# On Robots with Reasoning Capabilities and Human-like Appearance and Behaviour. Implications for Accident Investigations.

Pericle SALVINI[a,1], Jo-Ann PATTINSON[b], Lars KUNZE[c], Alan WINFIELD[c], Marina JIROTKA[a]
[a] *Department of Computer Science, University of Oxford, United Kingdom*
[b] *Faculty of Environment, University of Leeds*
[c] *School of Engineering, University of the West of England*
ORCiD ID: Pericle Salvini https://orcid.org/0000-0001-9602-6085
ORCiD ID: Lars Kunze https://orcid.org/ 0000-0001-5302-1938
ORCiD ID: Alan Winfield https://orcid.org/ 0000-0002-1476-3127
ORCiD ID: Marina Jirotka https://orcid.org/ 0000-0002-6088-3955

**Abstract.** AI-enhanced reasoning enables robots to create detailed accounts of their own situated behaviour as well as the behaviour of other people. This capability is currently employed by robot designers to achieve transparency, trust, and enhance robot social and communicative capabilities. Furthermore, robots may be designed to resemble humans both in their physical appearance and their behaviour. This approach is intended to facilitate more effective interactions with people. In this article we identify and examine some of the ethical, social and legal implications of these capabilities for the investigation of robot accidents. We consider two aspects in particular. The first of these is the role of robots as subjects in a testimony regarding an incident in which they are directly or indirectly involved. This can be described as a case of robots acting as witnesses. The second aspect is the role of robots as objects in a human testimony. This can be described as a case of robots being witnessed.

**Keywords.** Robot accident investigation, reasoning capabilities, explainable AI, human-like appearance and behaviour, robots as witnesses, robots as being witnessed

## 1. Introduction

The dictionary defines accountability as "the fact of being responsible for what you do and able to give a satisfactory reason for it, or the degree to which this happens" [1]. Artificial Intelligence (AI)-enabled reasoning allows robots to create detailed accounts of their own situated behaviour as well as the behaviour of other people. Drawing on data-driven insights and contextual understanding, these capabilities may contribute to transparency, accountability, and effective communication between robots and humans [2]. Indeed, this self-accountability aspect could be determinant for developing further the social capabilities of so called 'social robots'. Producing an account of one's

[1] Corresponding Author: Pericle Salvini, pericle.salvini@cs.ox.ac.uk.

behaviour or actions is an inherently social practice and, for adults of sound mind, being held accountable is central to sociality. To be fully considered social therefore, robots may need to be able to account for their actions and to do so in ways that are understandable and actionable by others. Verbal, and other practices of accounting can be embedded into the design of social robots. For instance, the inclusion of a natural language explainer system [3] could enable the robot to explain its own behaviour; robot's users and others would be able to ask the robot "why did you just do that?" or even "what would you do if…?" and the robot would give a simple intelligible explanation.

AI systems can process vast amounts of data, including their own interactions and experiences. By analysing this data, they can learn patterns, correlations, and context-specific behaviours. An AI robot equipped with reasoning capabilities could self-reflect by accessing its historical data (e.g., from a data recorder). For instance, a service robot in a household might analyse its past actions—such as cleaning routines, interactions with family members, and responses to various situations—to provide an account of its behaviour. This self-reporting function could be useful for maintenance, troubleshooting, in addition to accountability purposes.

Moreover, AI models can infer context from data. By combining sensory inputs (such as visual data from cameras, audio from microphones, or touch feedback) with stored knowledge, a robot can construct a narrative about its actions. Suppose a delivery robot encounters an obstacle while navigating a busy street. It can analyse sensor data, identify the obstacle (e.g., a pedestrian), and generate a contextual account: "I slowed down to avoid colliding with a pedestrian who suddenly stepped onto the crosswalk". Further, formal causal inference methods will also allow robots to answer counterfactual 'what if' questions. In case of an accident with the pedestrian this would allow an accident investigator to understand if the accident could have been prevented if the delivery robot had taken another action (i.e. made another decision). Such narratives or counterfactual narratives enhance transparency and confidence in AI systems, especially when they operate autonomously or interact with humans [4].

From a technological standpoint, to give an account of one own's actions, namely a written or spoken description of an event, or answer questions about what they have seen, done, plan to do and why, may not remain an exclusively humans-only competence. This capability introduces the possibility of using robots as witnesses, namely robots could be able to provide a testimony in a legal trial.

Furthermore, robots are designed to emulate not only human reasoning abilities, but also to replicate human appearance and behaviour. Such robots are referred to as androids or ginoids [5], which are robots that have been designed to be indistinguishable from a human being. The most successful results have been achieved in the reproduction of the human face, which has reached a high level of detail (e.g. the presence of wrinkles, eyelashes, teeth, tongue, etc., with the corresponding movements). The use of special artificial skins and very small servo-motors enables the reproduction of the minute facial movements typical of the human face, as well as expressions of emotion. The incorporation of artificial intelligence has endowed these robots with the capacity to emulate human-like behaviour and engage in discourse with humans, thereby enhancing the likelihood of being perceived as human. This design approach is intended to facilitate more effective interactions with humans in different domains.

Drawing on the existing literature, we will explore the consequences that these two types of simulation–the emulation of the human ability to recount and reason about facts and events and the replication of the human-like appearance and behaviour–could have

in the investigation of robot related accidents. This study will examine two cases in particular: firstly, robots as subjects of testimony (that is to say, robots as witnesses); secondly, robots as objects of a human's testimony (that is to say, humans witnessing robots).

## 2. Robots as witnesses

In this section we will focus on robots as the subjects of a testimony regarding an incident in which they are directly or indirectly involved. Let us consider the following scenarios involving robots with reasoning capabilities: a) an older person living alone is found in a state of unconsciousness, with her personal care robot exhibiting erratic behaviour; b) in the vicinity of a site where a criminal act was perpetrated, a last-mile delivery robot or a street cleaning robot was in operation. In the first scenario, the robot could be directly involved in the accident, for example, colliding with the person. In the second scenario, the robot might have witnessed the crime. In both scenarios, robots designed with explainable AI architecture could be capable of providing an account of their own actions and those of others. This gives rise to the question of whether robot testimony would be admissible in a court of law.

By law, a robot cannot testify because it is not (yet) either a natural or a legal person [6]. Nevertheless, a natural language account of a robot possessing reasoning capabilities may be accepted as evidence. It could be regarded as an alternative means of presenting the robot's stored data.

Indeed, all evidence is admissible in a law court, but it is up to a judge to decide if evidence is reliable. And how to decide whether robot evidence is reliable is a challenging question. The concept of robot evidence or machine evidence has already been addressed in the literature [7, 8].

Our case of self-accountable robots falls into the category of consumer products with so called 'function creep' [9], which is when information from non-forensic technological systems, such as a smart watch (its function is to measure beats not to be used as evidence in a trial), is used as evidence. A first relevant aspect emerges. In all these cases, it would be important to be aware that the device/robot capability to provide an account was not designed for legal purposes but for other purposes and this could have negative consequences if used as testimony in a trial. An illustrative example is provided in [7]: "car producers may wish to reduce their own potential liability by calibrating the system in a way that triggers an alarm at the very first sign of potential drowsiness". In [8], the author propose a framework for conceptualizing machine evidence. She points out that machine evidence as human testimony can be vulnerable to hearsay, though in a different way (i.e. the black box problem). Credibility testing is therefore needed. The paper offers a taxonomy of machine evidence, proposes testimonial safeguards for machines and explains why machine sources can be 'witnesses' under the Sixth Amendment (which is about the accused rights in criminal persecution).

Another possible risk related to the use of robot evidence is that investigators and Judges are currently completely ill-equipped and uneducated about how AI works and how explainable AI systems work. A judge unfamiliar with these concepts may accept anything a robot says as truth/accurate, firstly, because it is known that humans witnesses can lie and mis-remember events, and secondly because there may be a bias towards robot testimony as being more accurate, without appreciating the potential problems with how robot data is interpreted and communicated. These psychological

phenomena - of automation bias and complacency effects - have already been studied in aviation and recently in decision support systems used in medical practice. In other words, researchers warn that machine-based evidence may be attributed greater importance than human evidence (i.e. testimony) [10]. The author is interested in how it is possible to defend oneself when robot evidence is used in criminal proceedings, in particular when the defendant is underfunded and does not have access to expert technical advice or representation. Often in ML systems it is not possible to explain why the system arrived at a certain decision, even for experts. "This makes it very difficult to test the reliability of the evaluative data generated by systems using such techniques" [10]. Sometimes the need for help from the manufacturer might be necessary. However, if the manufacturer itself is involved in the litigation this can represent a problem.

We argue that whilst novel forms of robot self-accountability are interesting, they should be constituted and interpreted within a social frame. That is to say, investigators should not necessarily take the robot's word for what it did at face value. Would the robots own account be a neutral one? Moreover, how can we be sure that the reasoning framework is working correctly? In aviation it is the investigation, not the black box data per se, which seeks to discover why and how an air incident occurred. We anticipate this should also be true for incidents involving social robots, where an investigation will draw upon data recorders, robot's self-accountability or explainer interfaces amongst other information to determine the reason for an incident. Hence, alongside the technical parameters of what to record within a robot data recorder and the possibility to ask directly questions of the robot, investigators should also consider how the interpretation of those recordings and statements fit into the conduct of an investigation.

Incident investigations are social processes of reconstruction that need to be perceived as impartial and robust, and which (we argue) serve as a form of closure so that social robots do not acquire an enduring taint in the public's consciousness.

### *2.1. What distinguishes human 'evidence' for a fact from the recorded data of a robot*

There are different types of 'human' witnesses–one being a person who is giving a direct account of their experience of an event or conversation etc. (which is similar to data) and there are also an expert witnesses–used to interpret data or give an expert opinion on evidence or events. According to [8], in specific circumstances machine evidence can play the role of expert witnesses because it generates information that is beyond the court' knowledge.

In terms of robots being witnesses, they are not witnesses in the same way a human is a witness. Until AI, 'robots' were just data processors. If a robot is run on rule-based software then what they produce is just data which can be submitted as evidence in Court. In an adversarial system, evidence is submitted by a party to the proceedings and can be tested or challenged by another party. In other legal systems, such as in the UK there is a common law presumption that when a computer or machine is producing data that it was working properly. If you wanted to rebut this presumption, then to challenge this you would need to get an expert (human).

If a robot can speak in Court and is able to provide explanations on decisions made and interpretations on events, its testimony could be considered as a post-hoc explanation. There will still exist the problem of interrogating the evidence as if black box AI had been used, the post-hoc explanation is not necessarily faithful to the actual decisions or testimony made by the robot.

With human witnesses, like AI, it is not possible to look into their head, but it is possible to challenge and cross examine them. To challenge AI an expert on that particular model is needed. The reliability of post-hoc explanations as evidence is questionable. An AI-based robot can be challenged or interrogated with human experts. However, this can be very expensive and time consuming.

### *2.2. Why should we need robots as witnesses?*

This section aims to address the question of why there is a need for robots that are capable of providing testimony. This will be achieved by first identifying a potential ethical issue that may arise from the absence of such a capability. Secondly, a legal argument will be presented concerning the potential for exceeding privacy limits in specific circumstances.

#### *2.2.1. Ethical considerations*

It seems reasonable to posit that accidents between robots and humans are inevitable. The increasing prevalence of robots and the growing automation of tasks previously performed by people could result in accidents without human witnesses.

One illustrative example is the so-called 'Molly problem', which concerns self-driving vehicles. The problem has been described as follows: "A young girl called Molly is crossing the road alone and is hit by an unoccupied self-driving vehicle. There are no eye-witnesses. What should happen next?" [11]. To address this problem, researchers conducted a survey on the public expectations for the post-collision behaviour and recall capabilities of the self-driving software. The results show that respondents expect the autonomous vehicle to be aware of the collision, stop at the collision site, indicate a hazard to other road users, alert emergency services, etc. With regard to the recall capabilities necessary to reconstruct the causes of the accident, the majority of respondents pointed out the need to store data that could be used to reconstruct the accident, such as time and location of the collision, if and when human was detected as human, if mitigation action was taken, etc. Finally, another important result of the survey was that 73% of respondents "expect driving to be prohibited for software without recall capability" (15% unsure 12% don't) [11].

These findings demonstrate that an autonomous system's ability to store, recall and explain specific details regarding its operations is not only crucial for liability purposes but it is also needed for ethical reasons. Indeed, the post-collision behaviour of the driving software has the potential to save human lives (i.e. by alerting emergency services, warning other road users, etc.).

The results of the 'Molly Problem' survey indicate that the general public expects autonomous driving software to be aware of accidents and to be able to perform actions subsequent to the collision. In a different context, [12] considers autonomous systems, such digital home AI-assistants that are able to predict and prevent dangerous human behaviour. In other words, according to the author, robots could be designed with the capacity to report instances of dangerous human behaviour, such as alcoholism, depression or violence, with the aim of anticipating possible serious consequences. Such a robot would no longer have the passive capacity to bear witness to an event that has occurred and is therefore now irreparable, but would function proactively, to prevent potentially dangerous events. The author discusses the ethical and legal implications that would arise from this capability. Indeed, one of the main obstacles for robots with

reasoning capabilities–i.e. able to reason about their own and others' behaviour called to testify (or used as evidence)–would have to do with privacy.

### *2.2.2. Legal considerations*

Today, a significant proportion of cases involving accidents or criminal activity are resolved as a consequence of the evidence provided by technological devices, in particular surveillance cameras. Nevertheless, even commonplace technological devices such as a mobile phone or a credit card could be utilised to search for evidence, for instance in the electronic traces their use implies. In the near future, robots, such as delivery robots or personal assistant robots, could also be utilised to search for evidence in the data they record. However, access to data contained in a technological device that could provide valuable information for an investigation cannot be taken for granted. In the past, there have been a few cases where manufacturers have declined to provide data to law enforcement agencies in order to safeguard the privacy of device owners. As illustrative examples: i) Apple refused to unlock the iPhone of one of the killers of the San Bernardine massacre in California in 2015 and the FBI managed to get access only thanks to the help of a third party [13]; ii) in 2017, Amazon refused to give the police access to data from an Echo device owned by the alleged perpetrator of a murder. The company invoked the First Amendment to oppose the authorities' request to provide the data [14].

The question thus arises as to whether a robot designed with explainable AI functionalities, with the capacity to be aware of an accident (such as the 'Molly problem') and recognise dangerous behaviour (such as that described in [12]), including crimes, should also be designed with some pre- and post-accident/crime behaviour, such as the ability to intervene by alerting the police or the ability to report information to the authorities following the occurrence of an incident.

Let us consider the scenario in which a criminal act, such as robbery or rape, occurs in the presence of a robot. If a human were in the robot's place, it would be reasonable to expect at least recall and testimony, if not action. This is a moral as well as legal imperative. Would this capability, if applied to robots, enhance our sense of safety and security, or would it serve as an additional tool for surveillance?

## 3. Robots as the objects of a human's testimony

In the preceding chapter, we examined a number of potential considerations pertaining to the utilisation of robots with the capacity to emulate human reasoning as witnesses in a legal setting. This section will examine certain issues that arise from the replication of human appearance and behaviour in robots during the investigation of accidents, with a particular focus on instances where human witnesses are required to provide testimony about robots.

In a research project in which some of the co-authors of this paper are involved[2], preliminary findings have revealed that one of the main differences and possible future challenges related to humanoid robots implicated in an accident, such as androids or ginoids, would have to do with human witnesses. This concerns the special characteristic of these robots to mimic human beings. In other words, when robots are designed with

[2] RoboTIPS. Responsible Robotics for the Digital Economy: www.robotips.co.uk.

anthropomorphic appearances and behaviours, it is possible to encounter the problem of the 'cognitive attribution' phenomenon. The investigators interviewed in the study pointed out that humans interacting with social robots will compare the robots' interactions and attributes more similarly to other living entities (i.e. people or animals), such as the robot being shy, or behaving oddly. Conversely, such human terms are used less in other technical domains where this anthropomorphism or specific type of interaction is not fostered: one would be surprised, for example, to hear a flight crew state that "the flight control system seemed to be a bit aggressive at that time". While human participants may, on a cognitive level, understand that the robot is not alive, nonetheless they might still try to explain the robot behaviour using human attributions. This phenomenon has to do with the mental state attribution, namely a human's ability to understand their own and other person's mental states, such as beliefs, desires, feelings and intentions, and consequently to predict others behaviour [15].

So how do we engage with witnesses of these systems? Investigators should be aware that humans may use mental states rather than computational or physical states to describe the robot behaviour. In particular, they should be aware of the effects that mental states attribution could have on the human witness's behaviour too. Human-robot interaction studies confirm that participants could explain and predict robot behaviour by ascribing mental states to robots [15]. In principle, therefore, the attribution of mental states to robots (such as intentionality) could affect how people interact with them, by influencing their actions/decisions depending on the possibility of anticipating a robot reactions (e.g. "I did this because the robot did not want me to …").

A second issue that we would like to address here is related to ontological confusion, namely the uncertainty concerning the nature of the object in question. What if the uncertainty were not about the robot statements or its behaviour but the very nature of the entity producing such statements or actions? In other words, what if human witnesses could not say with certainty whether it was a robot or human they saw in the scene of the accident? In a research study, Hiroshi Ishiguro discovered that the capacity to distinguish between a human and an android is determined by the presence of micro-movements, which imbue a robot face with a sense of aliveness in comparison to a static face. A time threshold of approximately two seconds was identified as the point at which participants in the study were unable to discern that the entity in front of them was an android [5]. The potential for transparency and explainability to circumvent deception, especially in specific user categories (e.g. vulnerable individuals, including the elderly, individuals with disabilities, and children), has already been considered as a prospective solution [2]. The inability to ascertain whether the entity responsible for an accident was a human or a robot could serve as an additional incentive to emphasise the necessity of consistently disclosing the true nature of the object under scrutiny, whether it be a text, image, or a humanoid robot.

## 4. Conclusions

This paper focused on robots that exhibit human-like reasoning capabilities, physical appearance, and behaviour. We explored the potential implications of employing robots to provide accounts of their actions, observations, and experiences (i.e., robots as witnesses) as well as the role of human observers in describing the behaviour of these robots (i.e., robots as being witnessed), with a particular focus on the context of an accident investigation.

With regard to robots with reasoning capabilities and their potential use as witnesses, the following issues were identified:

- *Reliability of machine evidence*. Testing or challenging robot/machine evidence can be difficult (due for instance to the black box problem) and expensive (affordable only for a few). Legal guidelines and AI education are needed to regulate the potential use of robots as witnesses, taking into account that machine evidence should not affect humans' possibility to defend themselves. There should always be the possibility to inspect and extract raw data from a robot data recorder.
- *Function creep*. This phenomenon occurs when information derived from non-forensic technological systems is employed as evidence. It is erroneous to consider robots with reasoning capabilities as forensic tools unless the aforementioned function has been specifically designed for such a purpose.
- *Education*. A judge lacking familiarity with the functioning of AI may be inclined to accept the statements of a robot as truthful and accurate, thereby rendering robot-generated evidence admissible in a legal context.
- *Automation bias*. It is possible that robot testimony may be perceived as being more accurate than human testimony. However, causal inference frameworks may also be susceptible to bias or other forms of error. It would be inadvisable to rely on robot evidence as the sole basis for an investigation. Those involved in the investigation of robot accidents should be mindful that investigations are social practices.

In regard to the observation of robots with human-like characteristics and conduct, the following issues were identified:

- *Attribution of mental states to robots*. Although human participants may cognitively comprehend that the robot is not alive, they may nevertheless attempt to explain the robot's behaviour using human attributions, which could potentially lead to more ambiguous accident reconstructions.
- *Robot performance*. It is possible to design robots that are capable of exhibiting persuasive and deceptive behaviour. For instance, the physical embodiment of a robot, including the capacity to express emotion, has the potential to significantly influence the emotional response of humans to the robot.
- *Ontological confusion*. In certain instances, human witnesses may be unable to definitively ascertain whether the entity observed at the scene of an accident was a robot or a human. This may result in ambiguity regarding the identity of the perpetrator.

In conclusions, it is likely that future societies will be characterised by high levels of automation, which will result in a growing displacement of human beings by robots. This is likely to occur in a number of areas, including the provision of public services such as goods delivery, urban hygiene, human transportation, shops, restaurants, and so on. One of the consequences of the increasing robotization of societies will be the loss of the moral duty to report and intervene when someone is in need of help. We should consider whether robots should be equipped with ethical behaviours, such as the duty to report and intervene, even at the expense of privacy. Indeed, in the future, robots may well be the only witnesses to accidents and incidents involving humans.

## Acknowledgments

The authors wish to thank the anonymous reviewers of the first version of this paper for their constructive comments and for pointing out critical aspects.

This paper is part of the research project RoboTIPS - Responsible Robotics for the Digital Economy, an Established Career Fellowship awarded to Marina Jirotka by the EPSRC. Project reference: EP/S005099/1.

## References

[1] Cambridge Online Dictionary. Cambridge University Press & Assessment 2024. Available from: https://dictionary.cambridge.org/dictionary/english/accountability

[2] Winfield AFT, Booth S, Dennis LA, Egawa T, Hastie H, Jacobs N, Muttram RI, Olszewska JI, Rajabiyazdi F, Theodorou A, Underwood MA, Wortham RH, Watson E. IEEE P7001: A Proposed Standard on Transparency. Front. Robot. AI. 2021 Jul;8, doi: https://doi.org/10.3389/frobt.2021.665729

[3] Slack D, Krishna S, Lakkaraju H, et al. Explaining machine learning models with interactive natural language conversations using TalkToModel. Nat Mach Intell. 2023 Jul;5:873–883, doi: https://doi.org/10.1038/s42256-023-00692-8

[4] Omeiza D, Anjomshoae S, Kollnig K, Camburu OM, Främling K, Kunze L. Towards Explainable and Trustworthy Autonomous Physical Systems. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21); 2021 May 8-13; Yokohama, Japan: Association for Computing Machinery; 2021. p. 1-3, doi: https://doi.org/10.1145/3411763.3441338

[5] Ishiguro H. Android Science - Toward a new cross-interdisciplinary framework. In Cogsci-2005 workshop: Toward social mechanisms of android science; 2005, Stresa, Italy; 2005. p.1–6.

[6] Bennett B, Daly A. Recognising rights for robots: Can we? Will we? Should we? Law, Innovation and Technology. 2020;12(1):60–80, doi: https://doi.org/10.1080/17579961.2020.1727063

[7] Gless S, Lederer F, Weigend T. AI-Based Evidence in Criminal Trials? Tulsa Law Review. 2024;59(1), doi: https://digitalcommons.law.utulsa.edu/tlr/vol59/iss1/4

[8] Roth A. Machine Testimony. The Yale Law Journal. 2017;126(7). Available from: Yale Law Journal - Machine Testimony

[9] Koops BJ. The concept of function creep. Law, Innovation and Technology. 2021;13(1):29–56, doi: https://doi.org/10.1080/17579961.2021.1898299

[10] Di Gallo J. Defense Rights against Robot Testimony (PhD Project). Available from: https://ius.unibas.ch/de/human-robot-interaction/defense-rights-against-robot-testimony-phd-project/

[11] The Molly Problem. ITU-T Focus Group on AI for Autonomous and Assisted Driving. Available from: https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/MollyProblem.aspx#gsc.tab=0

[12] Zenor J. If you see something, say something: can artificial intelligence have duty to report dangerous behavior in the home? Denver Law Review. 2021;98(4):839-866.

[13] Nakashima E, Albergotti R. The FBI wanted to unlock the San Bernardino shooter's iPhone. It turned to a little-known Australian firm. The Washington Post. 2021. Available from: https://www.washingtonpost.com/technology/2021/04/14/azimuth-san-bernardino-apple-iphone-fbi/

[14] Eliott C, McLaughlin EC. Suspect OKs Amazon to hand over Echo recordings in murder case. CNN. 2017. Available from: Alexa, go ahead and hand over recordings in murder case | CNN Business

[15] Thellman S, de Graaf M, Ziemke T. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. ACM Trans. Hum.-Robot Interact. 2022;11(4), doi: https://doi.org/10.1145/3526112