a system integrating all these variables and developed a risk analysis tool with a form of calculation to obtain a trust rating to calibrate the system [7]. (Fig. 1) (see [7] for an extended explanation).

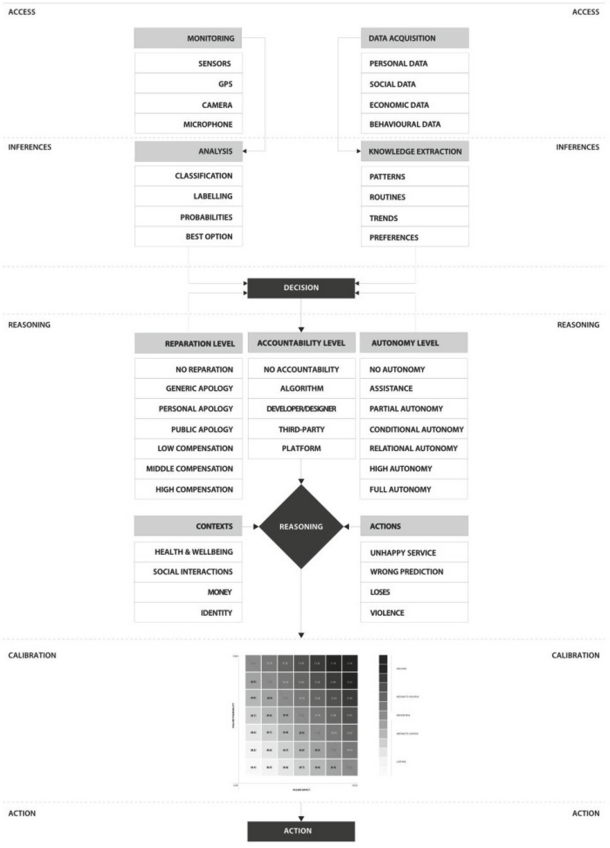



**Fig. 1.** Synthetic consequential reasoning framework

## 2 Method

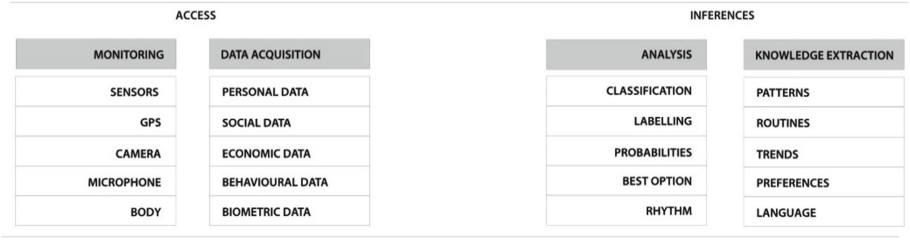
In order to test the proposed framework, a workshop with students from the Master in Research program at the Royal College of Art was implemented. They represented a mix of backgrounds in fashion, textile, architecture, computer science, industrial design, and engineering. The lead author defined the main area of intervention; health and wellbeing. This area was specifically selected due to its moral nature and impact. Then, a design task around a highly automated Virtual Assistant capable of diagnosing and providing treatment in the area of depression was structured.

As part of the workshop, the main author introduced a demo called Duplex to illustrated the nature of the system, and a small analysis underlined the key characteristics of upcoming Virtual Assistants. Students had 50 min to complete this task. We provided the aforementioned framework in the form of a calculator with all the variables. This tool provided a trust rating to calibrate interactions beforehand.

In order to understand the validity of the framework a comparative analysis was implemented to understand whether new elements not considered in the proposed framework emerged. Once the task was completed, the authors designed a semi-structured questionnaire to understand four elements; usefulness of the use of the calculator, whether the calculator helped them to improve their design, specific usefulness of the rating and whether the rating helped them to fine tune their decisions. The questionnaire consisted of two areas; a quantitative section asked participants to rate these elements by using an eleven point Likert scale and a qualitative section asking participants to expand why and how these elements have affected their design.

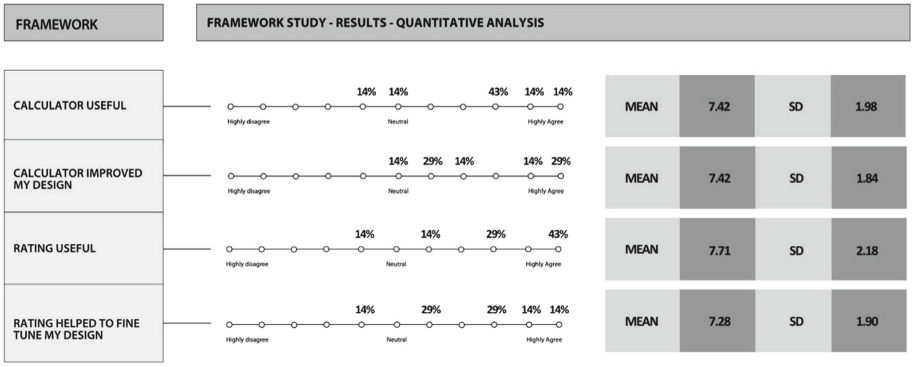
## 3 Discussion

In terms of differences (new elements not considered in the proposed framework). No alternative emerged to the proposed scales; autonomy, reparation and accountability. Neither differences emerged in terms of contexts and actions. However, subtle differences emerged in terms of access and inferences. Biometric data and the body present a categorical novelty to account in terms of data points, and language and rhythms emerged as new elements not previously considered in terms of inferences.



In terms of the usefulness of the framework in the form of a calculator participants rated it with a 7.42 mean value. In terms of product improvement, participants also rated the usefulness of the tool with 7.42 mean value. In terms of the rating usefulness, participant rated this element with 7.71 mean value. In terms of the effect of the rating to fine-tune decisions, participants rated it in with 7.28 mean value (Table 1).

**Table 1.** Quantitative analysis




In qualitative terms, participants described how these elements affected their decisions by understanding the impact on trust of the interaction beforehand. This exercise led to participants reducing risks by giving them a better perception of the implications their design may have on the user's trust. By the results presented we can establish that the framework and its mode of calculation is useful to facilitate the design of moral computational systems.

## 4 Conclusion

This paper presents an innovative multi-dimensional scalar system integrating post-interaction elements such as accountability and reparation, and integrating unintended actions, contexts, access and inferences as fundamental variables to facilitate the design of synthetic morality from a consequential perspective on unsupervised highly automated computational systems in the context of Human-Human-Interaction (HHI). This perspective has been traditionally missing in the design of computational systems and tools which fundamentally revolve around a priori strategies focusing on intentionality and monitoring. As part of this process, a form of calculation emerges to facilitate the calibration of moral computational reasoning in the context of HAS.

Finally, the study has underpinned four new elements. In terms of access, biometric data and the body, and in terms of inferences, language use and rhythm emerged as new sub-variables to account. These elements have been added to the previously proposed framework. Future work will be dedicated to build a functional prototype.

## References

1. Kohli, P., Gowal, S., Dvijotham, K., Uesato, J.: Towards robust and verified AI: specification testing, robust training, and formal verification. Deepmind. Medium, 28 March 2019. https://deepmind.com/blog/robust-and-verified-ai/. Accessed 29 Mar 2019
2. Wachter, S., Mittelstadt, B.: A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. Columbia Bus. Law Rev. SSRN (2018, forthcoming). https://ssrn.com/abstract=3248829
3. German Government Data Ethics Commission. https://datenethikkommission.de/wp-content/uploads/191023_DEK_Kurzfassung_en_bf.pdf
4. National Center for Biotechnology Information. Attention span statistics. Statisticbrain (2016). http://www.statisticbrain.com/attention-span-statistics/
5. Mullins, P.A.: Brain-scan shoppers project. Bangor University (2013). https://www.bangor.ac.uk/news/university/ground-breaking-project-to-brain-scan-shoppers-16874
6. Ortega, B.P.A.: Building safe artificial intelligence: specification, robustness, and assurance specification: design the purpose of the system. Medium (2018). https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f
7. Galdon, F., Wang, S.J.: Optimising user engagement in highly automated virtual assistants to improve energy management and consumption. In: Proceedings of the 2019 Applied Energy Symposium AEAB, MIT, Boston, Massachusetts (2019)