Human-robot relationships and the development of responsible social robots

Helena Webb
Department of Computer Science, University of Oxford, Oxford, UK, helena.webb@cs.ox.ac.uk

Marina Jirotka
Department of Computer Science, University of Oxford, Oxford, UK, marina.jirotka@cs.ox.ac.uk

Alan F.T. Winfield
Bristol Robotics Lab, University of the West of England, Bristol, UK, alan.winfield@brl.ac.uk

Katie Winkle
Bristol Robotics Lab, University of the West of England, Bristol, UK, k.winkle@brl.ac.uk

ABSTRACT

The contemporary development of social robots has been accompanied by concerns over their capacity to cause harm to humans. Our RoboTIPS study sets out to design and trial an innovative design feature that will advance the safe operation of social robots and foster societal trust. The Ethical Black Box (EBB) collects data about a robot's actions in real time and in context; when an incident occurs, this data can be used within a wider investigation process to determine what went wrong and prevent similar adverse events. In this paper we draw on Lucy Suchman's groundbreaking work on human-machine relationships to elucidate the goals, practices and potential impact of our study. We align with Suchman's positioning of safety as an accomplishment of situated action and draw on her analysis to describe the actions of the EBB-enhanced social robot as contingent on context and the robot's status as a social agent. We also describe shared priorities in our methodological approaches. We close with observations on how participatory design and an ethnomethodologically-informed stance towards data collection and analysis can contribute to the field of responsible innovation (RI), which seeks to ensure that innovations are undertaken in the public interest and provide societal value.

CCS CONCEPTS

• Human-centred Computing • Collaborative and Social Computing • Empirical studies in Collaborative and Social Computing

KEYWORDS

Human-robot interaction, Social robots, Ethnomethodology, Responsible Innovation

1 Introduction: social robots, plans and the practical accomplishment of safety

Over the last decade, advances in robotics and Artificial Intelligence (AI) have led to a rapid rise in the development and use of social robots – those which interact with humans to perform particular tasks in particular contexts [1]. The commercial sector has made significant investments into the production of robots with autonomous capacities that can perform social functions and interact with humans in order to fulfil roles in the home and workplace, as well as in settings connected to leisure, education, healthcare and social care. Perhaps the most well-known example is the driverless car, with over 50 companies approved by the

California Department of Motor Vehicles to test autonomous vehicles on the road [2]. Further examples include personal companion robots, which interact socially with humans for the purpose of education or overcoming isolation [3], and nursing care robots, which assist patients in healthcare environments [4].

Whilst commercial [5], government [6] and media discourses [7] often highlight the opportunities afforded by social robots, this positivity has sometimes been dimmed by the occurrence of negative events. Perhaps inevitably, the growing presence of social robots has been accompanied by a rise in incidents and accidents in which humans are harmed during their interactions with these technologies. Again, the most famous of these have involved driverless cars. In March 2018 during tests in Arizona, an Uber car in autonomous mode failed to identify a pedestrian crossing the road. The car hit the pedestrian, who subsequently died [8]. As of April 2019, there had been 6 recorded human fatalities in connection to driverless cars [9]. At times an atmosphere of fear and anger has developed in counterpoint to more optimistic viewpoints. Members of the public in Arizona have – in response to the fatality of 2018 – attacked self-driving cars being tested in Phoenix [10], and some public and media rhetoric constructs dystopian scenarios in which social robots cause mass unemployment and the usurpation of human autonomy, safety and authority [11].

Given this sometimes negative climate, it is also perhaps inevitable that public and policy discourses around social robots have increasingly called for processes and mechanisms to investigate, mitigate and prevent incidents and accidents. As described by Winfield and Jirotka [12], 10 different sets of ethical principles for robotics and AI, devised in academia and/or industry, had been published by December 2017, and more have been devised since. In addition, the UK government ran inquiries into Robotics and Artificial Intelligence [13], and Autonomous Vehicles [14] in the 2016-2017 parliamentary session. This demonstrates a clear commitment to good practice. However, we can question the extent to which guidelines and formal reports will assure the safer development and operation of these innovations. As Winfield and Jirotka state in [12] '*principles are not practice'* (page 3); the existence of the former does not guarantee standards of behaviour in the latter. Winfield and Jirotka call instead for a more agile and inclusive approach to the ethical governance of social robots by drawing on the field of Responsible Innovation (RI) [15]. The RI approach emphasises the value of seeking the viewpoints of multiple stakeholders in order to ensure that innovations are undertaken in the public interest, and of verification and validation processes to ensure that systems are safe and fit for purpose when they move into real world application.

Taking an ethnomethodological stance [16] helps us to move away from an over-reliance on abstracted principles and guidelines. Instead we can view the safe operation of a driverless car or other social robot as a lived practice. This position is expertly set out in the AI context in the pioneering work of Lucy Suchman [17] [18]. As Suchman describes, plans of action are not synonymous with purposeful action; they are "*constituent as an artefact of reasoning about action, not as the generative mechanism of action."* [18] (page 60). We can understand principles and guidelines for the safe development and operation of social robots as plans and thereby see that they do not generate practice. The safe conduct of a given encounter involving robot and human is not predetermined through any kind of plan but is rather the result of the orderliness constituted in and through the interactions occurring between them. Furthermore, the plans – in the form of the principles and guidelines – are artefacts open to interpretation and may be contested. This may particularly occur when plans are designed to be abstracted and generalised away from practical action and local order.
Following this logic, we argue that understanding and optimising safety in the interactions between humans and social robots requires a contextualised and in-depth understanding of how those interactions occur. When this is achieved, it becomes possible to embed mechanisms into social robots and the local order that can enhance safety in a practical way. In this paper, we describe a novel study of work that sets out to achieve this through the development and application of an innovative design feature. This is the 'RoboTIPS' project[1]

[1] https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/S005099/1

[15]. In this study, we develop an 'Ethical Black Box' (EBB) for social robots. Similar to the flight recorders used in aviation, the EBB collects data about a robot's actions in real time and in context. When incidents and accidents occur this data forms a vital resource to work out what has happened as part of a wider investigation process. This work moves away from dependence on abstract frameworks and guidelines in favour of eliciting a detailed understanding of the everyday practices of those involved in the development, use and governance of social robots.

Suchman's groundbreaking analysis provides a means to conceptualise and examine human-robot relationships in the context of our study. In the remainder of this paper, we describe how we draw on Suchman's arguments in our own work and how her understandings of human-machine interactions are reflected in our vision of the EBB-enhanced social robot. We start with a more detailed description of our RoboTIPS study and its aims. We then foreground context in our understanding of social robots and describe them as social agents, with consequences for agency and accountability. Our methodological approach, again mirroring Suchman, emphasises in-depth, context-based analysis and also incorporates participatory design. We argue that this approach has valuable synergies with Responsible Innovation (RI). In particular, work with stakeholders makes the notion of responsibility explicit and ethnomethodologically-informed analyses of accidents and accident investigation processes can serve to foster responsibility in innovation as practical action. Our paper illustrates the continuing relevance of Suchman's work, and ethnomethodology more broadly, to the increasing focus on AI in contemporary sociotechnical research and discussions of best practices in innovation.

2 The Ethical Black Box (EBB) for Social Robots

The RoboTIPS study forms part of the UK Engineering and Physical Sciences Research Council's Digital Economy programme. It takes up the challenge of exploring how issues of responsibility might be embedded into processes of technological design and development. The field of social robotics is used as a spotlight example to explore this question in detail and Responsible Innovation (RI) is drawn on as a resource to identify suitable activities for analysis and application. RI proposes anticipatory governance whereby the potential positive and negative consequences of an innovation may be anticipated and brought to developers' attention in order to influence the development trajectory of that innovation. The process of anticipation in RoboTIPS [19] identified the potential for harmful encounters between humans and social robots, and highlighted that these encounters would damage societal trust. The process also led us to look to a more trusted industry to see what might be learnt from it. The industry we considered was aviation.

The aviation industry has well-established safety procedures developed over time that help to generate (but of course cannot guarantee) trust with users of various kinds of aviation technologies [20]. At the centre of these procedures is the 'black box' flight recorder. This typically consists of a flight data recorder, which records the recent history of a flight in the form of key data parameters collected several times per second, and a cockpit voice recorder, which preserves the conversation in the cockpit between pilots as well as other sounds occurring. When an incident or accident occurs the flight recorder provides valuable evidence that can be accessed and interpreted. The data from the flight recorder is drawn on as part of a wider accident investigation process involving human experts and witnesses. These investigations establish the cause of individual accidents, help to rule out systematic failure modes, and produce lessons to maximise safety across a particular model or the entire industry. The development of the flight recorder has been described as one of the most crucial inventions in the history of safety engineering [21] and its role as a data source in the wider process of investigation helps to preserve trust in aviation.

In RoboTIPS we take the aviation context as a starting point and seek to translate this trust-enabling process into the domain of social robots. We do not assume that all the specifics of this process – several of which have developed over periods of time – can be transferred with immediacy but we are interested to see what we can learn from the practices involved and how they might be adopted and adapted in a new context. With

this mind, the study will develop a recorder for social robots. We call this recorder the Ethical Black Box (EBB). The EBB collects data about a robot's actions in real time and in context. This will include the AI decision-making of the robot and the environmental factors occurring *in situ*. When problematic incidents occur, the EBB provides a mechanism for the robot to produce an account of its own actions. The inclusion of a natural language explainer system will enable the robot to explain its own behaviour; eventually we want to allow the robot's users to be able to ask 'why did you just do that?' or even 'what would you do if…?' and the robot to give a simple, intelligible explanation. Of course the robot's own account will not be considered a neutral one to be straightforwardly taken at face value. Crucially, it is not the black box on its own that forms the safety mechanism; it is the inclusion of the black box within a social process of investigation. An investigation will draw upon EBB information amongst other information, including that of human witnesses and experts, to determine the reason for an incident and any lessons to be learnt from it. Hence, alongside the technical parameters of what to record within an EBB, our research will also consider how the interpretation of those recordings fits into the conduct of an investigation. We will devise a series of scenarios in which a robot with an EBB functions and then 'malfunctions' or 'misbehaves' in some way. We will then constitute a mock investigation with witnesses to reconstruct what happened and produce findings and recommendations as a real inquiry might. This mock scenario process forms a vital way for our study to trial the practical use of a safety mechanism for social robots in context and is, we believe, the first of its kind.

In order to develop a viable EBB and workable mock investigation procedures, the study will be grounded in understandings of the everyday practices and lived experiences of designers and developers, as well as the users of social robots themselves, plus those involved in investigation procedures. This approach builds on a rich tradition of sociological work that recognises the social side of technological risk. Most influential in this tradition is the work of Perrow [22], who provides a detailed analysis of complex systems and a framework for analysing the risks that emerge from them. Like Perrow, we explore the interweaving of the social and the technical in the occurrence of incidents and accidents, and their prevention. We also benefit from substantial bodies of work conducted in the fields of human-robot interaction [23] and human-computer interaction [24]. Since our focus is on 'organisation' in terms of local orderliness achieved moment by moment through social encounters involving human and machine, we find Suchman's work to be particularly illuminating. In the remainder of this paper we draw on her arguments to elucidate the goals, practices and potential impact of our study.

3 RoboTIPS: understanding the EBB-enhanced social robot in interaction

Published in 1987, Suchman's 'Plans and Situated Actions: The problem of Human-Machine Communication' provided a pivotal critique of then established understandings and practices in AI and HCI. Suchman highlighted the fallibility of the dominant planning model, which was anchored in cognitive science and which treated actions as a form of problem solving and plans as a sequence of actions designed to achieve a particular, preset goal. She described how this model fails to take into account the highly context specific and contingent nature of human action. Drawing on ethnomethodology, Suchman argued – and demonstrated through empirical analysis - that the situatedness and localness of practical action enables interactants to accomplish tasks and activities, often in collaboration with each other and with the use of different kinds of tools and technologies. We can therefore understand plans as resources for action, interpreted and drawn on by participants in their particular local contexts. The 2007 'Human-Machine Reconfigurations: Plans and Situated Actions' reproduces the earlier text with added commentary and contains further chapters to expand and update Suchman's arguments. Suchman widens the focus of her research interests to encompass a range of artefacts in addition to plans. She also describes her '*renewed interest in questions of machine agency…inspired by contemporary developments both in relevant areas of computing and in the discussion of human-nonhuman relations within social studies of science and technology*' [18] (page 2). She points to the need to avoid an essentialist machine human divide and is enthusiastic about bringing in new conceptual understandings to reconfigure human-machine relations, practices and projects.

Suchman's analysis of the relationships between human and machine provides a lens through which we can conceptualise human and EBB-enhanced social robot interaction in our study. We have already described how her critique of plans allows us to recognise the limitations of guidelines and frameworks and move towards a focus on safety as the outcome of practical action. Next we go deeper into Suchman's work to draw out how our vision of the EBB-enhanced social robot corresponds with her analysis of human-machine relationships. We begin by following Suchman to foreground context in our understanding of the social robot.

3.1 The contingencies of context

The actions of a driverless car, companion robot or robot nurse are contingent on the context in which they occur. They are not pre-set but emerge through interaction with others on the scene. Within any interaction with humans, the social robot will draw on its programming and information available in the local context – such as the utterances and visible actions made by humans – as resources to shape moment by moment its actions. Similarly, humans on the scene will draw on available contextual detail – such as the utterances and visible actions performed by the robots – and their own background experiences, perhaps including assumptions about how robots 'do' and 'should' behave. These will provide resources for the humans to draw on to shape their own actions. Since the way that a particular social robot interacts with humans when carrying out its 'functions' will be contingent on those interactions, as a consequence, so will the data the EBB collects. Therefore, we would expect the data collected by an EBB for a (mock) investigation to differ, even if the scenarios in which the robots were placed were formally designed to be identical. In turn, the ways in which the data provided by the EBB are drawn on in the investigation will be contingent on the context of that investigation and how those involved interact with each other. We can therefore see that understanding the meaningful actions of social robots relies on an appreciation of context, and that understanding how an EBB-enhanced social robot may function as a safety mechanism relies on an exploration of its actions and interactions as they occur within specific local contexts.

We also find it fruitful to take up Suchman's rejection of an essential human versus machine divide in the recognition that is through contextually grounded actions and interactions that 'machine-ness' and 'robotness' are constituted in any given encounter. The robot is a product of temporally and contextually grounded processes and the extent of its 'robotness' is not predetermined. For instance, in a scenario where a human user has a conversation with a companion robot, the category of 'robot', as relevant to their interaction, is a product of their interaction rather than merely the material features of the technology. In the 2007 edition of her work, Suchman takes up ideas from Actor Network Theory (ANT) [25] to argue that when we move away from the categorical human-machine divide, we also recognise that our unit of understanding (and analysis) is not the single machine, but the network that it exists within. If humans and machines are mutually constituted, we need to develop an understanding of their sociomaterial assemblages. Whilst acknowledging that significant differences do exist between them, we recognise that ANT and ethnomethodology share an interest in understanding the role of the non-human in interaction [26] and find this interest in networks of use to our own study to some extent. We can understand the EBB-enhanced robot as the product of a network of activities and as embedded within context-dependent networks encompassing human and machine. We particularly see the EBB-enhanced robot as belonging or potentially belonging to multiple strands of network simultaneously. When in use, the robot is embedded within a network that is related to and makes visible its particular function – for instance, a transport network, a hospital network and so on. It also carries within it the 'hidden labours' of the designers and developers who created it, and the way it functions is highly contingent on the work of those actors. At the same time, there also exists a potential future network of those involved in an incident or accident investigation scenario. This network includes witnesses, experts, regulators, adjudicators and perhaps other social robots or technologies. This potential network may be called into being at any time. Therefore, its potential to exist is of relevance to the ongoing activities in the other strands of the network; the robot's possible role within an investigation is of necessity embedded into its design and its use in everyday activities.

3.2 The social robot as a social agent

Once we have drawn on Suchman to understand the social robot as existing within a network and its actions as contingent on context, we can then consider her arguments about machine agency. As noted above, the 2007 work includes a renewed interest in machine agency; she follows ANT to position it as neither inherent to, or located within, humans or artefacts. Instead it is distributed and enacted. Again we recognise that differences exist between ANT and ethnomethodological positions. However, it is crucial to acknowledge that contemporary developments in AI have created new forms of machine human relationships, in which machines are able to exercise a degree of autonomy. A social robot is a social agent. It is not capable of intersubjective reasoning in the way a human is but nevertheless it is able to make decisions, and these decisions plus the actions that follow it have social consequences. We can usefully conceptualise the social agency of the robot as mutually constituted with its materiality [27] but our chief interest here is empirical. We are interested to observe how the EBB-enhanced robot is constituted as having agency in its interactions with humans; or as Suchman states '*how the effect of machines-as-agents is generated*' [18] (page 2).

The enacted agency of social robots is of central interest to the RoboTIPS study, with particular reference to how agency also confers accountability. As a social agent, a social robot might be oriented to by human interactants as accountable for its actions, with consequences for the ongoing encounter. Once again, this does not suggest that a social robot is capable of human-like reasoning, but that the humans in an encounter might treat a robot as accountable for some unexpected or disruptive behaviour and alter their own behaviours accordingly. For instance, a human might make a robot accountable for a nonsensical conversational turn by producing a quizzical response, or make a driverless car accountable for passing too close to a pedestrian through the conduct of evasive action and gesturing. The role of accounts lies at the heart of the EBB-enhanced social robot. When it is involved in an incident or accident, the robot is treated in the local context as accountable to produce an explanation for its actions by those investigating what happened. It is not regarded as the moral perpetrator of the adverse event and does not provide an explanation grounded in human reasoning. However, the robot does provide an account - in the form of information about its actions, the local context and the traces of its own design and development - that allows human investigators to piece together what happened. In this way the EBB serves to help map out where responsibility for an incident or accident lies. The specific processes through which this is achieved will be contingent on context. We note that the robot's role transforms as the situation develops. When conducting its first function as a social robot, the robot may be treated as accountable for its behaviours plus also accountable to collect information relevant to the immediate context and simultaneously to collect information to enable sense-making in the potential future context of an investigation scenario. Later, it may be treated as accountable to produce this information in a way that is meaningful to human participants in the investigation. This is an area of great fascination that draws parallels with the work of Garfinkel [28] and subsequent work in the ethnomethodological tradition on record keeping [29]. It is a necessary task of the social robot and the network of actors that develop and configure it to produce within the EBB a record of information that is ordered, orderly and intelligible in two separate contexts. At the empirical level we are interested in how the EBB-enhanced robot can produce an account that helps an ongoing investigation. As the project unfolds we will explore the interactions in which the robot produces these accounts and seek to identify how to best shape the development of the EBB so that the explanations the robot produces are locally coherent and aid the successful accomplishment of the investigation.

3.3 Methodological considerations

Suchman's work is groundbreaking methodologically as well as conceptually. Her original study included detailed analyses of interactions between users and an 'intelligent' machine. These demonstrated the value of naturalistic and quasi-naturalistic research methods, in particular the use of quasi naturalistic experiments to expose and identify unexpected aspects of social organisation when users interact with a new technology [30]. Following ethnomethodolgical principles, Suchman emphasises: i) conducting empirical work grounded in recognition of context; ii) observing interaction to identify how categories of machine and human

are enacted in and through actions occurring within the local sequential and material order; and iii) exploring how troubles are identified and dealt with in context. She also highlights the contribution of participatory design in technological development. Her own work demonstrates this value by capturing how the designers of expert systems came to consider their innovations more in terms of human-machine interaction and usability.

We similarly prioritise these elements and employ them in RoboTIPS to optimise our understanding of human -robot relationships and the operation of the EBB-enhanced social robot. Our empirical work begins with ethnographic studies of designers and developers of social robots. This will help us to understand their everyday activities and sense-making, as well as the constraints they operate under. In combination with the stakeholder engagement described below, this will also elicit a set of user-focused requirements for the EBB [31]. Our ethnographic work will focus on the particular domains in which the EBB will subsequently be tested, thereby ensuring we gather necessary context-relevant knowledge. The EBB will be trialled empirically. Our mock investigations will be quasi-naturalistic encounters beginning with interactions between the social robot and human users, and then involving the social robot in an investigation process that includes humans taking the roles of investigators, experts and other witnesses. Our emphasis throughout will be on understanding the configurations of human and EBB-enhanced social robot as the product of situated action. We will observe interactions between humans and the social robot within the specific context in which they occur and draw on these observations to understand the occurrence of adverse incidents and the process of investigation that follows them. This exploration of how troubles are identified and dealt is central to our study. In the first instance, we are interested to observe how the 'misbehaviour' or 'malfunctioning' of a social robot becomes visible and is dealt with in the moment by moment unfolding of interaction. Next we are then interested to observe how this misbehaviour or malfunction is addressed in the later interactional context of the investigation process, in particular the ways in which the EBB provides contextually relevant information to aid this investigation. This includes a keen interest in how accountability for the event, and accountability for the production of information relevant to the event, is established.

We will embed participatory design activities across the entire timescale of the project. These will encourage reflexive awareness amongst developers and others of their own role in the innovation process. We see great resonance between the aims of participatory design and Suchman's deployment of it, and the RI approach described earlier in the paper. The notion of responsibility for the development of an innovation is implicitly embedded into participatory design. RI makes this explicit and identifies steps to make recognising responsibility practically achievable [31]. We will include a wide range of relevant stakeholders - policy makers, regulators, researchers, and educators amongst others - alongside the designers and users of the technology in various workshops, roundtables, surveys and focus groups. At times these activities will focus on specific design issues connected to the EBB and/or the requirements of a particular domain. At other times they will be broader in scope, for instance eliciting and circulating issues associated with being responsible in the contemporary digital economy. Towards the end of the project, we will run 'speculative design' exercises where we re-imagine different sorts of novel institutions, technologies, and policy environments as a tool for fostering an appreciation of the issues and generate creative solutions amongst designers, innovators, policy makers and so on. These speculative designs will take a number of forms, including video material and demonstrator systems, and will be used across our networks of stakeholders to stimulate discussion and impact.

4 Conclusions: fostering the development of responsible robots

Suchman's groundbreaking work offers a vivid account of human-machine relationships. It also provides a resource to conceptualise and analyse sociomaterial arrangements in different contexts. We have drawn on Suchman's arguments to set out the work of our RoboTIPS study, which develops an innovative safety design feature, the Ethical Black Box (EBB) for social robots, and embeds social robots into the incident and

accident investigation process. Suchman's description of plans '*as an artefact of reasoning about action'* rather than the '*generative mechanism of action*' [18] (page 60) helps us to move away from a focus on formalised safety and ethical principles for social robots and to understand that safety is a practical accomplishment, achieved within a local context. As human-robot encounters are contingent on the interactions that occur between them, so too the data collected by the EBB in the social robot, and the outcomes of the investigation process drawing on information provided by the EBB, will be context-dependent. The incident investigation process we describe is a safety and transparency mechanism that can serve to foster trust. The EBB-enhanced social robot plays a pivotal, and pivoting, role in this process. In its first function it serves as a driverless car, companion or nurse etc. in interaction with humans but later transforms to provide information about an event that it was itself part of.

Our study will interrogate the technically constructed EBB from within a social frame of reference. We are very interested in the EBB-enhanced social robot as a social agent that may be oriented to as accountable to collect information in its EBB that is both coherent in real time and that can also be meaningfully interpreted during an investigation at any future time. This accountability is enhanced by the existence within the EBB of a natural language explainer, meaning that the social robot can produce a verbal account of its actions. By extension, we can also consider the developers of the robot as also accountable to some degree, and our study seeks to establish processes through which responsibility is mapped out following an adverse event. Our mock investigations play a key role here. It is when something goes wrong with a technology that we are able to see what the chains of responsibility are; we are able to do this by observing how these chains are made visible and are enacted across the networks of human-machine actors. Our aim isn't to develop the 'perfect' explainer in an EBB, but to polish a state-of-the-art explainer and to explore how it fairs within the social context of a simulated incident and investigation process. It is to 'learn-by-doing' what the parameters are for an explainer to be integrated into a social process and to learn what role it comes to play. Taking this further, our work can contribute to current debates about trust and transparency in Artificial Intelligence (AI). The EBB is a motivation for exploring how investigations around the failure of AI might be constituted as a future tool in the wider AI regulatory framework.

Suchman highlights the methodological approaches needed to understand human-machine configurations in context and in depth. We agree that empirical work and observations of interaction provide a fruitful approach to ground understandings of human-machine relationships in context and to identify how categories of machine and human are enacted moment by moment. We also agree that the exploration of how troubles occur and are resolved provides great analytic value and share Suchman's enthusiasm for participatory design in technological development. We take this further by making the notion of responsibility explicit as part of an agile process that takes the viewpoints of societal stakeholders into account in order to develop and implement an innovative design feature. We regard this as an instance of responsible innovation (RI) in action. We close with some observations on the synergies between ethonomethodology and the vision of responsible innovation.

The RI agenda has been strongly promoted by the EU, and national funding institutions are also increasingly adopting RI principles and encouraging RI for their funded research[2]. RI is positioned as an opening up the innovation space to a range of stakeholders and facilitating practices to engage with them. This in turn allows the opportunities and challenges inherent to the innovation process to become visible and to be addressed creatively. RI proposes a new approach to research and innovation governance and seeks to facilitate a more adaptive, human-centred, inclusive and sustainable research and innovation process. This is to be applied from fundamental research through to application design and beyond. RI work so far has produced

[2] https://epsrc.ukri.org/research/framework/

frameworks that researchers and innovators can draw on as a resource to pursue responsibility in their work. The contribution of ethnomethodology can take this further. The ethnomethodological focus on the detailed observation and understanding of context aligns closely with RI as a deep understanding of context is necessary to achieve responsibility in innovation. Attention to situated action allows us to understand human and machine practices, gauge user needs, and observe how innovations are used in action, including how problems are emerged and resolved. Drilling down into context combines with the ethnomethodolgical focus on practical action. This provides a way to avoid broad, overarching plans such as guidelines etc. to govern innovation in favour of more flexible and responsive mechanisms, such as the EBB for social robots. In our RoboTIPS study, we draw on social robots as a spotlight example of how responsible innovation may be achieved in action; in doing so, we hope also to illuminate ways that RI can be informed by ethnomethodology.

ACKNOWLEDGMENTS

The 'RoboTIPS: Developing Responsible Robots for the Digital Economy' study is an EPSRC Established Career Digital Economy Fellowship awarded to Marina Jirotka. Project ref: EP/S005099/1.

REFERENCES

[1] Satyandra K. Gupta (2015). Six recent trends in robotics and their implications *spectrum,ieee.org,* 8th Sep 2015, https://spectrum.ieee.org/automaton/robotics/home-robots/six-recent-trends-in-robotics-and-their-implications.

[2] Andrea Miller (2018). Some of the companies that are working on driverless car technology. *abcnews.go.com*, 21st March 2018. https://abcnews.go.com/US/companies-working-driverless-car-technology/story?id=53872985.

[3] The Medical Futurist (2018). The Top 12 Social Companion Robots, *medicalfuturist.com,* 31st July 2018, https://medicalfuturist.com/the-top-12-social-companion-robots.

[4] Riken (2015). The strong robot with the gentle touch, riken.jp, 23rd February 2015, http://www.riken.jp/en/pr/press/2015/20150223_2/.

[5] KPMG Advisory N.V. (2016). Social Robots, *kpmg.home,* https://assets.kpmg/content/dam/kpmg/pdf/2016/06/social-robots.pdf.

[6] Tom Kalil and Sridhar Kota (2011). Developing the Next Generation of Robots, *obamawhitehouse.archives.gov,* June 24th 2011, https://obamawhitehouse.archives.gov/blog/2011/06/24/developing-next-generation-robots.

[7] Marc Ambasna-Jones (2016). How social robots are dispelling myths and caring for humans, *theguardian.com,* 9th May 2016, https://www.theguardian.com/media-network/2016/may/09/robots-social-health-care-elderly-children.

[8] Sam Levin and Julia Carrie Wong (2018). Self-Driving Uber kills Arizona woman in first fatal crash involving pedestrian, the*guardian.com,* 19th March 2018, https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

[9] Wikipedia (2019). List of self-driving car fatalities, *wikipedia.org,* accessed 24th June 2019, https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities.

[10] Oliver Rudguard (2018). Arizona residents attack self-driving cars, *telegraph.co.uk,* 13th December 2018, https://www.telegraph.co.uk/technology/2018/12/13/arizona-residents-attack-self-driving-cars/.

[11] Dom Galeon (2017). Bill Gates: Benefits of robots, healthcare AI, will outweigh pitfalls, *futurism.com,* 16th November (2017), https://futurism.com/bill-gates-benefits-robots-healthcare-ai-outweigh-pitfalls/.

[12] Alan F.T. Winfield and Marina Jirotka (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems, *Phil. Trans. R. Soc. A.* 376:20180085. doi.org/10.1098/rsta.20180085.

[13] Parliament UK (2016). Robotics and artificial intelligence inquiry launched, *parliament.uk,* 24th March 2016, https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/news-parliament-2015/robotics-and-artificial-intelligence-inquiry-launch-15-16/

[14] Parliament UK (2016). Driverless vehicles – where are we going?, *parliament.uk,* 15th September 2016, www.parliament.uk/business/committees/committees-a-z/lords-select/science-and-technology-committee/news-parliament-2015/autonomous-vehicles-inquiry/
[15] Rene Von Schomberg (2013). A vision of responsible research and innovation. *Responsible innovation: Managing the responsible emergence of science and innovation in society* (2013), 51–74.
[16] Harold Garfinkel (1967). *Studies in Ethnomethodology*, Englewood Cliffs, N.J.: Prentice Hall.
[17] Lucy Suchman (1987) *Plan and Situated Actions: the problem of human-machine communication,* Cambridge, Cambridge University Press.
[18] Lucy Suchman (2007). *Human-Machine Reconfigurations: Plans and Situated Actions, 2nd Edition,* Cambridge, Cambridge University Press.
[19] Alan F.T. Winfield and Marina Jirotka (2017). The Case for an Ethical Black Box. In: Gao Y., Fallah S., Jin Y., Lekakou C. (eds) Towards Autonomous Robotic Systems. TAROS 2017. Lecture Notes in Computer Science, vol 10454. Springer, Cham.
[20] UK Civil Aviation Authority (2013). *Global Fatal Accident Review 2002 to 2011,* June 2013 http://publicapps.caa.co.uk/docs/33/CAP%201036%20Global%20Fatal%20Accident%20Review%202002%20to%202011.pdf.
[21] Krishna M. Kavi (2010) Beyond the Black Box, *IEEE Spectrum,* August 2010, https://csrl.cse.unt.edu/kavi/Research/Spectrum-Aug-2010.pdf.
[22] Charles Perrow (1984). *Normal Accidents,* New York:Basic Books.
[23] Takayuki Kanda and Hiroshi Ishiguro (2017) *Human-Robot Interaction in Social Robotics,* Boca Raton, CRC Press, doi:10.1201/b13004.
[24] Alex Roney Mathew, A. Al Hajj and A. Al Abri, Human-Computer Interaction (HCI): An overview, *2011 IEEE International Conference on Computer Science and Automation Engineering*, Shanghai, 2011, pp. 99-100. doi: 10.1109/CSAE.2011.5953178
[25] Bruno Latour (1987). *Science in Action: How to follow scientists and engineers through society* Cambridge, MA.:Harvard University Press
[26] Eric Laurier et al. (2019) The "Studies in Ethnomethodology" Are a Way of Understanding and Handling Empirical Materials and Thoughts. Eric Laurier in Conversation With Hannes Krämer, Dominik Gerst and René Salomon**.** *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, [S.l.],20,2, May 2019. ISSN 1438-5627.
[27] Morana Alač (2016). Social robots: Things or agents? *AI & Soc* 31, 519, doi 10.1007/s00146- 015-0631-6.
[28] Harold Garfinkel and Egon Bittner (1967). Good organizational reasons for bad clinic records. In Harold Garfinkel (Ed.) *Studies in Ethnomethodology*, Englewood Cliffs, N.J.: Prentice Hall, 186-207.
[29] Christian Heath and Paul Luff (1996). Documents and Professional Practice: 'bad' organizational reasons for 'good' clinical records, *Proceedings of the Conference on Computer Supported Cooperative Work*, Boston, ACM Press 1996, 354-363
[30] Christian Heath and Paul Luff (2017) The Naturalistic Experiment: Video and Organisational Interaction, *Organizatuonal Research Methods,* 21, 2, 466-488.
[31] Barbara Grimpe, Mark Hartswood and Marina Jirotka. 2014.Towards a closer dialogue between policy and practice: responsible design in HCI. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (Toronto, Canada, April 26-May 01). ACM Press, New York, 2014, 2965-2974