



# The Right to Reparations: A New Digital Right for Repairing Trust in the Emerging Era of Highly Autonomous Systems

Fernando Galdon<sup>(✉)</sup> and Ashley Hall

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

**Abstract.** This paper argues that a new digital right, the ‘right to reparation’, is needed to address the accountability gap presented by highly autonomous complex systems (HACS) incapable of fully monitoring their actions in real-time due to the increasing complexity of these advanced systems. The ‘Right to reparation’ follows the articulation of the ‘Right to be forgotten’, the ‘Right of access’ or more recently the ‘Right to Reasonable Inferences’, and aims to ensure that emerging HACS interactions remain accountable as current highly autonomous technologies cannot fully guarantee the effect of their behaviors. Building from an integrative review of previously published surveys specifically designed to address the rising concerns of artificial intelligence in the context of HACS, this paper presents indications by which introducing reparation and accountability strategies increase trust and engagement in the system in the context of unexpected events. Thus, building a case for the introduction of the newly proposed digital right.

**Keywords:** Human factors · Human-systems integration · Systems engineering · Digital rights · Ethics · Reparation · Accountability · Highly autonomous complex systems

## 1 Introduction

Skepticism and a lack of trust in AI has increased recently with citizens believing that the overall system is neither accountable nor transparent. To rebuild trust and restore faith in the system, some experts suggest that institutions must step outside of their traditional roles and work toward a new, more integrated operating model that puts people and the addressing of their fears—which mainly revolve around technological developments in artificial intelligence—at the centre of everything they do.

In this context, recent developments in computing prompted Peter Hancock to raise a warning to the human factors community by which attention must be focused on the appropriate design of a new class of technology: Highly Autonomous Systems (HAS) [1].

With the development and combination of machine learning and deep learning techniques a new paradigm is emerging; Machine-Human-Interaction (MHI). In this

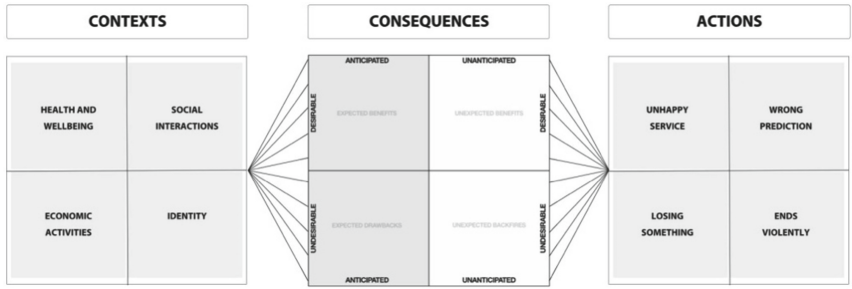
paradigm, technology controls the initiation of interaction. This approach positions highly autonomous systems at the centre and tries to address the implications of trust from their perspective [2]. Traditionally, HACS required the human operator to appropriately calibrate their trust in the automation in order to achieve performance and safety goals. However, a recent statement from DeepMind, the most advance AI company in the world states that “No amount of testing can formally guarantee that a system will behave as we want. In large-scale models, enumerating all possible outputs for a given set of inputs...is intractable due to the astronomical number of choices for the input perturbation” [3]. Recognising the impossibility of fully calibrating HACS, the authors note the challenge and propose the ‘Right to reparation’ as a human-centred strategy directly aimed at ensuring that emerging HACS interactions remain accountable to the user’s needs and preferences.

This paper argues that a new digital right, the ‘right to reparation’, is needed to address the accountability gap presented by highly autonomous complex systems incapable of fully monitoring their actions in real-time. The ‘Right to reparation’ follows the articulation of the ‘Right to be forgotten’ [4], the ‘Right of Access’ [5] or more recently the ‘Right to Reasonable Inferences’ [6], and aims to ensure that emerging HAS interactions remain accountable while the development of highly autonomous technologies cannot fully guarantee their behaviours. Building from an integrative review of previously published surveys specifically designed to address the rising concerns of artificial intelligence in the context of Highly Automated Systems [6, 7], this paper presents indications by which introducing reparation and accountability strategies increase trust and engagement in the system in the context of unexpected events, these results build a case for the introduction of the newly proposed digital right.

## 2 Method









A preliminary co-design workshop with students from the Royal College of Art was structured to analyse the emerging concerns of Highly Automated Complex Systems where we wouldn’t be able to fully guarantee their behaviours via a case study of Virtual Assistants (VA). It was approached from a consequential perspective to underpin its implications.

This activity underpinned two fundamental elements. On one side the four main highly sensitives areas where HACS may impact users significantly. As a result health and wellbeing, identity, economically related activities and social interactions emerged as the most highly sensitive areas. On the other hand, four major unintended consequences; unhappy services, wrong predictions, unintended losses related to the service and actions unexpectedly ending violently emerged (Fig. 1).



**Fig. 1.** Co-design workshop outputs - Fernando Galdon

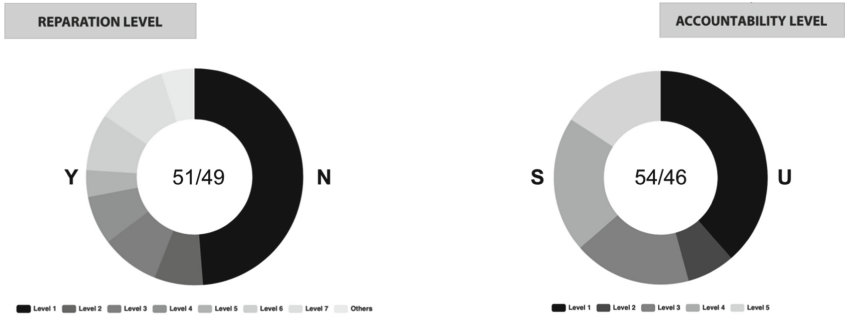
From the areas aforementioned and based on demos, patents and prototypes, eight case studies were built to address different contexts and unintended consequences (Fig. 2). Two cases addressed each highly sensitive area ranging from low to high impact. Then, a survey was designed to establish whether a posteriori strategies such as reparation and accountability in highly automated virtual assistants were relevant to address the rising concerns of HACS for each case.

IDENTITY	HEALTH + WELLBEING	SOCIAL INTERACTIONS	MONEY ACTIVITIES
PREDICTING POLITICS 	PREDICTING COUGH 	PREDICTING BEST DATE 	PREDICTING BEST JOB 
PREDICTING SEXUALITY 	PREDICTING DEPRESSION 	PREDICTING DOMESTIC VIOLENCE 	PREDICTING INVESTMENT 

**Fig. 2.** Case studies distribution - Fernando Galdon

Participants were asked two questions for each case;  
 + the VA predicts ... but something goes wrong ... who would be accountable?  
 + the VA predicts ... but something goes wrong ... what would be the right level of reparation?

The survey was answered by 50 participants including 21 men, 27 women and 2 who didn't want to gender identify. They represented 14 different nationalities with an age range between 18–67 years old from different professions. The survey was live for four weeks and distributed to maximise a robust distribution via dissemination with relevant profiles and relevant forums.



**REPARATION**

AREA	UNHAPPINESS	UNHAPPINESS	END VIOLENCE	END VIOLENCE	WRONG PRED.	WRONG PRED.	LOSE SMTHG.	LOSE SMTHG.	TOTAL
	medicine	newspaper	addiction	rape	sexuality	jailed	money	job	
LEVEL 1 No reparation	36%	48%	56%	64%	46%	56%	40%	44%	48.75%
LEVEL 2 Basic apology	10%	8%	10%	6%	10%	4%	4%	6%	7.25%
LEVEL 3 Personal apology	22%	12%	2%	0%	24%	2%	2%	6%	8.75%
LEVEL 4 Public apology	6%	20%	6%	4%	4%	4%	6%	8%	7.25%
LEVEL 5 Low compensation	6%	2%	4%	4%	2%	0%	10%	4%	5.00%
LEVEL 6 Medium compensation	6%	0%	10%	2%	10%	8%	14%	18%	8.50%
LEVEL 7 High compensation	4%	6%	10%	12%	2%	20%	20%	10%	10.50%
OTHER -	10%	4%	2%	8%	2%	6%	4%	4%	5.00%

**ACCOUNTABILITY**

AREA	UNHAPPINESS	UNHAPPINESS	END VIOLENCE	END VIOLENCE	WRONG PRED.	WRONG PRED.	LOSE SMTHG.	LOSE SMTHG.	TOTAL
	medicine	newspaper	addiction	rape	sexuality	jailed	money	job	
LEVEL 1 Platforms	20%	26%	12%	6%	14%	22%	18%	8%	15.75%
LEVEL 2 Designer	18%	28%	18%	14%	22%	24%	32%	8%	20.50%
LEVEL 3 Algorithm	16%	20%	8%	6%	38%	30%	16%	8%	17.75%
LEVEL 4 User	38%	24%	56%	38%	24%	22%	34%	74%	38.75%
OTHER -	8%	2%	6%	36%	2%	2%	0%	2%	7.25%

Fig. 3. Survey results - Fernando Galdon

### 3 Discussion

Building from the eight case studies aforementioned, in average 48,75% of participants did not demanded any type of reparation as part of interacting with Highly Automated Systems in Highly sensitive areas. However, 51,25% of the participants demanded some type of reparation strategy. In this context, 23.25% of participants would accept some sort of apology, and 23.00% in average would demand some kind of compensation to repair their trust in the system. The remaining 5,00% demanded a combination of apology and compensation (Fig. 3)

In terms of accountability, in average from the eight case studies addressing unexpected consequences in the interaction, 46% placed the accountability out of the system, 38.75% place the accountability in the user, 7.25% placed it in third parties delivering the service (for instance a pharmacy delivering some drugs to customers) and, 54% of participants placed the main accountability on the system side. Specifically, 20.50% of participants pointed to designers/developers, 17.75% pointed to the algorithm, finally, 15.75% pointed to the platform as accountable (Fig. 3).

The survey aimed to understand whether unexpected consequences derived from unsupervised Highly Autonomous Complex Systems where we cannot guarantee its behaviour/output affected users trust and engagement, to understand whether a posteriori reparative strategies such as reparation and accountability could provide a framework to address the rising concerns in these systems.

From the surveys conducted, contexts (highly sensitive areas) and actions (unintended consequences) played a role in determining user engagement. The 50/50 in average split presented by this research presents an empirical need for approaching the design of these system equally from preventive a priori strategies around simulation and calibration strategies to reparative a posteriori strategies around accountability and reparation.

These results present indications by which introducing reparation and accountability strategies increase trust and engagement in the system in the context of unexpected events. Thus building a case for the introduction of the newly proposed digital right.

### 4 Conclusion

In the results presented, the authors recommend the articulation of ‘the right to reparation’ to successfully build, maintain and repair trust in highly autonomous complex systems. This paper argues that this new digital right, the ‘Right to reparation’, is needed to address the accountability gap presented by highly autonomous complex systems incapable of fully monitoring its actions in real-time. The ‘Right to reparation’ follows the articulation of the ‘Right to be forgotten’, the ‘Right of access’ or more recently the ‘Right to Reasonable Inferences’, and aims to ensure that emerging HACS interactions remain accountable while the development of highly autonomous technologies cannot fully guarantee its behaviour.

## References

1. Hancock, P.A.: Imposing limits on autonomous systems. *Ergonomics* **60**(2), 284–291 (2017). <https://doi.org/10.1080/00140139.2016.1190035>
2. Ortega, B.P.A., Maini, V.: Building safe artificial intelligence: specification, robustness, and assurance Specification: define the purpose of the system. *Medium* (2018). <https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f>
3. 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 (2019). <https://deepmind.com/blog/robust-and-verified-ai/>. Accessed 29 Mar 2019
4. Weber, R.H.: The Right to Be Forgotten: More Than a Pandora’s Box? 2 (2011). *JIPITEC* 120, para. 1
5. EU GDPR: GDPR Regulation: Art. 15 GDPR: Right of access by the data subject (2019). <https://gdpr-info.eu/art-15-gdpr/>
6. Wachter, S., Mittelstadt, B.: A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *C. Bus. Law Rev.* **2019**(2) (2018). SSRN <https://ssrn.com/abstract=3248829>
7. Galdon, F., Wang, S.J.: From apology to compensation: a multi-level taxonomy of trust reparation for highly automated virtual assistants. In: *Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHiet 2019) Conference*, 22–24 August 2019, Nice, France (2019)
8. Galdon, F., Wang, S.J.: Addressing accountability in highly autonomous virtual assistants. In: *Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHiet 2019) Conference*, 22–24 August 2019, Nice, France (2019)