Check for updates

Ethical Issues and Unintended Consequences of Digitalization and Platformization

JIT

Journal of Information Technology
2023, Vol. 0(0) 1–29
© Association for Information Technology Trust 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02683962231207108
Journals.sagepub.com/jinf
Sage

Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems

Frantz Rowe[1], Maximiliano Jeanneret Medina[2,3], Benoit Journé[1], Emmanuel Coëtard[1] and Michael Myers[4]

Abstract

Autonomous driving systems (ADS) operate in an environment that is inherently complex. As these systems may execute a task without the permission of a human agent, they raise major safety and responsibility issues. To identify the relevant issues for information systems, we conducted a critical and scoping review of the literature from many disciplines. The innovative methodology we used combines bibliometrics techniques, grounded theory and a critical conceptual framework to analyse the structure and research themes of the field. Our findings show that there are certain ironies in the way in which responsibility for apparently safe autonomous systems is apportioned. These ironies are interconnected and reveal that there remains significant uncertainty and ambiguity regarding the distribution of responsibility between stakeholders. The ironies draw attention to the challenges of safety and responsibility with ADS and possibly other cyber-physical systems in our increasingly digital world. We make seven recommendations related to (1) value sensitive design and system theory approaches; (2) stakeholders' interests and interactions; (3) task allocation; (4) deskilling; (5) controllability; (6) responsibility (moral and legal); (7) trust. We suggest five areas for future IS research on ADS. These areas are related to socio-technical systems, critical research, safety, responsibility and trust.

Keywords

autonomous driving systems, autonomous cyber-physical systems, artificial intelligence, safety, responsibility, trust, predictability, literature review

Introduction

Although autonomous driving systems (ADS) are often introduced with the rationale that they are safe, this is not always the case. People can be killed using semi or fully autonomous cars. Following one of the first fatal Tesla crashes in 2016 (Banks, Plant and Stanton, 2018), there have been more than 17 fatalities and 700 crashes involving Tesla's Autopilot feature since then (*Washington Post*, 10 June 2023, '17 fatalities, 736 crashes: The shocking toll of Tesla's Autopilot'). The digital world in which ADS and other autonomous cyber-physical systems (ACPS) operate can be viewed as dystopian if the collaboration required with humans is too complex (Jiao et al., 2020).

For years Tesla, Waymo and many traditional automotive manufacturers have been experimenting with autonomous cars. Autonomous cars are presented as one of the next disruptive changes in our lives. However, beyond the remarkable technical achievements needed to operate such cars, human and societal adaptation to them requires solving considerable ethical and legal challenges related to safety and responsibility. As most research on this topic ADS has been published in other disciplines, we want to understand the relevance of this research for IS. As a socio-technical discipline (Sarker et al., 2019), we suggest that IS can contribute to solving some of these challenges, such as those related to trust. For autonomous cars to be accepted, people need to be convinced that autonomous cars will be

[1]Institut d'Administration des Entreprises (IAE, Economics and Management), Nantes University, LEMNA, Nantes, France
[2]Institute for the Digitalisation of Organisations, HEG Arc, HES-SO, University of Applied Sciences Western Switzerland, Neuchâtel, Switzerland
[3]Human-IST Institute, University of Fribourg, Fribourg, Switzerland
[4]Department of ISOM, University of Auckland, Auckland, New Zealand

Corresponding author:
Frantz Rowe, Management, Nantes University, LEMNA, Chemin de la Censive du Tertre, BP 52231, 44322, Nantes Cedex 3, France.
Email: Frantz.rowe@univ-nantes.fr

trustworthy and safe. Hence, in this paper we discuss trustworthiness and the allocation of responsibility for autonomous driving, focusing on the ethical and legal challenges related to safety.

In our study we consider autonomous driving systems (ADS) as systems with specific agentivity (or agentic IS artifacts) such as capabilities of anticipation and prescription (Baird and Maruping, 2021). Naturally, autonomy (Vagia et al., 2016) and automation (Parasuraman et al., 2000) may vary within this category of systems. Operationally, fully ADS would be equipped with artificial intelligence and an Internet of Things (IoT) infrastructure enabling them to interact with the environment without any human intervention.

A recent report by MIT says that 'few sectors better illustrate the promises and fears of robotics than autonomous cars and trucks. Autonomous vehicles (AVs) are essentially highspeed industrial robots on wheels, powered by cutting-edge technologies of perception, machine learning, decision-making, intelligence ethics, regulation, and user interfaces' (Autor et al., 2020, p. 39–40). There are many challenges associated with the increasing use of robots and ADS.

First, in the past autonomous systems such as robots were confined to a closed environment such as a factory. Today, however, ADS operate in an environment that is open. This environment is uncertain, unpredictable and inherently complex. This complexity is a feature of the environment, not the system per se. For example, an autonomous car might be safe and reliable on a freeway, but what about on a country road or city centres? There might be animals, pedestrians, bikes, or scooters, and people might not always obey the road rules. People's driving habits vary depending on the country.

Second, ADS may execute a task without the permission of a human agent. Although a human driver might hope that the ADS will always follow directions, the autonomous nature of the system raises issues of control and delegation. Will an ADS always obey the human driver? Sometimes it might not.

Third, how can we understand responsible design in such a scenario? What does responsibility mean when the environment is so uncertain? Responsible design raises numerous questions related to ethics, such as who is morally responsible or legally liable in the case of an accident.

This suggests that there are many societal challenges associated with the introduction of ADS (Ketter et al., 2022). Hence, the purpose of this paper is (1) to provide a better understanding of the issues of safety and responsibility associated with the introduction of ADS and (2) to propose a research agenda for the information systems field. This paper thus contributes to answering the call for more IS research on ADS (Ketter et al., 2022; Lyytinen et al., 2022). Our two research questions are as follows: (1) *How do ironies of automation manifest when levels of automation of ADS increase beyond what we currently experience?* And (2) *Who is responsible and what does responsibility mean in the context of autonomous driving systems?* To answer these questions, we review the literature with a focus on (a) autonomous driving systems, (b) responsibility and (c) safety. This review reveals there are certain ironies in the way in which responsibility for apparently safe autonomous systems is apportioned (Noy et al., 2018). These ironies draw attention to the challenges of safety and responsibility with ADS in our increasingly digital world. Although this paper focuses on autonomous driving only, we believe our findings might be relevant to researchers doing work on robotics and ACPS in general.[1]

This paper is organized as follows. After proposing a conceptual framework that will be used as a critical lens to assess the literature, we explain our methodology. The findings section reports on the research related to safety and responsibility for ADS. We then make eight recommendations for future research and identify six specific areas for future IS research.

Conceptual framework

Our conceptual framework considers automation and autonomy as synonymous. At certain levels of automation, endowed with certain capabilities, the system will perform certain tasks autonomously (Hancock, 2019). However, human experience will encounter ironies of automation stemming from the delegation of certain tasks and from the ensuing confusion between who has control and who is morally or legally responsible. Inevitably, these ironies evoke issues of safety, responsibility, predictability and trust. Hence, to address these issues and to answer our research questions, we propose an integrative framework (see Figure 1).

ADS levels and ironies of automation

ADS levels. The ADS literature builds upon that of automated systems (Parasuraman et al. 2000; Sheridan, 1992; De Winter and Dodou, 2014). A typology of six levels of automation proposed by SAE International (2016) –previously known as the Society of Automotive Engineers –is largely used in the autonomous driving literature. The five levels start at level zero 'no automation' through to level 5 'full automation' (see Table 1 for a condensed version, where level zero is not presented). Level 1 is driver assistance: some assistance can be provided for either acceleration/deceleration or steering to keep safe longitudinal and lateral distance. Level 2 relates to partial automation where the car assumes control of both steering and acceleration/deceleration. The driver remains responsible for monitoring the driving environment and fallback

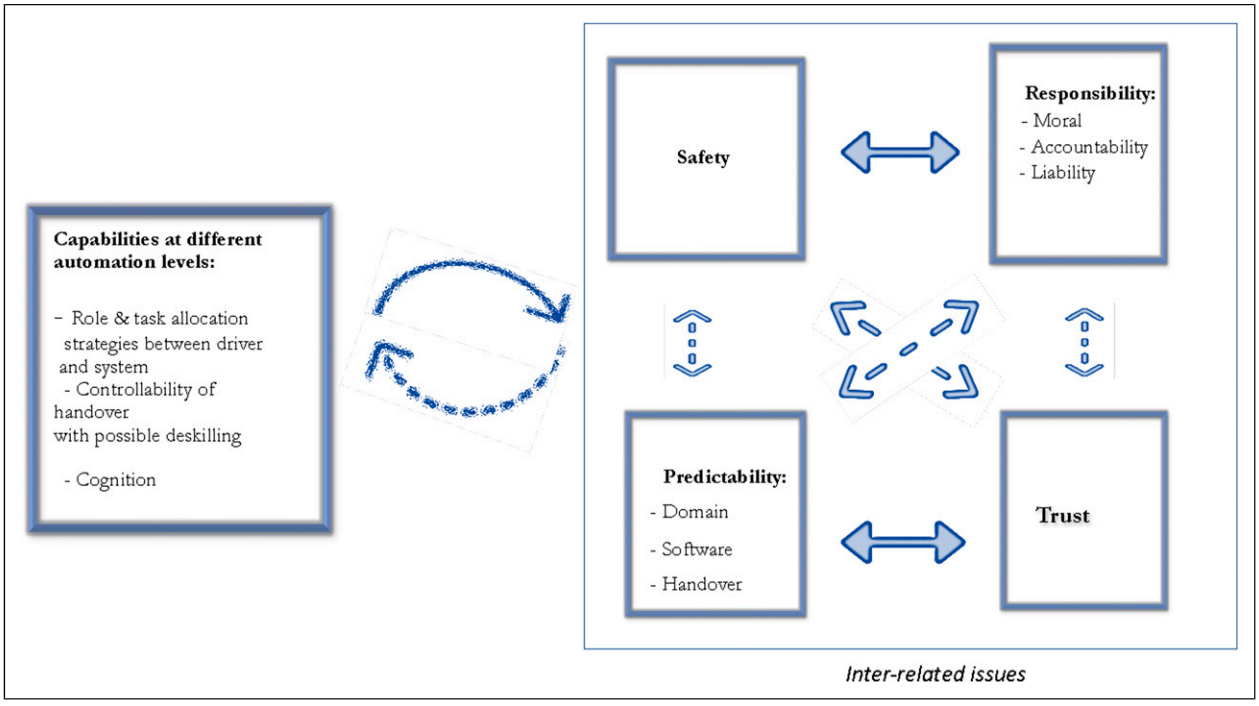

Figure 1. The ironies of ADS: a conceptual framework.

performance. Level 3 is conditional automation: the vehicle is responsible for monitoring the environment. However, the driver is required to be receptive to alerts, or other driving relevant system outputs, and is expected to respond if there is a request to intervene. In level 4, called high automation, the system assumes control for specific dynamic driving tasks and/or within a specified area (Operational Design Domain – ODD) even if the human driver does not respond to a request to intervene. Fallback performance now lies with the vehicle, which means that in case of an emergency, or if the driver does not respond to a request to intervene, the vehicle automatically assumes control. Finally, level 5 signifies full automation: all driving tasks are undertaken by the system and the driver has no responsibility for monitoring the environment. Essentially, it is level 4 with an unlimited ODD. Fallback performance lies with the vehicle, but the driver can intervene and manually request the vehicle to achieve a minimal risk condition.

In the human factors literature (e.g. Sheridan, 2011) as in the IS literature (Baird and Maruping, 2021), the allocation of control to a human or to an artificial agent is to accommodate changes in the conditions of either the physical environment or the human. Fully ADS are still at the experimental stage, while partial automation driving systems are already used in mixed traffic conditions (Cabrall et al., 2019). As of July 2023, all commercially available cars have achieved level 2 only. It is important to clarify that in this context, autonomous 'only makes sense if it refers to the relationship between human driver-passenger and the vehicle –the vehicle is increasingly autonomous from the driver, not to the relationship between the vehicle and the traffic environment'. (Lee and Hess, 2020, p.87). In fact, ADS need to be constantly connected to sensors equipping other vehicles surrounding them. A major difference between level 2 and levels 3 and above is that for latter levels it is the ADS not the driver that monitors the driving environment, except when the system requires the driver to take over (fallback performance) at level 3 (Hancock, 2019). This typology has been criticized like most automation typologies for considering that automation can only be conceptualized as a growing delegation of tasks to the machine from full human control to full machine control. Shneiderman (2020a) argues that, rather than seeing this as a one-dimensional controllability problem, safe and trustworthy systems should highlight both high human control and high technology control, at least for certain complex and life-critical situations. The SAE typology also presents a narrow view of the 'Dynamic Driving Task' which only includes operational control and a part of tactical control. It does not include strategic control where drivers determine trip goals, route and levels of automation and corresponding functions they may want to change for the trip (Zhang Y et al., 2021). Recent research proposes that the role for the human in a joint system is to allow both human and machine to work as a team. In this team the human is in charge and

Table 1. Automation levels in the autonomous driving literature (after McCall et al., 2019; Hancock 2019).

ADS levels	Execution of steering and Acceleration/Deceleration	Monitoring of driving environment	Fallback performance of dynamic driving task
1. Driver assistance	Human driver and system	Human driver	Human driver
2. Partial automation (feet-off driving)	System	Human driver	Human driver
3. Conditional automation (hands-off driving)	System	System	Human driver
4. High automation (eyes-off driving)	System	System	System
5. Full automation (brains-off driving)	System	System	System

delegates responsibility to the system as a crew member while the human retains command and control. The commander is fully responsible at each level: accepting (strategic), taking (tactical) and coping with (operational) risk, but does not operationally control everything.

Ironies of automation. The literature about ADS highlights the ironies of automation because 'rather than relieving driver workload and vigilance, they can actually place greater demands on the user or they can lead to outcomes that manifest themselves as unintended consequences' (Noy et al. 2018, p. 72). A typical example of this is the difficulty of finding appropriate task allocation strategies (see Table 2).

Moreover, 'the more advanced the automation, the more challenging the role of the driver under critical conditions' (idem). It is ironic that the problem automation intends to solve (e.g. allowing individuals to relax instead of driving) still requires the driver to actively monitor the system (if it is not fully autonomous). To make it fully autonomous would require allowing the system highly developed learning capabilities, which would itself require the driver to fully trust the system.

The ironies of automation we discuss below are rooted in a wider theoretical stream of ironic and paradoxical approaches of technology. 'Technologies of many kinds perform in ways that are ironic, perverse and paradoxical. That is to say, a certain technology applied in a certain way in a certain context may have consequences or implications of one kind but may be implicated in a contrary set of consequences or implications in another context'. (Arnold, 2003, p. 231). The research literature concurs to say that the only viable response to paradoxes and ironies is to accept them and attempt to 'cope' (Mick and Fournier, 1998). Ironies and paradoxes are very close concepts. However, the main difference is that ironies are more rooted in critical approaches reflecting negative and/or problematic outcomes. We summarize some of the ironies of ADS below. Please note that we highlight in italics the terms used in the literature (Bainbridge, 1983; Noy et al., 2018):

1. In terms of *task allocation*, ADS should offer greater value to the human if they are applied to tasks that are too complex for the human or to do well under certain circumstances. Yet, *what is automated are often tasks that human can easily do, not the most difficult ones.*
2. Automation leads to *deskilling*, which in turn, *leads to 'reduced intervention effectiveness when disengagement of automation is requested or necessary*' (Noy et al., 2018, p. 72). Handover from the automated mode to the human operated mode is fraught with difficulties if people have forgotten the required skills (Merat and Lee, 2012). These difficulties are related to potential cognitive overload or lack of understanding due to a learned overreliance on the system.
3. From a *cognition* viewpoint, *the monitoring of a complex ADS requires capabilities and understanding of system operations that might be beyond the requisite diagnostic skills of human operators.* Understanding is a necessary condition for moral responsibility or accountability (Van de Poel, 2011), but it is rarely met in semi or fully autonomous systems. Moreover, users may at times stop monitoring the situation when they should and may experience insufficient Level of Situational Awareness (LSA) or make errors induced by biased representation (mental) models. The National Transportation Safety Board in the USA has said that humans are notoriously inefficient at monitoring automated systems (Banks, Eriksson, O'Donoghue and Stanton, 2018). However, some also consider that awareness should be understood as shared rather than the duty of either driver or system (Hoc, 2000; Hoc and Amalberti, 2007).

Table 2. Task allocation strategies (after Cabrall et al., 2019).

Task allocation strategies	Comments
1. Avoid the role of sustained human supervision of automation	Human attention becomes ineffective at some point if they must concentrate too long
2. Reduce the supervising role along an objective dimension (e.g. duration or envelope of automated operations)	Human supervision may falter if time is excessive or if applied to too many operations
3. Reduce the supervising role along a subjective dimension (e.g. share responsibilities and/or alter the end user experience and impressions)	Human supervision may falter if they have too many responsibilities
4. Support the supervising role from the behaviourism paradigm	Behaviourism paradigm: Conditions the desired target behaviours through training and selection
5. Support the supervising role from the dyadic cognitivism paradigm	Dyadic cognitivism paradigm: Informs designs to support cognitive processes and mental models
6. Support the supervising role from the triadic ecological paradigm	Triadic ecological paradigm: Informs designs to leverage external environment contexts and task considerations

4. From an action *control* viewpoint, the dynamics of ADS are different from those of conventional systems. *If a person lacks recent 'in-the-loop experience' the ADS may not know who is in control at the time.* Both the driver's knowledge about the ADS, and ADS knowledge about the driver (and whether the human driver has sufficient competencies to react) point to uncertain action control in certain situations.
5. *Trust* depends on several factors including predictability, ability, perceived organizational and institutional support among many others. *Trusting ADS when trajectories are unknowable or unpredictable is difficult, notably when systems seem opaque and operate in complex environments* (Burton et al., 2020; Zhang Z et al., 2021). There may be unintended consequences of use.
6. The *liability* gap is also ironic because *conventional law currently considers the human operator liable when an accident happens at all automated levels,* the only exception being full autonomy (McCall et al., 2019). However, this may be unreasonable given the above ironies. It is a case of responsibility without power, making owning (not necessarily operating) the automated equipment rather unattractive from a liability standpoint (McCall et al., 2019). In the case of the fatal Tesla crash in 2016, the National Transportation Safety Board in the USA initially concluded that human error (inattention) was to blame. Subsequently, however, the board revised its decision and criticized Tesla for allowing the autopilot feature to be activated on roads it had not been designed for and for the way in which it determines whether drivers are engaged (Banks, Plant and Stanton, 2018).

The six ironies of automation listed above may manifest differently at *different levels of automation* (Parasuraman et al., 2000) and interconnectedness or autonomy (Talebpour and Mahmassani, 2016). One contribution of this paper is to advance how they manifest at different levels of ADS.

Safety and responsibility

Safety concerns associated with the use of dangerous products, or the functioning of complex socio-technical systems, raise issues about the allocation of responsibility. The Safety Science literature as well as professional and institutional standards (e.g. nuclear industry, aviation industry) point to the necessity to identify responsibilities as clearly as possible to prevent accidents, and/or to handle the consequences, and/or learn from them. A tension exists between the responsibility of the human or the operator of a control room and the responsibility of the whole system (including the interactions between its components). Early approaches in the 1960's and 70's focused on 'human errors' and considered human beings as the 'weak' link responsible for accidents. Over the years, however, attention has shifted to the complexity of the system and the responsibility of the designers, with the notion of a 'normal accident' (Perrow, 1984) or 'organizational accident' (Reason, 1993) becoming accepted. Perrow suggests that the engineers' knowledge, preferences, practices and professional rules for the design of risky technologies (such as nuclear power plants) induce a high level of interactive complexity between tightly coupled components that retrospectively appear impossible to control by the operators and may end up in a major accident. More recently, integrated approaches of safety show the positive contribution of human factors and analyse the organizational processes that produce or reduce safety in complex and unstable environments (Rasmussen, 1997; Lecoze, 2015). 'High Reliability Organizations' (Cantu et al., 2020) safely operate complex and high-risk technologies in extreme

contexts (Hällgren et al., 2018). They succeed in keeping major accidents rare despite the enormous potential for accidents in their respective industries. Responsibility is embedded in a strict definition of occupational roles and is at play every time the obligation to act in accordance with formal procedures competes with the obligation to take the initiative when unexpected events threaten safety (Weick and Roberts, 1993). Therefore, the allocation of responsibility for safety is dynamically distributed in time and space among various actors and organizations (especially between the licensee, the regulator and the public). This evolution in thinking paves the way for a stakeholder approach of responsibility allocation, each one considering safety as a common good and influencing the allocation of responsibility.

The variety of risks (moral, legal, social and psychological) associated with autonomous driving makes responsibility a key issue for the development and adoption of this technology. Key questions include: Who is responsible for what and to whom in the case of an accident? What does responsibility mean in the context of ADS? Such apparently simple questions unveil the complex network of actors potentially involved. It appears that if drivers are considered liable in ADS, adoption will be hampered. This problem has been examined from both legal and ethical viewpoints. Philosophers, ethicists and law scholars distinguish several meanings of responsibility. To say that someone is responsible might mean:

1. To attribute causality (i.e. someone is causally responsible for the outcomes of the use of her system); or
2. To attribute fault or blame (i.e. someone is at fault because of neglect or incompetency); or
3. To attribute liability (i.e. someone should be held legally liable for the outcomes of using the system); or
4. To attribute role assignments (i.e. someone failed to do something required by her role, for example, warn clients of some potential limitation of the system) (Johnson and Mulvey, 1995).

Each of these uses of 'responsibility' is complex, and they are often interdependent. Moral philosopher Van de Poel (2011) distinguishes nine notions of responsibility. The first four are primarily descriptive (as cause, as task, as authority, as capacity), whereas the last five are normative and imply evaluation or prescription (as virtue, as moral obligation, as accountability, as blameworthiness, as liability). Based on Van de Poel (2011), in this paper we focus on only three types of responsibility to integrate the diversity of meanings: 'moral responsibility' (moral obligation and blameworthiness), 'accountability' and 'liability'. Van de Poel (2011) introduces an interesting distinction between 'forward-looking' and 'backward-looking' to characterize the way responsibility operates. He underlines that the first two normative meanings are primarily forward looking, in the sense that someone feels the obligation 'to see to it that something is the case' when it is not yet the case; and that the last three are backward looking 'in the sense that they usually apply to something that has occurred' (p. 40). Since Aristotle responsibility as cause and as capacity has been considered a precondition for holding someone accountable or liable. Responsibility as obligation is generally closely related to the task or authority; however, a task is not necessarily moral. Thus, moral responsibility (obligation or blame) can be either forward looking (i.e. related to obligations attached to the role, for example, to inform passengers about risks) or backward looking (e.g. to accept blame, accountability, or financial liability if monitoring of the semi-autonomous system was poor).

From these fundamental considerations of responsibility and from high reliability theory, two cases of forward vs backward responsibility can be distinguished for moral responsibility and for legal liability. Moral or legal responsibility can be attributed either to the role or status of the person or to how tasks were executed. With forward-looking responsibility, strict liability would suggest that whatever the automation level the user is responsible, since driving a semi-autonomous car in itself implies potential risks and damage (Hevelke and Nida-Rümelin, 2015); similarly, product liability implies that there could be a product design defect that relates to the manufacturer (Abbott, 2018). Alternatively, with backward-looking responsibility neglect liability could be invoked by the law if the user or the designer neglected to perform something that would have enabled the avoidance of risk (Abbott, 2018) (e.g. warning consumers about safety risk level or not complying with regulations (Geistfeld, 2017)). Under neglect theory, legal liability is based not just on causality, but also on blameworthiness which is itself related to the examination of intentions, understanding and power. Such additional conditions make it harder to establish liability than with strict liability (Abbott, 2018).

Accountability, defined as to be open to demands for justification and to be answerable for actions, can be addressed in two different ways (Van de Poel, 2011). In ethics we generally distinguish between deontic positions based on general principles and consequentialists positions based on a calculus (Mingers and Walsham, 2010). In the following, we discuss only the consequentialist position of accountability, because with moral and legal responsibility, we already considered the deontic route. According to the consequentialist position, 'A is accountable if A has the capacity to act responsibly, is causally involved in X and did something wrong'. Understanding accountability in a consequentialist way makes actions by stakeholders' key. As a result, transparency (visibility and accessibility for

those in charge of checking transparency) can be considered as an important dimension if not a good proxy for accountability (Boos et al., 2013).

Trust and predictability

The 'liminal experience of using [autonomous] tools forces us to confront issues such as technology-based trust and ethics' (Zhang Z et al., 2021, p. 18). A liminal experience is characterized by a 'state of emergence marked by ambiguity and multifariousness' (ibid, p. 15). To be used an ADS must be trusted. The perception of trust is itself related to perceptions of dependability or reliability of operations (Kalra and Paddock, 2016; McKnight et al., 2011) and whether the system will respect privacy. The stronger the relational trust (Tax et al., 1998) towards the manufacturer, the weaker the privacy concerns (Smith et al., 1996) for personal driving data. However, while relational trust qualifies trust in an organization, ADS trustworthiness refers mostly to trust in the technology itself (McKnight et al., 2011). Such trustworthiness has three main components: (1) the belief that the technology has the functional attributes to be able to do certain things, (2) the belief that it can provide adequate and responsive help and (3) that it is reliable (McKnight et al., 2011). Trustworthiness in technology is notably explained by situation normality and by structural assurance, that is, the fact that success is likely because the situation is normal and contract guarantees and regulations are in place (McKnight et al., 2011).

If the system is both reliable and responds predictably to situations, it becomes controllable, which seems a crucial factor for trustworthiness and safety in ADS. However, controllability does not mean control by the user. Control can be remotely exercised by a regulating centre which then has control and reduces the agency of the user. In addition, predictability may not be warranted, in the sense that not all conditions can be known in advance. Compared with the famous trolley problem (Bonnefon et al., 2016), which amounts to deciding who should be killed with certainty, the moral responsibility problems facing autonomous driving are characterized by uncertainty (Nyholm et al., 2016). More generally, as ADS are connected to the external environment, they cannot be tested in every situation since the environment continues to evolve; in this sense they are unpredictable.

Ensuring the safety of ADS means that they respond to certain situations in a predictable manner. That they do so under the same normal conditions without causing an accident is key to their trustworthiness (McKnight et al., 2011). However, accidents sometimes happen because normal conditions are not met. Wildlife, pedestrians, or cyclists may occasionally interact with vehicles, and they may not always respect the rules. Weather conditions and infrastructure maintenance can also introduce hazards. Thus, in some abnormal conditions or infrequent situations, ADS may become unpredictable and dangerous (Goodall, 2020). While at high levels of automation, drivers should be able to anticipate that the car will avoid accidents, they cannot reliably predict the trajectories and actions that will be chosen by ADS (Zhang Z et al. 2021) given their self-learning characteristics. Research projects using reinforcement learning (simulation) or supervised learning based on hundreds of human drivers have found that ADS at level 4 of automation seem unlikely to reach safety requirements (Russell, 2019).

It may be that people's trust in ADS will vary depending on the predictability of the systems and the ethical decisions that they might make (ideally known in advance) (Karnouskos, 2021). Such decisions could be personalized by the driver or the owner or made mandatory by public authorities (Gogoll and Müller, 2017). The difficulties of providing complete specifications of ADS are related to 'three root causes: the complexity and unpredictability of the system's operational domain; the complexity and unpredictability of the system itself; and the increasing transfer of decision-making function from human actors to the system. These three issues also affect the safety assurance of autonomous systems'. (Burton et al. 2020, p. 2). From complexity theory we know that the more complex the system, the more it is unpredictable, and particularly in the domain of high-risk technologies (Perrow, 1984). We can use complexity as a proxy for predictability. Typically, regarding the operational domain, autonomous driving in a short segment with no pedestrians crossing will exhibit more predictable outcomes (i.e. less prone to accidents) and may be considered low complexity compared with autonomous driving on a freeway. However, driving on a freeway is low complexity compared with autonomous driving mixed with human-driven cars in a city (Burton et al., 2020). Regarding the system itself, whether systems are rule based or trained with machine/deep learning would be important. Opaque systems based on self-learning algorithms cannot be predictable (Zhang Z et al., 2021). Regarding handovers from human to machine or vice versa, it both depends on the learning capabilities of the system and on the users (if not fully automated). Interestingly, predictability difficulties also create a moral responsibility gap and a liability gap in the sense that if normal conditions are not met, do manufacturers, operators or users deserve moral blame? And if so, should they be liable to pay compensation for those injured by an autonomous system (Burton et al., 2020)?

Methodology

As the issue we wish to study (responsibility and safety in ADS) is by nature interdisciplinary (Koopman and Wagner, 2017), we formed an interdisciplinary team of two senior IS researchers, one with a background in transportation and the

other with publications in digital transformation and ethics, one senior researcher in organizational studies specializing in safety in nuclear power plants, and two PhD students, one specializing in HCI with a strong background in bibliometrics, and one in human resources and IS. We used a variant of a method that the first author and one of the PhD students specializing in bibliometrics have used previously to conduct a literature review. This method named BIBGT interlaces bibliometrics with a grounded theory approach (Walsh and Rowe, 2023). We used (a) a bibliographic technique called Documents Bibliographic Coupling Analysis (DBCA) as proposed by Walsh and Renaud (2017) and (b) a systematic assessment of a subset of documents derived from DBCA that uses our critical analytical framework as a lens. To further classify the literature, we located them on the socio-technical continuum and identified the stakeholders they consider. Our approach thus combines the inductive approach of BIBGT with a final deductive reasoning touch. An interpretive literature review coupled with bibliometrics can be illustrated by the flesh and bones metaphor, whereby researchers' interpretation of documents (the flesh) is added to the bibliometric analysis (the bones) to reveal the structure of a field. DBCA can help identify current themes/trends of a field[2] (Zupic and Čater, 2015) and thus contribute to scoping reviews (Rowe et al., 2023). A high value in bibliographic coupling strength indicates a similar subject relationship between two documents.

Bibliometric techniques introduce some objectivity into the classification of the publications of a research field and are valuable to investigate a subfield that has been studied from the perspective of different disciplines (Walsh and Renaud, 2017). Since Walsh and Renaud's (2017) publication, major improvements have been made to automate the use of bibliometrics techniques and facilitate BIBGT. We used ARTIREV software to conduct the entire bibliometric workflow (i.e. to collect bibliometric data, calculate and normalize co-occurrences matrices, cluster documents and generate science mappings). ARTIREV automates the procedure while also enabling researchers to make easy choices about thresholds, normalization methods and clustering algorithms (Walsh et al., 2022).

Details of the BIBGT Iterative Process

First, we explored the ethics of and responsibility in various[3] ACPS to become familiar with the general literature and develop our 'theoretical sensitivity' (Glaser, 1978). Second, we narrowed our scope to ADS and designed our critical conceptual framework for understanding the ironies of responsibility in ADS. We used this framework to undertake a review for understanding (Rowe, 2014) what autonomous driving means and how responsibility can be allocated at different automation levels. This review can be considered a hybrid between a scoping review through our use of BIBGT (focused on identifying gaps on the research front with DBCA) and a critical review through our ironies-based framework.

The first step of BIBGT consists of defining the boundaries of the review. Using Scopus as a data source, we extracted 217 documents –and their bibliographic references – focused on autonomous driving systems/vehicles that were published during the period 2016–2022.[4] Our query, shown in Appendix A, includes the most common terms denoting ADS (Gandia et al., 2018).

The second step of BIBGT involves cleaning the bibliographic data, choosing a citation threshold, and using normalization techniques to reduce the dataset. Because scientific data sources such as Scopus contain significant data quality problems (Van Eck and Waltman, 2017), we used fuzzy string similarity algorithms provided by ARTIREV to help merge similar references. Hence, we curated in a semi-automatic way the 10,192 single references included in these documents. We finally obtained 7491 single references that corresponds to 27% of curated references. We also performed a manual verification of the documents selected for our analysis.

Regarding the selection of the second order samples, we iteratively and theoretically sampled the literature to obtain the most suitable thresholds and resulting samples. We used the normalized citation count (NCC) and the citation count (CC) as thresholds. Because the citation count increases over time, the NCC prevents the disadvantaging of recent literature. When the NCC of a document is more than 1, it means that the publication received more attention in terms of citations than others published the same year. Our CC threshold allows us to exclude recent publications cited once (i.e. typically those published in 2022) that are not filtered by the NCC threshold. After several trials, we retained three analyses for comparison: the entire set (n = 217), an intermediate set (subset A, NCC >0.5 and CC > 1, n = 76 after a manual verification), and a restrictive set (subset B, NCC >1 and CC > 1, n = 44 after a manual verification) (see Figure 2). A manual screening was carried out to retain only the most relevant publications for our objective (see BIBGT step 3).

In the third step of BIBGT – Clustering/mapping/interpreting –and for each analysis, we first calculated a co-occurrence matrix in which we applied the association strength normalization method. Then, we applied the Leiden clustering algorithm and the ARTIREV mapping on the three matrices produced. As recommended, we produced science mappings at different citation thresholds and interpreted them (Walsh and Rowe, 2023). We started with the entire literature and coarse grain clusters. We noticed three primary topics (governance, Responsibility Sensitive Safety (RSS) and human factors), but when analysed in detail considerable noise was present.

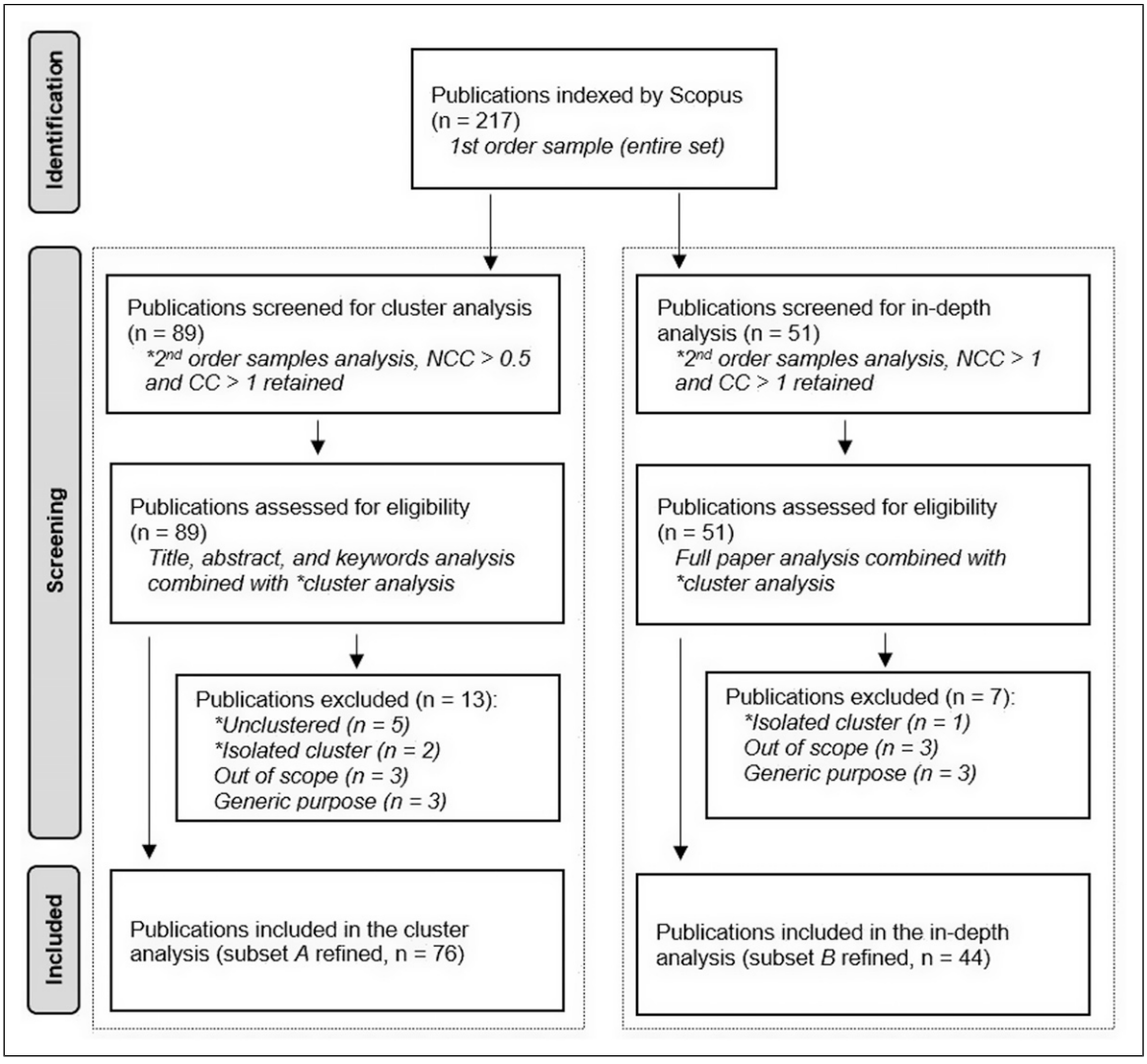

Figure 2. Screening and Selection Process. The broader sample (subset A, on the left) has lower citation thresholds. Hence, it includes documents of the more restrictive bibliometric analysis (subset B, on the right). The (*) denotes that the processing has been supported by ARTIREV.

Subsequently, we performed an intermediate analysis (subset *A*) and pursued a more restrictive set (subset *B*). This smaller set helped us to clearly delimit the clusters and their relations. Subsets *A* and *B* were retained because the documents faithfully represent the subject of this study, and their respective mappings were the most clearly interpretable. We refined both subsets by removing documents not aligned with our study objective, were unclustered, or isolated (list available upon request). Regarding review papers, we assigned them to the closest cluster. If no cluster was affected, we discarded the paper to gain clarity (e.g. Gandia et al., 2018).

We highlight that our interpretation has been performed for the three analyses. However, the findings section is focused on our interpretation of the subsets *A* and *B*, while the more limited set was analysed in-depth. Our interpretation was supported by qualitative coding of the title, abstracts and keywords of all documents. This also led to the identification of the core category, that is, responsibility under uncertainty in autonomous driving.

In the fourth and last step of BIBGT, using our conceptual framework and the perspectives taken, we coded the 44 selected papers of the subset *B* (see Appendix B for methodological details). We synthesized this literature and discovered some research gaps. For the latter, to ensure convergence of the coding, each paper was coded by pairs of coders. We formed duos of one senior researcher and another researcher of the team to code subsets of papers separately. The assignment of papers was performed with regards to domain specificity. Once a

subset was coded, the two researchers compared their coding, discussed their differences and calculated the a priori convergence rate. Then, the entire team conferred to discuss divergences in the coding. Out of 616 codes to be compared, the convergence rate grew from 64% before discussion to 100% afterwards. To present the results of the coding, we ranked the papers, first, by cluster (see Appendix C). Subsequently we classified them according to whether they addressed transparency. The articles that refer to it are ranked before those that do not. Indeed, we think that transparency is one of the central ethical values from which responsibility and trust can be discussed.

Perspectives: Types of approaches and stakeholders

Finally, following both Noy et al. (2018) and Sarker et al. (2019, p. 712) we coded the problem of safety and responsibility in autonomous driving as 'consisting of both social and technical aspects and locating them on a social-technical continuum'. To that end, Sarker et al. (2019) suggest engaging in literature reviews to help identify research gaps with respect to 'types' of socio-technical research. They define these types as follows (ibid, p. 712):

1. Type 1 'predominantly social in a technological context', 'where the investigation focuses […] on the social […] aspects related to the phenomenon of interest, with technological or informational considerations serving as the context'.
2. Type 2 "social imperative "treats technology as the product of human choice and organizational processes.
3. Type 3 'the social and technical as additive antecedents to outcomes… Both the social component and the technical component are seen as separate antecedents to certain outcomes; [with….] no evidence of any interaction between the components themselves in producing these outcomes'.
4. Type 4 'the social and technical interplay to produce outcomes' closest to the socio-technical perspective represented by structuration theories, fit or misfit theories, the socio-material perspective and value sensitive designs.
5. Type 5 'the technical imperative', a soft form of technological determinism where the properties [..] of technology influences socio-economic outcomes.
6. Type 6 'predominantly technical' aims at advancing problem solving capabilities of technology, typically through IS design research.

These types of approach are high level abstractions about how social and technical aspects interact. Various stakeholders play an essential role in the acceptance, governance and implementation of ADS.

Findings

Mapping the current research about safety and responsibility in autonomous driving systems

Figure 3 shows a cluster analysis based on the DBCA. Each of the six node documents represents the most central work for that cluster. For each cluster, we present only the most important works, based on their number of citations. We describe them clockwise starting from the top of Figure 3.

Cluster 1: Adoption of autonomous driving. Cluster 1, coloured brown in Figure 3, contains 12 publications concerned with the adoption of ADS. Factors influencing ADS adoption can be classified into four areas: technology infrastructure (communication, technology of roads and traffic signs and cost of infrastructure), legal (liability, privacy and cybersecurity), ethical principles and regulations, and user behaviour influence factors (marketing and advertising, cost and trust) (Alawadhi et al., 2020). Shabanpour et al. (2018) showed that people are much more sensitive to the purchase price and incentive policies (such as taking liability away from the driver in case of accidents), as well as provision of exclusive lanes for ADS, compared with other factors such as fuel efficiency, safety, or environmental friendliness. Wang and Zhao (2019) analysed the relationship between individualized risk preference (e.g. economic, psychometric) and ADS adoption, and showed that risk preference parameters are significantly associated with socio-economic variables. Finally, Wu et al. (2020) demonstrated that consumers display a positive attitude toward autonomous, connected, electric vehicles. They can bring widespread benefits, including reducing driver fatigue, environmental friendliness and increased accessibility of travel for non-drivers. On the other hand, participants had concerns about vehicle safety and legal liability. Finally, ethical preferences built into systems influence the adoption of ADS (Karnouskos, 2021).

Cluster 2: Governance of autonomous driving. Cluster 2, coloured red in Figure 3, contains 23 publications that are mainly concerned with governance of ADS. The high number of documents and the high number of mean citation counts suggests that governance of ADS is a hot topic (see Table 4). Many scholars have explored the legal challenges related to highly automated driving systems compared with those related to vehicles of lower automation (Leiman, 2020). So far, governments have not constrained ADS developments but have explored ADS implications (De Bruin, 2016; Taeihagh and Lim, 2019). Numerous initiatives have been undertaken to resolve blame attribution.

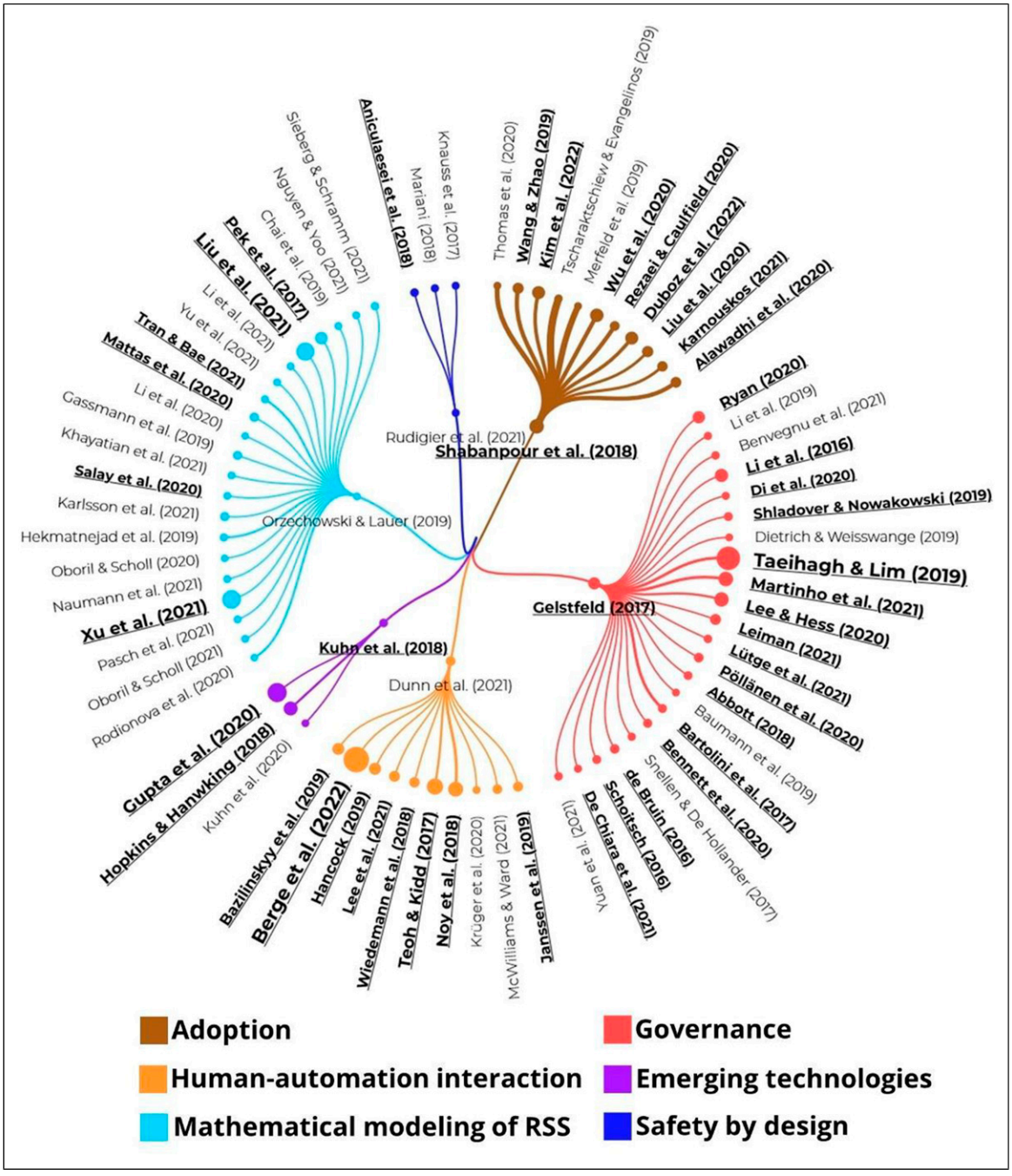

Figure 3. Map presenting current research themes on safety and responsibility for autonomous driving systems. The subset A (n=76) is presented and corresponds to documents in the cluster analysis. The subset B (n=44), more restrictive (in bold and underlined), has been analysed in-depth. Central nodes denote documents with the most important connectedness in their respective cluster. The node size denotes the NCC of documents. (b) Cluster 1: Adoption of autonomous driving.

These include changing the paradigm of tort liability (Abbott, 2018); adding new federal safety regulations to established tort doctrines (Geistfeld, 2017); investigating public opinion regarding the attribution of blame in crashes (Bennett et al., 2020; Pollanen et al., 2020); and providing an analytical tool based on a game-theoretic model to assist policymakers (Di et al., 2020).

Citizens tend to assign the obligation of setting moral norms for self-driving cars to ethics researchers and to car manufacturers (Li et al., 2016). Hence, both scholars and

practitioners are engaged on the topic. For instance, an initiative joining both industry experts and researchers developed practical guidelines for ethical principles related to ADS (Lütge et al., 2021). Martinho et al. (2021) performed a review of the ethics of autonomous technology in the scientific literature and in industry reports published by companies testing ADS in California. Discussions about ethics related to ADS tend to focus on the trolley problem, while practitioners do not address this problem in industry reports. Both scholars and practitioners prioritize safety and cybersecurity and agree that ADS will not eliminate the risk of accidents.

Cluster 3: Human-automation interaction in a safety context. Cluster 3, coloured orange in Figure 3, comprises 11 publications concerned with human-automation interaction in a safety context. Both sides can interact within the vehicle or in traffic. The safety benefits of ADS are mostly emphasized in this cluster (Noy et al., 2018; Teoh and Kidd, 2017). Highly autonomous driving is safer than human drivers in certain conditions but will continue to be involved in crashes with human-driven vehicles (Teoh and Kidd, 2017). Noy et al. (2018) argue that regardless of the level of automation, a driver will continue to have a role. More critically, Hancock (2019) discusses issues surrounding driverless vehicles with a human factors/ergonomic perspective. Other publications investigate the interactions between humans and various levels of automated vehicles under particular conditions, such as the effect of different alcohol levels on handover performance in conditional ADS (Wiedemann et al. 2018), and interruption and interleaving processes (Janssen et al., 2019). Amongst a variety of external human-machine interfaces, those sending textual egocentric messages from the viewpoint of a pedestrian (rather than that of the ADS) are regarded as the clearest, which poses a dilemma because textual instructions are associated with practical issues of liability, legibility and technical feasibility (Bazilinskyy et al., 2019). More recently, to enhance the safety of cyclists interacting with ADS, Berge et al. (2022) explored on-bike human-machine interfaces. They found that cyclists are hesitant about such interfaces because the utility value is unclear, and the responsibility of safety should not be imposed on the more vulnerable road user.

Cluster 4: Emerging technologies to guaranty traceability and accountability. Cluster 4, coloured violet in Figure 3, contains four publications concerned with emerging technologies helping to guarantee safety, traceability, liability and accountability in manufacturing processes and the resulting products. These publications focus on the cause of the problem and/or responsibility attribution. These articles identify the characteristics that safety assurance should exhibit regarding different stakeholders. Blockchain (Kuhn et al., 2018; Gupta et al., 2020), IoT (Kuhn et al., 2018), and big data (Hopkins and Hawking, 2018), alone or in combination, are presented as promising solutions to address the above-mentioned issues. Blockchain technology is mentioned as a valuable technology to record, share and trace specific data in the supply chain and consequently prevent problems by tracing accountability. Kuhn et al. (2018) investigated the electrical supply industry (producer of electrical components that transmit the vehicle's energy and communication flow) to derive current challenges as well as future requirements for production processes of safety critical products in the age of ADS.

Cluster 5: Mathematical modelling of Responsibility Sensitive Safety. Cluster 5, coloured light blue, contains 22 publications with the lowest mean citation count. Most of these publications are concerned with RSS. These publications are focused on mathematical modelling of RSS (Salay et al., 2020), formalizing traffic rules to solve liabilities of traffic participants if a collision occurs (Pek et al., 2017), and on simulations that increasingly benefit from naturalistic data (Pek et al., 2017; Xu et al., 2021).

Cluster 6: Safety by design. Cluster 6, coloured blue in Figure 3, contains four publications concerned with safety by design (i.e. requirements, framework, design and resulting infrastructure). These publications introduce a holistic and iterative software engineering approach to develop dependable autonomous systems (Aniculaesei et al., 2018). Mariani et al. (2018) describe the status of the ISO 26,262 functional safety standard with a specific focus on its application to semiconductors. Testing the functionality and safety of automated vehicles are also investigated (Knauss et al., 2017).

Synthesis and gap analysis: Recommendations for interdisciplinary research

Taking our conceptual framework as a guide, we now present some gaps and recommendations. Many articles are published in research fields other than IS, such as transport, safety, justice, accident analysis and prevention. Hence, responsibility in autonomous systems is an interdisciplinary subject for many social sciences and socio-technical disciplines. Our findings in this section are relevant to ADS researchers in all these disciplines including IS. We first highlight the frequencies of the codes found in Table 3.

Types of approaches and types of stakeholders considered. All types of approaches mentioned by Sarker et al. (2019) have been used for studying safety and responsibility in ADS. This diversity is particularly reflected in the governance

Table 4. The IS Type continuum by cluster. MCC: Mean Citation Count; TDC: Total number of documents within cluster; TDA: Total number of documents analysed in-depth. T1 to T6 denote the six system types according to Sarker et al. (2019).

#	Cluster	MCC	TDC	TDA	T1	T2	T3	T4	T5	T6
1	Adoption	25.8	12	9		2	7			
2	Governance	25.4	23	17	2	6	1	8		
3	Human-automation interaction	40.0	11	8			2	6		
4	Emerging technologies	49.5	4	3					2	1
5	Math. Mod. RSS	11.4	22	6						6
6	Safety by design	13.5	4	1						1

cluster (see Table 4). Two papers fit into Type 1, where the social aspect is predominant. The psychological aspect is critical in the study of Bennett et al. (2020) where the respondents' representations of the attribution of blame and ultimately of responsibility in the event of an accident with an autonomous car are investigated. In our sample, the eight Type 2 papers mostly concern the development of a legal framework for ADS (Abbott, 2018; Geistfeld, 2017) and reflect upon improving social/legal norms (e.g. in an accident involving autonomous cars, what is the effective liability rule for allocating losses among road users so that the total social cost is minimized? (Di et al., 2020)). Type 3 is also well represented in our sample with 11 articles. For example, Karnouskos (2021) studied the role of technology (technical factor), and the ethical preference for self-safety vs. utilitarianism (social factor) on self-driving car acceptance.

As we found 29% of papers to be Type 4, we refined this type into subtypes where authors position them as follows:

1. HCI: human-computer interaction models in general or specific situations of automated use and respective agencies (6 papers, for example, Janssen et al. (2019) present a 10-step model to explain the control transition between the autonomous car and the human operator. His attention is divided between driving and non-driving tasks (e.g. watching a movie on the mobile phone).
2. FIT: multidimensional adaptation between characteristics of autonomous systems and societal constraints and demands is challenging, yet possible (4 papers; it is typically used by policy perspectives from different countries providing best practices such as driver training programs for safety (Lee and Hess, 2020).
3. SYT: socio-technical systems theory with interaction feedback effects between social and automated components at different hierarchical levels (2 papers: (Noy et al., 2018) propose three levels: (1) the Socio-technical Environment reflecting the domestic and the international legal, political and economic (e.g. trade) milieu that influences the approach to autonomous driving. (2) 'Transportation System Planning' articulating mobility objectives, norms, regulations, public health priorities and road culture. (3) Automated driving as a human-centric cyber-physical system where not only drivers' interactions with the automated vehicle, but interactions between vehicles and other stakeholders influence each other. Taeihagh and Lim (2019) focus on interactions at this third level.
4. VSD: value sensitive design of systems involving respect of ethical principles (Winkler and Spiekermann, 2021). Here, the Lütge et al.'s (2021) paper stresses the importance for both policy and industry to respect certain ethical principles: human agency depending on the level of automation, safety and resilience of the system to attacks (for instance, hijacking), or explicit consent of the driver required for the collection of certain personal data which could be used for marketing purposes or shared with third parties.

Our sample also includes papers with a very technical dimension: type 5 and 6. Type 5 treats the technology as a structural change in an organization, such as Hopkins and Hawking (2018) who explain the role of Big Data and IoT to improve driver safety in a logistics firm. Finally, Type 6 focuses mainly on how to develop or improve the technology (Sarker et al., 2019) (e.g. Salay et al., (2020) propose RSS models).

Recommendation 1 (R1): develop systems theory, value sensitive design and socio-material approaches on ADS

The selected papers show an aggregation of actors, risks and safety issues associated with ADS. Surprisingly, however, few focus on the complex interactions between manufacturers, users and regulatory agencies. The dominant point of view refers mainly to the manufacturers (20 occurrences) of ADS. This is followed by the perspective of the human driver and regulatory agencies (respectively 17 and 15 occurrences). They are clearly identified as key actors in the adoption/rejection of ADS. Each stakeholder is

Table 3. Codes count of the 44 papers for levels of analysis, capabilities and outcomes (see codes in Appendix B).

Category	Codes	Frequency
Clusters	1 (adoption)	9
	2 (governance)	17
	3 (human-automation interaction)	8
	4 (emerging technologies)	3
	5 (mathematical modelling of RSS)	6
	6 (safety by design)	1
Type of approach	T1 (predominantly social)	2
	T2 (social imperative)	8
	T3 (social and technical as separate antecedents)	11
	T4 (social and technical as producing outcomes through their interplay)	13
	VSD (value sensitive design)	1
	FIT ((mis)fit theory)	4
	HCI	6
	SYT (systems theory)	2
	T5 (technical imperative)	2
	T6 (predominantly technical)	8
Stakeholders perspective	A (academics)	3
	AV (autonomous vehicle)	8
	CIT (citizens)	12
	D (driver)	17
	I (insurer)	2
	M (manufacturer)	20
	O (owner)	8
	RA (regulatory agency)	15
	TO (traffic operator)	2
Task allocation	D (driver)	1
	D-4 (behaviourism support)	1
	DSY (driver and system)	6
	DSY-1 (avoiding human supervision)	1
	DSY-2 (reducing human supervision on an objective dimension)	1
	DSY-3 (reducing human supervision on a subjective dimension)	5
	DSY-4 (behaviourism support)	1
	DSY-5 (cognitivism support)	2
	SY (system)	1
	SY-1 (avoiding human supervision)	6
	SY-3 (reducing human supervision on a subjective dimension)	1
	NA (not applicable)	18
Deskilling	1 (scheduled handover)	1
	2 (non-scheduled system-initiated handover)	4
	3 (non-scheduled driver-initiated handover)	5
	4 (non-scheduled driver-initiated emergency handover)	4
	5 (non-scheduled system-initiated emergency handover)	6
	C (concerned)	3
	NA	33
Cognition	LSA (lack of situational awareness)	15
	O (overload)	2
	RE (representation based error)	11
	NA	24
Automation level	1 (driver assistance)	0
	2 (partial automation)	1
	3 (conditional automation)	16
	4 (high automation)	15
	5 (full automation)	24
	Any	16

(continued)

Table 3. (continued)

Category	Codes	Frequency
Risk	C (concerned)	21
	COL (collision)	13
	D (death)	14
	I (physical injury)	14
	DI (physical discomfort)	3
	M (mental (psychological) problems)	1
Moral responsibility	C (concerned)	6
	N (neglect)	14
	S (strict)	10
	NA	21
Accountability	T (transparency)	24
	NA	20
Legal liability	L (liability considered)	11
	N (neglect)	16
	SP (strict product)	21
	NA	9
Trust	T (trust)	11
	R (reliability)	7
	Both	11
	NA	15
Domain predictability	C (concerned)	3
	HC (high complexity)	22
	MC (medium complexity)	3
	LC (low complexity)	0
	NA	16
Software predictability	C (concerned)	25
	C* (paper particularly interesting)	1
	NA	18

viewed as a risk producer for others as well as a potential victim of the behaviour and decisions of others. Although some articles address more than one stakeholder, most papers focus on the perspective of a single stakeholder.

This leads us to the identification of a second gap in the literature: the non-political treatment of the stakeholders. Almost nothing is said about the lobbying and negotiating activities of various stakeholders. There is no systematic analysis of the political and economic interests of the actors. This is problematic given that some papers can be interpreted as support provided to the interests of a particular stakeholder. This is particularly the case for liability issues faced by the manufacturers (Schoitsch, 2016). The safety lens seems to orient the study of ADS in a technical and legal direction rather than a political one.

Recommendation 2 (R2): Develop a political focus on stakeholders' interests and interactions.

Capabilities. Task allocation is mentioned in 59% of the articles in our sample. The key questions are about the complexity of handover and the related choices made by the manufacturers and the habits and preferences of the users that pose new safety risks (Baumann et al., 2019). For individuals, giving up control completely and trusting the car is an obstacle that has yet to be overcome (Bruckes et al., 2019). A solution often mentioned in the literature is for the driver to have a button to regain control. The admissibility of such a function depends on the level of automation of the car as well as on the state and behaviour of the driver (Lütge et al., 2021). Thus, controllability can be viewed as increasingly allocated to the machine as automation levels increase. The most frequently mentioned locus of control in our sample is shared between driver and system. Shneiderman's (2020a) position is that both should have high control with humans having authority and responsibility. Control should be coordinated between humans and systems (Baird and Maruping, 2021), while responsibility should be either that of the driver or of the ADS designer. However, and surprisingly, rather than focusing on the human-machine cooperation suggested by the cognitivist paradigm (strategy 5) and beyond as in the ecological paradigm (strategy 6), the most common strategies mentioned by the authors are strategies 1 (avoiding the role of sustained human supervision of automation) and 3 (reducing human supervision on a subjective dimension). This may be due to the gap between the existing conceptual

cognitivist framework in Ergonomics (Hoc, 2000) and IS (Biondi et al. 2019) and the lack of empirical research.

Recommendation 3 (R3): Investigate options for sharing control with digital agents beyond a unidimensional problem.

Deskilling is rarely addressed in the articles, with three articles being concerned and only eight mentioning it precisely (considering equally whether unplanned handover is initiated by the driver or by the system, see Appendix B). If humans are no longer required to steer the vehicle, human drivers are not a good backup in case of an imminent accident, at least when they are busy with other tasks (Baumann et al., 2019). Three clear reasons why the deskilling of humans seems to be inevitable are as follows (Janssen et al., 2019). First, research shows that in early stages of the transfer from the system to the human (i.e. stages 2 to 5 (ibid), drivers might not immediately direct their attention to driving and they might not have sufficient awareness of their environment to act appropriately. Second, studies have shown that earlier tasks might negatively impact later tasks (such as driving) even if those earlier tasks were discontinued. Third, there is no empirical evidence that human drivers completely and systematically disengage from other activities when they take control of the vehicle, whereas there is evidence that they engage in other non-driving tasks under conventional driving conditions.

The literature fails to address the double ironies of task allocation and deskilling (Noy et al., 2018) generated by the opposite tendencies between the extra load of complex activities assigned to the human user (due to the limitations of the ADS when operating under critical circumstances) and the deskilling tendency resulting from the absence of regular practice of driving. Such a combined irony has significant consequences for safety and liability issues. On the one hand, the risk of an accident becomes more likely in critical situations, exposing users' lives, because by design the deskilled driver is supposed to take control of the situation under certain critical circumstances for which she has no training. This potentially leads to failure. In a sense, this state of affairs mechanically 'produces' human errors (that feeds the statistical assumption that 90% of the crashes are caused by the human driver). On the other hand, such an irony could expose the human driver to liability (in most countries humans are still considered as the ones responsible for safety (Bennett et al., 2020; McCall et al., 2019)). Finally, the human appears to be potentially trapped in the manufacturer's design choices.

The cognitive irony identified by Noy et al. (2018) is mainly addressed in the literature through LSA (Lack of Situation Awareness). However, it is mainly treated from a technical perspective and less from an ironical one: little is said about the LSA created by the ADS design options that reinforces inattentive, sleepy and distracted drivers. We suggest a critical approach should also consider a meta-LSA issue: if the driver is not aware of the ironies identified by Noy et al. (2018), they are more likely to engage in inappropriate behaviours, putting safety at risk. This is related to the stakeholder political approach mentioned above.

Recommendation 4 (R4): Investigate deskilling and the associated ironies to improve driver awareness.

We did not encounter papers analysing the action control irony in our sample, in the sense of either the anticipation of human cognition by ADS or controllability of ADS actions by the human, which is why we did not code this irony. Behaviourism support appears in our sample but not specifically on this aspect. Hence,

Recommendation 5 (R5): Address ADS cognition anticipation and ADS dynamics controllability.

Safety risks and responsibility. The risks considered in the articles focus mainly on collision and on the physical integrity of the user: injuries, road accidents and death. One of the fundamental arguments for the adoption of autonomous driving is that it will cause far fewer accidents. But proving the safety of ADS may take a long time (Kalra and Paddock, 2016; Shladover and Nowakowski, 2019). RSS papers mostly use mathematical models and simulations. Ironically, these papers focus on the prevention of accidents but there is no explicit discussion about the nature of the 'agent' who could be responsible for safety (i.e. human drivers, hybrid automata, multi-agent systems, the car itself) and no discussion about the assignment of that responsibility.

Moral responsibility is mentioned in just over half of the sample, whereas legal liability is addressed in nearly 80% of the sample. Both are almost evenly split between 'neglect' and 'strict' conceptions of moral responsibility and legal liability. It appears that a clearly agreed doctrine for responsibility (moral and legal) is lacking and is still under elaboration. The public tends to attribute responsibility to the human for an accident, regardless of the level of automation of the car, and thus the human can be legally prosecuted (Bennett et al., 2020). However, liability is sometimes attributed to manufacturers based on strict product liability. Merfeld et al., (2019) say that *'scholars have reached the shared conclusion that elimination of a human driver will shift responsibility onto manufacturers as a matter of products liability law'* (p. 1619). This is clearly the case at level 5. But such a morally justified evolution can also be considered as counterproductive for the development of ADS because it increases the liability risk taken by the manufacturers. At levels 3 and 4 the law is still favourable to them (McCall et al., 2019). However, regarding moral and legal responsibility, designing systems and interfaces for a commander role (Zhang Y et al., 2021) paves the way for allocating responsibility to the human from level 2 to level 4, unless strict product liability applies. In fact, such designs would give the driver the ability and authority

to assume controllability and thus proper conditions would exist (notably regarding endowments, preferences and roles) for effective delegation mechanisms (Baird and Maruping 2021). If, on the other hand, the driver is mostly considered as a fallback mechanism and since responsibility without authority leads to stress, responsibility may be allocated to the manufacturer, or at least shared with it. In short, if an algorithm '*is designed to preclude individuals from taking responsibility within a decision, then the designer [*manufacturer*] of the algorithm should be held accountable for the ethical implications of the algorithm use*' (Martin, 2019, p. 835). Accountability is regularly addressed in the articles and considered as an ethical issue: transparency is required for autonomous cars. Algorithms should not be designed as non-transparent black boxes (Lütge et al., 2021).

Recommendation 6 (R6): Contribute to the elaboration of a universal doctrine for responsibility at levels 3 and 4

Trust and Predictability. In 41% of the articles, the notion of trust or reliability appears, 61% of which refer to both. Perceptions related to ADS vary between countries depending on the culture, the diversity of backgrounds, technological awareness and the social interactions of people (Cho and Jung, 2018). Whether keeping to limits (e.g. speed), or enjoying comfort (e.g. temperature), customizing system parameters is also important in technology adoption (e.g. Alawadhi et al., 2020). There are strong positive effects of perceived technical protection of trust (Bruckes et al., 2019). Among older individuals, situational normality significantly influences trust towards autonomous cars, while this effect is not significant for younger participants.

Finally, predictability of the (driving) environment is low: it is defined by its inherent complexity (e.g. changing weather conditions, unpredictable behaviour of some road users, etc.). Added to this is the complexity of the system itself. Predictability of the ADS is a concern in 59% of our papers. As ADS use environmental signals captured by sensors, the relationship between inputs (signals) and trajectories (outputs) is unknowable (Zhang Z et al., 2021) for users, meaning that moral and legal responsibility of the driver at level 3 or four cannot be determined. For the moment, it is not certain that future users will fully understand how autonomous cars will work, although this seems an important condition for the adoption of this technology. User perception of safety will play an important role in ADS adoption (Alawadhi et al., 2020). Because the lack of predictability in complex environments threatens trust, a solution will have to be found by manufacturers and system designers to make the introduction of autonomous cars a success.

Recommendation 7 (R7): Investigate how trust in ADS can be gained or lost

A research agenda for information systems

As ADS are a new frontier for IS research (Ketter et al., 2022; Lyytinen et al., 2022), in this section we suggest a research agenda for the IS field. We believe the five areas we suggest for future research below might enable IS researchers to shed light on, if not resolve, some of the ironies we have identified. For instance, safety experts recommend that when control of a task is delegated to a machine, legal responsibility should be allocated to the manufacturer (Shneiderman, 2020b; Lee and Hess, 2020). Delegation to a machine creates a need for accountability and transparency, topics that have been discussed within IS along with privacy issues (Cichy et al., 2021). Like other disciplines, IS has engaged in research on the adoption of ADS and related HCI.[5] However, with few exceptions (Bruckes et al., 2019; Bornholt and Heidt, 2019) IS papers do not cover *both* safety and responsibility issues. Even when they do, their treatment misses the ironies. Hence, we suggest a set of research themes that IS researchers could study. These research themes build on existing research themes in IS as well as our findings from the literature on ADS in other fields. Whereas existing IS research tends to have a narrow focus on the technology itself, future IS research on autonomous systems needs a broader focus. Table 5 lists the recommendations we provided earlier along with our future suggested research themes.

Socio-technical perspective on ADS

In their review of the IS research literature, Sarker et al. (2019) found that IS research has largely neglected the socio-technical perspective over the past 20 years. Most studies (91%) have focused exclusively on instrumental goals, such as efficiency and effectiveness. They suggest that the socio-technical perspective can serve as a distinctive and coherent foundation for the IS discipline. We agree. The ADS literature tends to have an instrumental focus on technology without considering the wider implications. In our review of the literature on ADS, we found that most articles neglect the interaction between the various parts of a system –they only take an HCI or a fit approach in a minority of cases. Since the delegation of tasks is central to several ironies, IS researchers could use the framework of delegation to and from the IS agentic artifact (Abbass, 2019; Baird and Maruping, 2021). Coordination between the latter and other stakeholders can be interpreted as a delegation mechanism problem under uncertainty, that is, where trust in technology depends on limited human knowledgeability of the ADS endowments (ability, preferences, roles) and where the appraisal by the IS agentic artifact (ADS) of the respective human driver endowments is limited. Thus, coordination to and from the ADS inevitably raises the cognitive irony (human knowledgeability of the ADS

Table 5. Future research themes in IS

Recommendations	Future research themes in IS
R1 develop systems theory, value sensitive design and socio-material approaches on ADS	Socio-technical perspective on ADS
R2 develop a political focus on stakeholders' interests and interactions	Critical research studies on ADS
R3 investigate options for sharing control with digital agents beyond a unidimensional problem	Safety and risks of ADS
R4 investigate deskilling and the associated ironies to improve driver awareness	
R5 address ADS cognition anticipation and ADS dynamics controllability	
R6 contribute to the elaboration of a universal doctrine for responsibility at levels 3 and 4	Responsibility for ADS
R7 investigate how trust in ADS can be gained or lost	Trust in ADS

endowments), the control irony (appraisal by the IS agentic artifact (ADS) of the respective human driver endowments), and the trust irony (unpredictability in certain circumstances leading to low willingness to adopt ADS).

The three delegation mechanisms put forth by Baird and Maruping (2021) –distribution of responsibility, coordination and appraisal – are interdependent in autonomous driving. In fact, coordination depends on appraisal, and responsibility depends on coordination.

Critical research studies on ADS

One antidote to the almost exclusive focus on the instrumental use of technology would be critical research studies on ADS. If we take a critical perspective, issues such as the political interests of actors come to the fore (Myers and Klein, 2011). Considering the tensions, ironies and discourse in which we are enframed (Arnold, 2003), critical studies could focus on wider societal issues such as lobbying of politicians by companies promoting ADS and the cultural changes associated with automation.

Safety and risks of ADS

IS has a long history of studying the adoption and implementation of technology, but few IS scholars have focused on safety. For example, in a recent article by two leading IS scholars that summarized a 6-year research program that identified principles 'leading to successful intelligent automation programs', the word 'safety' is not mentioned once (Lacity and Willcocks, 2021). Yet in many industries (e.g. airline industry and nuclear industry), safety is the most important aspect. As autonomous systems increasingly take over tasks that used to require substantial human experience (Koester and Salge, 2020), and as our lives start to depend on them, it is obvious that the risks to human safety will increase. If some autonomous systems prove to be unsafe, the business value of such systems will be negligible in any case. How to design interfaces and roles in a joint optimized way (Sarker et al., 2019; Biondi et al., 2019) so that semi-autonomous systems reflect a joint and integrated system is an important issue 'rather than trying to perfect the machine to meet the human's safety expectations [through technology]' (Zhang Y et al., 2021, p. 7). Future IS research could look at whether safety risks come from artificial intelligence (Burton et al., 2020), from connectivity (lack of or too much), or from socio-technical interactions. Notably, there is a need to avoid 'death by GPS' related to the unintended effect of the loss of skill of a trip commander because the navigation is not united with 'other primary driving tasks and controlled by the driver in a commander-like role' (Zhang Y et al., 2021, p. 5).

Responsibility for ADS

As discussed above, ADS raises major ironies about responsibility because its mixes (1) a high level of risks, calling for a clear and simple (but possibly unfair) identification of 'who' is fully in charge and responsible for safety, with (2) a high level of socio-technical complexity. This complexity involves the distribution of responsibilities between several agents, companies and institutions operating in different spaces and times (designers, developers, manufacturers, regulators, users, owners, insurance companies, contractors…). These ambiguities could lead a dilution of responsibility with the potential for long court procedures in the case of accidents along with strategic and political bargaining that may affect the development and adoption of ADS. The rich literature of moral philosophy (Van de Poel, 2011) could be a great help in clarifying the various meanings of the concept of responsibility and perhaps could help to find some innovative solutions. Facing such issues for decades, the nuclear industry has elaborated a doctrine based on the clear identification of 'who' is responsible for safety to avoid the risk of dilution of responsibility: 'The prime responsibility for safety must rest with the person or organization responsible for the facilities and activities that give rise to radiation risks (…) Authorization to operate a facility or conduct an activity may be granted to an operating organization or to an individual, known as the licensee. The licensee retains the prime responsibility for safety throughout the lifetime of facilities and activities, and this responsibility cannot be delegated'

(IAEA, 2006, p.6). This doctrine could inspire the reflection on responsibility for ADS, the problem being to collectively define the equivalent of the 'licensee' and to make it acceptable to all parties.

Clarifying the issue of responsibility leads to questions about the attribution of blame and legal liability. As mentioned earlier, scholars in other disciplines have explored the legal and ethical challenges related to highly automated driving systems. But as cyber-physical systems become more autonomous, it is the information system that becomes more important for safety. IS designers thus may bear huge responsibility for the use of such systems. Even with partially autonomous systems, when circumstances lead to unintended consequences (such as death or neuropsychological harms (Clegg et al., 2020)), should the designer of the fallback mechanism or the commander (Zhang Y et al., 2021), or even government (Pöllanen et al., 2020) be to blame? Encouraged by Stahl and Markus (2021), we suggest IS researchers should not simply leave the resolution of these issues to the courts, but actively engage in the debate regarding responsibility and liability. Governance of automated systems and associated legislation should become an important topic for IS research.

Trust in the wider ADS environment

A perennial topic in the IS research literature is trust in technology (McKnight et al., 2011; Bruckes et al., 2019). Koester and Salge (2020) say that for people to delegate full control to ADS, they need to establish sufficient initial trust in the automation's functionality, reliability and transparency. However, what makes trust a challenging topic with ADS is the complex nature of the environment. For example, if you purchase an automated car, do you trust the manufacturer (Cichy et al., 2021)? Do you trust other drivers to drive safely around you? Do you trust the software to perform correctly after a software update? Do you trust the law to protect your interests in the event of an accident? Will the insurance company cover your loss? These questions and more potentially affect a person's trust in ADS (Wiefel and Buxmann, 2021). It involves trust, not just in a particular technology, but in the wider social, political and regulatory environment as well (McKnight et al., 2011). We suggest IS researchers are well placed to examine these broader questions related to trust. Shneiderman (2020b) suggests that to obtain trustworthiness at a societal level, technology should be reliable, and a culture of safety developed with organizational controls and regulations.

Conclusion

It seems that we are rushing headlong into automation without properly understanding the consequences. To contribute to the nascent debate within IS about adopting autonomous cyber-physical systems such as autonomous driving systems (Lyytinen et al., 2022), we have sought to answer two research questions: *How do ironies of automation manifest when levels of automation of ADS increase beyond what we currently experience? Who is responsible and what does responsibility mean in the context of autonomous driving systems?* In answering these questions, we have identified some ironies related to the introduction of ADS and how they manifest at various level of automation. We hope we have contributed to a better understanding of moral responsibility (Van de Poel, 2011)) along with the corresponding legal aspects. There remains significant uncertainty and ambiguity regarding the distribution of responsibility between stakeholders and the coordination mechanisms that affect the delegation of control (Baird and Maruping, 2021). In this respect responsibility and controllability should not be confused.

The relentless increase in use of autonomous systems in our daily lives leads us to suggest that a reorientation is needed in IS research. While there is nothing wrong with a focus on the instrumental value of technology, we believe it would be a mistake for the IS discipline to neglect the broader societal issues associated with autonomous systems. Over time, as more tasks that used to rely on human experience and intuition become automated, these broader societal issues will take precedence. Hence, although we have focused on ADS in this paper, we call for IS research to critically examine the social, political and technical aspects of autonomous cyber-physical systems more generally. While a few IS papers have moved in this direction (e.g. Wiefel and Buxmann, 2021; Shneiderman, 2020b), we propose that the IS field needs to actively engage in research that considers the design, management and broader societal implications of all kinds of autonomous systems.

Such research has many practical implications. Facing the complexity of the interrelated outcomes (see Figure 1), and their uncertainty related to what would be a fair distribution of responsibilities, ADS is unlikely to be largely adopted unless their design and implementation is debated with the public. However, the issue of the modalities of participation and framing of such debate is difficult. Our critical approach based on the ironies might be an interesting avenue for such deliberation which might be applicable to other autonomous cyber-physical systems. That very significant uncertainty remains does not mean that some applied configurations operationalizing capabilities in certain environments cannot be successful. However, responsible design calls for more clearly identifying and specifying them. In other words, 'responsible design' should be open to some forms of public and/or expert inquiries (Dewey, 1927) that could be based on the ironies developed in this paper. Finally, the issue of whether we can continue using the traditional way software has been implemented –pushed by suppliers with bugs fixed later – becomes highly questionable.

Developing software this way raises huge concerns when systems put human lives at risk.

This paper has some limitations. First, it is limited by the broad interdisciplinary scope of our literature review. Second, our choice of bibliographic method meant that we had to ignore some papers that we considered as not central to our topic. Despite these limitations, we believe we have identified the most important ironies associated with ADS, and by extension, autonomous cyber-physical systems in general.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Frantz Rowe https://orcid.org/0000-0001-8520-1570
Maximiliano Jeanneret Medina https://orcid.org/0000-0003-4203-7680

Notes

1. Cyber-physical systems are those systems which integrate computational and physical capabilities, such as vehicles, aircraft and air-conditioning units (Tuunanen et al., 2019). ACPS are those cyber-physical systems that directly act upon the world without any direct human input. Autonomous driving systems are thus one type of ACPS.
2. When two documents cite the same document, they are bibliographically coupled. The higher the number of common documents cited by two documents, the higher is the bibliographic coupling strength of the two documents.
3. Autonomous/automated driving, smart/intelligent home, IoT.
4. We extracted the dataset at two different times, first on the 21th of December 2021, second on the 3rd of January 2023 to include documents published in 2022. The actualization of the citation count used as thresholds slightly impacted the documents screened for the in-depth analysis which has been performed during the year 2022, and actualized in early 2023 with three additional documents.
5. Identified on the AIS Library.

References (* = included in the 44 DBCA papers)

*Abbott R (2018) The reasonable computer: disrupting the paradigm of tort liability. *George Washington Law Review* 86(1): 1–45. DOI: 10.2139/ssrn.2877380.

*Alawadhi M, Almazrouie J, Kamil M, et al. (2020) A systematic literature review of the factors influencing the adoption of autonomous driving. *International Journal of Systems Assurance Engineering and Management* 11(6): 1065–1082. DOI: 10.1007/s13198-020-00961-4.

*Aniculaesei A, Grieser J, Rausch A, et al. (2018) Towards a holistic software systems engineering approach for dependable autonomous systems. In: *ACM/IEEE 1st International Workshop on Software Engineering for AI in Autonomous Systems Towards*, pp. 23–30. DOI: 10.1145/3,194,085.3194091.

Abbass H (2019) Social Integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11: 159–171.

Arnold M (2003). On the phenomenology of technology: the "Janus-faces" of mobile phones. *Information and Organization*. 13(4): 231–256. DOI: 10.1016/S1471-7727(03)00013-7.

Autor D, Mindell D and Reynolds E (2020) *The Work of the Future: Building Better Jobs in an Age of Intelligent Machines*. MIT Press.

Bainbridge L (1983) Ironies of automation. *Automatica* 19(6): 775–779. DOI: 10.1016/0005-1098(83)90046-8.

Baird A and Maruping LM (2021) The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly* 45(1): 315–341.

Banks VA, Eriksson A, O'Donoghue J, et al. (2018) Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics* 68: 138–145.

Banks VA, Plant KL and Stanton NA (2018) Driver error or designer error: using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety Science* 108: 278–285.

*Bartolini C, Tettamanti T and Varga I (2017) Critical features of autonomous road transport from the perspective of technological regulation and law. *Transportation Research Procedia* 27: 791–798. 10.1016/j.trpro.2017.12.002.

Baumann MF, Brändle C, Coenen C, et al. (2019) Taking responsibility: a responsible research and innovation (RRI) perspective on insurance issues of semi-autonomous driving. *Transportation Research Part A: Policy and Practice* 124: 557–572. 10.1016/j.tra.2018.05.004.

*Bazilinskyy P, Dodou D and de Winter J (2019) Survey on eHMI concepts: the effect of text, color, and perspective. *Transportation Research Part F: Traffic Psychology and Behaviour* 67: 175–194. doi:10.1016/j.trf.2019.10.013.

*Bennett JM, Challinor KL, Modesto O, et al. (2020) Attribution of blame of crash causation across varying levels of vehicle automation. *Safety Science* 132: 104968. doi:10.1016/j.ssci.2020.104968.

*Berge SH, Hagenzieker M, Farah H, et al. (2022) Do cyclists need HMIs in future automated traffic? An interview study. *Transportation Research Part F: Traffic Psychology and Behaviour* 84: 33–52.

Biondi F, Alvarez I and Jeong KA (2019) Human–vehicle cooperation in automated driving: a multidisciplinary review and appraisal. *International Journal of Human-Computer Interaction* 35(11): 932–946. DOI: 10.1080/10447318.2018.1561792.

Bonnefon JF, Shariff A and Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352(6296):1573–1576.

Boos D, Guenter H, Grote G, et al. (2013) Controllable accountabilities: the Internet of Things and its challenges for organisations. *Behaviour and Information Technology* 32(5):449–467.

Bornholt J and Heidt M (2019) To drive or not to drive - a critical review regarding the acceptance of autonomous vehicles. *ICIS Proceedings* 1–17.

Bruckes M, Grotenhermen J-G, Cramer F, et al. (2019) Paving the way for the adoption of autonomous driving: institution-based trust as a critical success factor. In: *Twenty-Seventh European Conference on Information Systems (ECIS 2019)*, 2019. https://aisel.aisnet.org/ecis2019_rp/87

Burton S, Habli I, Lawton T, et al. (2020) Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence* 279: 1–16.

Cabrall C, Eriksson A, Dreger F, et al. (2019) How to keep drivers engaged while supervising driving automation? A literature survey and categorisation of six solution areas. *Theoretical Issues in Ergonomics Science* 20(3): 332–365.

Cantu J, Tolk J, Fritts S, et al. (2020) High reliability organization (HRO) systematic literature review: discovery of culture as a foundational hallmark. *Journal of Contingencies and Crisis Management* 28(4): 399–410.

Cho E and Jung Y (2018) Consumers' understanding of autonomous driving. *Information Technology and People* 31(5): 1035–1046.

Cichy P, Salge TO and Kohli R (2021) Privacy concerns and data sharing in the internet of things: mixed methods evidence from connected cars. *MIS Quarterly* 45(4): 1863–1892.

Clegg FM, Sears M, Friesen M, et al. (2020) Building science and radiofrequency radiation: what makes smart and healthy buildings. *Building and Environment* 176: 1–15.

*de Bruin R (2016) Autonomous intelligent cars on the European intersection of liability and privacy: regulatory challenges and the road ahead. *European Journal of Risk Regulation* 7(3): 485–501. DOI: 10.1017/S1867299X00006036.

*De Chiara A, Elizalde I, Manna E, et al. (2021) Car accidents in the age of robots. *International Review of Law and Economics* 68: 106022. DOI: 10.1016/j.irle.2021.106022.

de Winter JC and Dodou D (2014) Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology and Work* 16(1): 1–11.

Dewey J. *The Public and its Problems*. Alan Swallow; 1927.

*Di X, Chen X and Talley E (2020) Liability design for autonomous vehicles and human-driven vehicles: a hierarchical game-theoretic approach*Transportation Research Part C: Emerging Technologies* 118: 102710. DOI: 10.1016/j.trc.2020.102710.

*Duboz A, Mourtzouchou A, Grosso M, et al. (2022) Exploring the acceptance of connected and automated vehicles: focus group discussions with experts and non-experts in transport. *Transportation Research Part F: Traffic Psychology and Behaviour* 89: 200–221.

Gandia RM, Antonialli F, Cavazza BH, et al. (2018) Autonomous vehicles: scientometric and bibliometric review. *Transport Reviews* 39(1): 9–28. 10.1080/01441647.2018.1518937.

*Geistfeld MA (2017) A roadmap for autonomous vehicles: state tort liability, automobile insurance, and federal safety regulation. *California Law Review* 105(6): 1611–1694. DOI: 10.15779/Z38416SZ9R.

Glaser BG (1978) *Theoretical Sensitivity*. Mill Valley, CA: Sociological Press.

Gogoll J and Muller J (2017) Autonomous cars: in favor of a mandatory ethics setting. *Science and Engineering Ethics* 23(3): 681–700.

Goodall NJ (2020) Machine ethics and automated vehicles. Preprint version. Published in Meyer G and Beiker S (eds.), *Road Vehicle Automation*, Springer, 2014; pp. 93–102. DOI: 10.1007/978-3-319-05990-7_9.

*Gupta R, Tanwar S, Kumar N, et al. (2020) Blockchain-based security attack resilience schemes for autonomous vehicles in industry 4.0: a systematic review. *Computers and Electrical Engineering* 86: 106717. DOI: 10.1016/j.compeleceng.2020.106717.

Hällgren M, Rouleau L and De Rond M (2018) A matter of life or death: how extreme context research matters for management and organization studies. *The Academy of Management Annals* 12(1): 111–153.

*Hancock PA (2019) Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics* 62(4): 479–495. DOI: 10.1080/00140139.2018.1498136.

Hevelke A and Nida-Rümelin J (2015) Responsibility for crashes of autonomous vehicles: an ethical analysis. *Science and Engineering Ethics* 21(3): 619–630.

Hoc JM (2000) From human–machine interaction to human–machine cooperation. *Ergonomics* 43(7): 833–843.

Hoc JM and Amalberti R (2007) Cognitive control dynamics for reaching a satisficing performance in complex dynamic situations. *Journal of Cognitive Engineering and Decision Making* 1(1): 22–55.

*Hopkins J and Hawking P (2018) Big data analytics and IoT in logistics: a case study. *International Journal of Logistics Management* 29(2): 575–591. DOI: 10.1108/IJLM-05-2017-0109.

IAEA. *Safety Standards Series No. SF-1*. Vienna: IAEA; 2006.

*Janssen CP, Iqbal ST, Kun AL, et al. (2019) Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. *International Journal of Human-Computer Studies* 130: 221–233. DOI: 10.1016/j.ijhcs.2019.07.004.

Jiao J, Zhou F, Gebraeel NZ, et al. (2020) Towards augmenting cyber-physical-human collaborative cognition for human-automation interaction in complex manufacturing and operational

environments. *International Journal of Production Research* 58(16): 5089–5111.

Johnson DG and Mulvey JM (1995) Accountability and computer decision systems. *Communications of the ACM* 38(12): 58–64.

Kalra N and Paddock SM (2016) Driving to safety: how many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94(C): 182–193.

*Karnouskos S (2021) The role of utilitarianism, self-safety, and technology in the acceptance of self-driving cars. *Cognition, Technology and Work* 23(4): 659–667. doi:10.1007/s10111-020-00649-6.

Ketter W, Schroer K and Valogianni K (2022) *Information Systems Research for Smart Sustainable Mobility: A Framework and Call for Action*. Information Systems Research: 1–21. DOI: 10.1287/isre.2022.1167.

Kim JH, Lee G, Lee J, et al. (2022) Determinants of personal concern about autonomous vehicles. *Cities* 120: 1–11.

Knauss A, Schröder J, Berger C, et al. (2017) Paving the roadway for safety of automated vehicles: an empirical study on testing challenges. *IEEE Intelligent Vehicles Symposium* (IV) 1873–1880.

Koester N and Salge TO (2022) Building trust in intelligent automation: insights into structural assurance mechanisms for autonomous vehicles. In: *International Conference on Information Systems 2020 Proceedings, 7*.

Koopman P and Wagner M (2017) Autonomous vehicle safety: an interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine* 9(1): 90–96.

*Kuhn M, Nguyen HG, Otten H, et al. (2018) Blockchain enabled traceability - securing process quality in manufacturing chains in the age of autonomous driving. In: *2018 IEEE International Conference on Technology Management, Operations and Decisions* (ICTMOD), 2018, pp. 131–136. DOI: 10.1109/ITMC.2018.8691242.

Lacity M and Willcocks L (2021) Becoming strategic with intelligent automation. *MIS Quarterly Executive* 20(2): 1–14.

Le Coze JC (2015) Reflecting on Jens Rasmussen's legacy. A strong program for a hard problem. *Safety Science* 71: 123–141.

*Lee D and Hess DJ (2020) *Regulations for On-Road Testing of Connected and Automated Vehicles: Assessing the Potential for Global Safety Harmonization. Transportation Research Part A: Policy and Practice 136*. Elsevier: 85–98. DOI: 10.1016/j.tra.2020.03.026.

*Lee YC, Momen A and LaFreniere J (2021) Attributions of social interactions: driving among self-driving vs. conventional vehicles. *Technology in Society* 66: 101631. DOI: 10.1016/j.techsoc.2021.101631.

*Leiman T (2020) Law and tech collide: foreseeability, reasonableness and advanced driver assistance systems. *Policy and Society* 40(2): 250–271. DOI: 10.1080/14494035.2020.1787696.

*Li J, Zhao X, Cho MJ, et al. (2016) From trolley to autonomous vehicle: perceptions of responsibility and moral norms in traffic accidents with self-driving cars. In: *Society of Automotive Engineers World Congress*, Detroit, MI, 2016. DOI: 10.4271/2016-01-0164.

*Liu N, Nikitas A and Parkinson S (2020) Exploring expert perceptions about the cyber security and privacy of Connected and Autonomous Vehicles: a thematic analysis approach. *Transportation Research Part F: Traffic Psychology and Behaviour* 75: 66–86.

*Liu S, Wang X, Hassanin O, et al. (2021) Calibration and evaluation of responsibility-sensitive safety (RSS) in automated vehicle performance during cut-in scenarios. In: *Transportation Research Part C: Emerging Technologies* 125: DOI 10.1016/j.trc.2021.103037.

*Lütge C, Poszler F, Acosta AJ, et al. (2021) AI4people: ethical guidelines for the automotive sector-fundamental requirements and practical recommendations. *International Journal of Technoethics* 12(1): 101–125. doi:10.4018/IJT.20210101.oa2.

Lyytinen K, Nickerson JV, Svahn F and Straub E (2022) Automated Driving Systems as a New Frontier for Information Systems. In: *International Conference on Information Systems 2022 Proceedings, 3*.

Mariani R (2018) An overview of autonomous vehicles safety. In: *2018 IEEE International Reliability Physics Symposium* (IRPS), 2018, pp. 6A. DOI: 10.1109/IRPS.2018.8353618.

Martin K (2019) Ethical implications of accountability of algorithms. *Journal of Business Ethics* 160: 835–850.

*Martinho A, Herber N, Kroesen M, et al. (2021) Ethical issues in focus by the autonomous vehicles industry. *Transport Reviews* 41(5): 556–577. DOI: 10.1080/01441647.2020.1862355.

*Mattas K, Makridis M, Botzoris G, et al. (2020) A Study Based on Empirical Observations. *Accident Analysis and Prevention*, 148: DOI: 10.1016/j.aap.2020.105794.

McCall R, McGee F, Mirnig A, et al. (2019) A taxonomy of autonomous vehicle handover situations. *Transportation Research Part A: Policy and Practice* 124: 507–522.

McKnight H, Carter M, Thatcher JB, et al. (2011) Trust in specific technology: an investigation of its components and measures. *ACM Transactions on Management Information Systems* 2(2): 25.

Merat N and Lee JD (2012) Designing highly automated vehicles with the driver in mind: prologue to a special section. *Human Factors* 54(5): 681–686.

Merfeld K, Wilhelms MP and Henkel S (2019) Being driven autonomously – a qualitative study to elicit consumers' overarching motivational structures. *Transportation Research Part C: Emerging Technologies* 107: 229–247. DOI: 10.1016/j.trc.2019.08.007.

Mick DG and Fournier S (1998) Paradoxes of technology: consumer cognizance, emotions, and coping strategies. *Journal of Consumer Research* 25(2): 123–143.

Mingers J and Walsham G (2010) Toward ethical information systems: the contribution of discourse ethics. *MIS Quarterly* 34(4): 855–870.

Myers MD and Klein HK (2011) A set of principles for conducting critical research in information systems. *MIS Quarterly* 35(1): 17–36.

*Noy IY, Shinar D and Horrey WJ (2018) Automated driving: safety blind spots. *Safety Science* 102: 68–78. doi:10.1016/j.ssci.2017.07.018.

Nyholm S and Smids J (2016) The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory & Moral Practice* 19(5): 1275–1289.

Parasuraman R, Sheridan TB and Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30(3): 286–297.

*Pek C, Zahn P and Althoff M (2017) Verifying the safety of lane change maneuvers of self-driving vehicles based on formalized traffic rules. In: *IEEE Intelligent Vehicles Symposium (IV)*, Redondo Beach, CA, 2017, pp. 1477–1483. DOI: 10.1109/IVS.2017.7995918.

Perrow C (2020) *Normal Accidents*. Princeton university press; 1984.

*Pöllänen E, Read GJM, Lane BR, et al. (2020) Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system. *Ergonomics* 63(5): 525–537. 10.1080/00140139.2020.1744064.

Rasmussen J (1997) Risk management in a dynamic society: a modelling problem. *Safety Science* 27(2): 183–213.

Reason J (1993) Managing the management risk: new approaches to organisational safety. In: *Reliability and Safety in Hazardous Work Systems: Approaches to Analysis and Design*. UK: Lawrence Hove: 7–22.

*Rezaei A and Caulfield B (2020) Examining public acceptance of autonomous mobility. *Travel Behaviour and Society* 21: 235–246. doi:10.1016/j.tbs.2020.07.002.

Rowe F (2014) What literature review is not: diversity, boundaries and recommendations. *European Journal of Information Systems* 23(4): 241–255.

Rowe F, Kanita N and Walsh I (2023). The importance of theoretical positioning and the relevance of using bibliometrics for literature reviews. *Journal of Decision Systems*. doi:10.1080/12460125.2023.2217646.

Russell S (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Ryan M (2020) The future of transportation: ethical, legal, social and economic impacts of self-driving vehicles in the year 2025. *Science and Engineering Ethics* 26(3): 1185–1208. DOI: 10.1007/s11948-019-00130-2.

SAE International (2016) *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.

*Salay R, Czarnecki K, Alvarez I, et al. (2020) PURSS: towards perceptual uncertainty aware responsibility sensitive safety with ML. In: *CEUR Workshop Proceedings*, 2020, pp. 91–95.

Sarker S, Chatterjee S, Xiao X, et al. (2019) The sociotechnical Axis of cohesion for the IS discipline: its historical legacy and its continued relevance. *MIS Quarterly* 43(3): 695–719.

*Schoitsch E (2016) Autonomous vehicles and automated driving: status, perspectives and societal impact. In: *24th Interdisciplinary Information Management Talks* (IDIMT), 2016, pp. 405–423.

*Shabanpour R, Golshani N, Shamshiripour A, et al. (2018) Eliciting preferences for adoption of fully automated vehicles using best-worst analysis. *Transportation Research Part C: Emerging Technologies* 93: 463–478. doi:10.1016/j.trc.2018.06.014.

Sheridan TB (1992) *Telerobotics, Automation, and Human Supervisory Control*. MIT press.

Sheridan TB (2011) Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 41(4): 662–667.

*Shladover SE and Nowakowski C (2019) Regulatory challenges for road vehicle automation: lessons from the California experience. *Transportation Research Part A: Policy and Practice* 122: 125–133. doi:10.1016/j.tra.2017.10.006.

Shneiderman B (2020a) Human-centered artificial intelligence: reliable, safe and trustworthy. *International Journal of Human Computer Interactions* 36(6): 495–504.

Shneiderman B (2020b) Human-Centered artificial intelligence: three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12(3): 109–124.

Smith HJ, Milberg SJ and Burke SJ (1996) Information privacy: measuring individuals' concerns about organizational practices. *MIS Quarterly* 20(2): 167–196.

Stahl BC and Markus ML (2021) Let's claim the authority to speak out on the ethics of smart information systems. *MIS Quarterly* 45(1): 485–488.

*Taeihagh A and Lim HSM (2019) Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews* 39(1): 103–128. DOI: 10.1080/01441647.2018.1494640.

Talebpour A and Mahmassani HS (2016) Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C: Emerging Technologies* 71: 143–163.

Tax SS, Brown SW and Chandrashekaran M (1998) Customer evaluations of service complaint experiences: implications for relationship marketing. *Journal of Marketing* 62: 60–76.

*Teoh ER and Kidd DG (2017) Rage against the machine? Google's self-driving cars versus human drivers. *Journal of Safety Research* 63: 57–60. 10.1016/j.jsr.2017.08.008.

*Tran DQ and Bae S-H (2021) Improved responsibility-sensitive safety algorithm through a partially observable markov decision process framework for automated driving behavior at non-signalized intersection. *International Journal of Automotive Technology* 22(2): 301–314. doi:10.1007/s12239–021–0029–z.

Tuunanen T, Kazan E, Salo M, et al. (2019) From digitalization to cybernization: delivering value with cybernized services. *Scandinavian Journal of Information Systems*. 31(2): 83–96.

Vagia M, Transeth AA and Fjerdingen SA (2016) A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics* 53(A): 190–202.

van de Poel I (2011) The relation between forward-looking and backward looking responsibility. In: Vincent N, van de Poel I and van den Hoven J, eds. *Moral Responsibility : Beyond Free Will and Determinism*. Dordrecht: Springer: 37–70.

van Eck NJ and Waltman L (2017) Accuracy of citation data in web of science and Scopus. *Journal of Informetrics* 9(3): 570–576.

Walsh I and Renaud A (2017) Reviewing the literature in the IS field: two bibliometric techniques to guide readings and help the interpretation of the literature. *Systèmes d'Information et Management* 22(3): 75–115.

Walsh I and Rowe F (2023). BIBGT: combining bibliometrics and grounded theory to conduct a literature review. *European Journal of Information Systems* 32(4): 653–674. doi:10.1080/0960085X.2022.2039563.

Walsh I, Renaud A, Jeanneret Medina M, et al. (2022) ARTIREV: an integrated bibliometric tool to efficiently conduct quality literature reviews. *Systèmes d'Information et Management* 27(4).

*Wang S and Zhao J (2019) Risk preference and adoption of autonomous vehicles. *Transportation Research Part A: Policy and Practice* 126: 215–229. doi:10.1016/j.tra.2019.06.007.

Weick KE and Roberts KH (1993) Collective mind in organizations: heedful interrelating on flight decks. *Administrative Science Quarterly* 38(3): 357–381.

*Wiedemann K, Naujoks F, Wörle J, et al. (2018) Effect of different alcohol levels on take-over performance in conditionally automated driving. *Accident Analysis and Prevention* 115: 89–97. doi:10.1016/j.aap.2018.03.001.

Wiefel J and Buxmann P (2021) *Automated Mobility as a Service: Development of a Hierarchical Quality Scale*. ECIS Proceedings.

Winkler T and Spierkermann S (2021) Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology* 23(4): 17–21.

*Wu J, Liao H and Wang JW (2020) Analysis of consumer attitudes towards autonomous, connected, and electric vehicles: a survey in China. *Research in Transportation Economics*, 80: 100828. DOI: 10.1016/j.retrec.2020.100828.

*Xu X, Wang X, Wu X, et al. (2021) Calibration and evaluation of the Responsibility-Sensitive Safety model of autonomous car-following maneuvers using naturalistic driving study data. *Transportation Research Part C: Emerging Technologies* 123: DOI: 10.1016/j.trc.2021.102988.

Zhang Y, Angell L and Bao S (2021) A fallback mechanism or a commander? A discussion about the role and skill needs of future drivers within partially automated vehicles. *Transportation Research Interdisciplinary Perspectives* 9: 1–31.

Zhang Z, Yoo Y, Lyytinen and Lindberg A (2021) The Unknowability of Autonomous Tools and the Liminal Experience of Their Use. *Information Systems Research* 32(4): 1192–1213.

Zupic I and Čater T (2015) Bibliometric methods in management and organization. *Organizational Research Methods* 18(3): 429–472.

Appendix

The table below gives complementary details about the bibliometric workflow in relation to the three first steps of BIBGT: 1) defining boundaries, 2) treating bibliographic data and theoretical sampling, and 3) clustering, mapping, and interpreting.

Appendix A.
Bibliometric workflow complementary details

Step	Task	Details
1	Data extraction	Scopus advanced query: *TITLE-ABS-KEY (('autonomous car' OR 'autonomous vehicle' OR 'autonomous automobile' OR 'automated car' OR 'automated vehicle' OR 'automated automobile' OR 'driverless car' OR 'driverless vehicle' OR 'driverless automobile' OR 'self-driving car' OR 'self-driving vehicle' OR 'intelligent car' OR 'intelligent vehicle' OR 'intelligent automobile' OR 'autonomous driving' OR 'automated driving' OR 'driving automation') AND safety AND (responsibility OR accountability OR liability)) AND DOCTYPE (ar OR cp OR re) AND SRCTYPE (j OR p) AND PUBSTAGE (final) AND LANGUAGE(english) AND PUBYEAR > 2015 AND PUBYEAR < 2023*. Considered only publications in journals and conference proceedings, written in English. This query resulting in 217 documents on the 3rd Jan 2023
2	Cleaning bibliographic references	Reference exclusion (e.g. no author or no title, blank, short references), fuzzy string similarity algorithms provided by ARTIREV and manual verification. On the initial 10192 single bibliographic references, 27% have been curated. We obtained 7491 single references that were used in further steps
2	Theoretical sampling	Processing of three alternatives based on bibliometric indices and a manual verification. a) Entire set of 217 documents b) Subset a of 89 documents, NCC >0.5 and CC > 1, 76 documents after a manual verification (35% of the set, retained); c) Subset B of 51 documents, NCC >1.0 and CC > 1, 44 documents after a manual verification (20% of the set, retained); Normalized citation count (NCC) and citation count (CC)
3	Clustering, mapping and interpretation	For each alternative, we performed the following steps: a) Clustering with association strength normalization, Leiden clustering algorithm with default parameters; b) Mapping with a clustered and hierarchical radial dendrogram. c) Interpretation of the resulting map and documents (title, abstract, keywords, cluster and bibliometric indices)

Appendix B.
Analytical category codes

Codes	Comments	Major coding issues/precision
Type of approach 1 social (in a hidden tech context) 2 social imperative 3 social and tech as separate antecedents 4HCI – Human-computer interaction 4SYT – Systems theory 4FIT – (mis)Fit theory 4VSD –Value sensitive design 5 – Technical imperative 6 – Technical	From Sarker et al. (2019) and above	Type 2. History and social norms in a specific setting (e.g. California (US) or Victoria (Australia) guide the building of regulations for autonomous driving
Task allocation NA – Not applicable D – driver DSY – driver and system SY – system 1 – Avoiding human supervision 2 – Reducing human supervision on an objective dimension 3 – Reducing human supervision on a subjective dimension 4 – Behaviourism support 5 – Cognitivism support 6 – Ecological support	From Cabrall et al. (2019). Coded as a combination of (D, SY, DSY) and (1–6). However, sometimes we could not identify a task allocation strategy at the level of Cabrall's typology and only coded D, SY or DSY. For 18 papers it was not applicable	Almost no author explicitly proposes a solution. Therefore, we coded the type of task allocation based on our shared interpretation against this grid We also coded (S, NS, SNS) for satisfying, non-satisfying and both (not shown in Table 3). We ended up with most papers coded SNS. This indicates that regardless of the strategy defended by authors, results are complex or ambiguous in relation to the ironies
Deskilling NA – Not applicable 1 scheduled handover 2 non-scheduled system-initiated handover 3 non-scheduled driver-initiated handover 4 non-scheduled driver-initiated emergency Handover 5 non-scheduled system-initiated emergency Handover C – concerned	User's decreasing ability to control the system. The more automated and autonomous the system becomes, the less qualified the user becomes due to a lack of practice The five handover types are from McCall et al. (2019)	Deskilling is linked to task allocation and handover issues *Multiple answers are possible*
Cognition NA – Not applicable O – Overload LSA – Level of situational awareness RE – representation based error	Cognitive factors involved in accidents	*Multiple answers are possible*
Trust NA – Not applicable T – trustworthy R – reliable Both – both trustworthy and reliable	Reliability does not mean trustworthiness. A measure can be reliably wrong	Reliability (as regularity) is a necessary condition for trustworthiness. We coded 'overreliance' as trust, not as reliability. Reliable always when 4HCI. Reliable in our coding means that not only it is analysed, but AD is considered more reliable (which is not the case for leiman)
Automation level 1 2 3 4 5 Any	Please refer to table 1	'Any' means any level from 1 to 5 We did not code 0, when papers were considering interactions between AVs and cars at 0 level

(continued)

Appendix B. (continued)

Codes	Comments	Major coding issues/precision
Moral responsibility NA – Not applicable C – concerned N – Neglect S –Strict	Neglect implies wrongdoing, besides causation and the negative consequence, while strict responsibility is related to being in charge	Establishing moral responsibility can be more difficult than legal liability because this would require access and understanding of AI algorithm
Accountability NA – Not applicable T – Transparency	Accountability practically refers to transparency	
Legal liability NA – Not applicable L – liability considered N – Neglect SP –Strict product	Neglect implies wrongdoing, besides causation and the negative consequence, while strict product responsibility is related to design or manufacturing defects	Multiple coding possible, including N and SP such as in Germany (see Lee and Hess, 2020). However, in a given paper it would be difficult to code both L and either N or SP.
Financial liability NA – Not applicable C – Concerned	Financial consequences attached to liability	We have decided to remove the column of financial responsibility from the table as its importance is not central to our research. This is an important gap
Stakeholder perspective AV – automated vehicle D– driver O – owner M – manufacturer TO – traffic or system operator RA – regulatory agency or government I – insurer CIT – citizens A – academics	Defense of a point of view and interest of one or several actors involved in autonomous systems	From the software in the vehicle can embed an RSS algorithm which will impact the behaviour of the automated vehicle. Most papers investigating RSS in different scenarios analyse outcomes from the AV perspective A when academics are actively involved in a proposed solution Passengers, like cyclists or pedestrians are all citizens
Stakeholder studied AV – automated vehicle D – driver O – owner M – manufacturer TO – traffic or system operator RA – regulatory agency or government I – insurer CIT – citizens	Analysis of the main stakeholders involved in autonomous systems be it at a design stage or in an experimental or natural setting. On a given paper there may be only a limited number of stakeholders	From the software in the vehicle can embed an RSS algorithm which will impact the behaviour of the automated vehicle. Most papers investigating RSS in different scenarios analyse outcomes from the AV studied notwithstanding other stakeholders but do not take account of interactions with users in the AV.
Domain predictability NA – Not applicable LC – Low complexity MC – Medium complexity HC – High complexity C – Concerned	Domain predictability becomes more complex with the number of interactions and with the heterogeneity of agents	We use complexity as a proxy for predictability as follows - Low complexity (LC) (typically isolated-road section) - Medium complexity (MC) (typically freeway) - High complexity (HC) (typically urban environment, or severe weather conditions)
Software predictability NA – Not applicable C – Concerned	Decisions taken by software are explainable in all circumstances	If particularly interesting code S*, but given the paper is unique we did not add a specific code
Handover predictability NA – Not applicable C – Concerned	Switch in control over the system (from human to machine and conversely). It can be unpredictable in case of emergency (see McCall et al., 2019)	We removed the column of HP from the table as very few papers treated it in the covered literature. This is an important gap
Safety risk DI – physical discomfort M – mental (psychological) problems I – physical injury D – death COL – collision C – concerned		All combinations possible. If we could not identify some precise argument or focus on one or several of the different types of risks from DI to COL, then we code C (concerned)

Appendix C.
Coding of the 44 research front papers

Cluster	Author	S persp	S studied	Type of approach	TA	Desk	Cogn	Auto lvl	Risk	MR	Ac	LL	Trust	Dom pred	Soft pred
1	Karnouskos (2021)	CIT/D	CIT/D	3	SY-S-1	NA	LSA	3, 4, 5	D/I	C	T	NA	T	NA	C
1	Alawadhi et al. (2020)	A/CIT	CIT/D/M/RA/TO	3	DSY-S-3	NA	LSA	Any	C	C	T	N/SP	both	HC	C
1	Liu et al. (2020)	D/M/O/RA/TO	AV	3	NA	NA	NA	5	C	NA	T	N	T	HC	C
1	Rezaei and Caulfield (2020)	O	O	3	NA	3, 4	NA	3,4,5	C	NA	T	L	T	NA	NA
1	Duboz et al. (2022)	CIT/D	AV/CIT/D/M/O/RA/TO	2	DSY-S-3	NA	LSA	Any	COL/D/DI	C	NA	L	both	HC	C
1	Kim et al. (2022)	D/O	CIT/D/O	3	NA	NA	NA	5	COL	NA	NA	L	NA	HC	NA
1	Shabanpour et al. (2018)	O	O	3	NA	NA	NA	3,4.5	C	NA	NA	L	NA	NA	NA
1	Wang and Zaho (2019)	I/O	O	2	NA	NA	NA	5	C	NA	NA	NA	NA	NA	NA
1	Wu et al. (2020)	CIT	CIT	3	NA	NA	NA	Any	C	NA	NA	L	NA	NA	NA
2	Shladover and Nowakowski (2019)	RA	D/M/RA	2	SY-1	NA	NA	3,4,5	D/I	S	T	N/SP	both	HC	C
2	Bennett et al. (2020)	CIT	CIT/D/I/M/O/RA/SO	1	DSY-S-3	NA	LSA	Any	I	N/S	T	N/SP	both	HC	C
2	Taeihagh and Lim (2019)	RA	D/I/M/RA	4SYT	NA	NA	NA	4, 5	C	N	T	N/SP	NA	HC	C
2	Geistfeld (2017)	M/RA	CIT/D/I/M/RA	2	DSY-S-3	NA	NA	Any	I	N/S	T	N/SP	both	HC	C
2	Lütge et al. (2021)	A/M/RA	M/RA	4VSD	DSY-SNS	NA	NA	Any	C	N/S	T	N/SP	both	NA	C
2	Pöllänen et al. (2020)	RA	CIT	3	NA	NA	NA	3,5	COL/D/I	S	T	SP	NA	HC	C*
2	Lee and Hess (2020)	RA	CIT/D/I/M/O/RA/TO	4FIT	SY-S-3	3	LSA/RE	3,4,5	D/I	NA	T	N/SP	T	HC	C
2	Ryan (2020)	RA	CIT/D/I/M/O/RA/TO	4FIT	DSY-SNS	2,5	LSA/O/RE	3,4	C/D/I	C	T	N (3)/SP (4,5)	T	NA	C
2	Leiman (2020)	D/RA	D/m/RA	4FIT	DSY-SNS-2	C	LSA/RE	Any	C/D/I	N	T	N/SP	T	C	NA
2	Martinho et al. (2021)	CIT/D/M	AV/M	4HCI	NA	2,5	LSA	Any	C/D	N	T	N	both	NA	C
2	De Bruin (2016)	CIT/D/M/O	CIT/D/M/O	4FIT	NA	NA	NA	Any	C	N/S	T	N/SP	T	NA	C
2	De Chiara et al. (2021)	D/M	D/M	2	D	NA	NA	5	C	N/S	T	N/SP	NA	NA	NA
2	Di et al. (2020)	CIT/D/M/RA	CIT/D/M/RA	2	NA	NA	NA	5	COL	N/S	NA	N/SP	NA	HC	NA

(continued)

Appendix C. (continued)

Cluster	Author	S persp	S studied	Type of approach	TA	Desk	Cogn	Auto lvl	Risk	MR	Ac	LL	Trust	Dom pred	Soft pred
2	Abbott (2018)	CIT/RA	D/M/I	2	DSY-SNS	C	NA	3, 5	D/I	N/S	NA	N/SP	NA	HC	NA
2	Schoitsch (2016)	RA	D/I/M/RA	3	DSY-SNS	NA	LSA/RE	Any	COL/D/I	NA	NA	SP	T	HC	C
2	Li et al. (2016)	CIT/D/M/RA	CIT	2	NA	NA	NA	5	COL/I	S	NA	SP	NA	HC	C
2	Bartolini et al. (2017)	D/I/M/O/RA	RA	1	NA	NA	NA	5	D	NA	NA	SP	NA	NA	NA
3	Noy et al. (2018)	D/M/RA	D/M/RA	4SYT	DSY-NS	3,4,5	LSA/RE	Any	D	C	T	SP	both	HC	C
3	Hancock (2019)	D/M	D	4HCI	DSY-SNS-5	1,2,3,4,5	RE	Any	C	C	T	L	both	C	C
3	Berge et al. (2022)	AV/CIT	CIT	4HCI	NA	NA	LSA/O/RE	5	C	NA	NA	SP	T	HC	C
3	Wiedemann et al. (2018)	D/M	D	4HCI	DSY-SNS-4	C	LSA	3	COL	N	NA	N	R	MC	NA
3	Teoh and Kidd (2017)	AV/M	AV/M	3	DSY-SNS-1	NA	NA	4,5	COL/I	NA	NA	NA	NA	HC	NA
3	Janssen et al. (2019)	D/M	D/M	4HCI	DSY-SNS-5	2,3,4,5	LSA	3 and 4	C	NA	NA	NA	R	NA	NA
3	Bazilinskyy et al. (2019)	CIT/M	CIT/D	4HCI	NA	NA	RE	3,4,5	C	NA	NA	L	T	NA	NA
3	Lee et al. (2021)	AV/D	D	3	DSY	NA	RE	3,4,5	C/M	N	NA	NA	NA	HC	C
4	Gupta et al. (2020)	A	AV	6	NA	NA	NA	Any	C	NA	T	NA	T	NA	NA
4	Hopkins and Hawking (2018)	O	O	5	D-4	5	LSA	2	D/I	NA	T	L	NA	NA	NA
4	Kuhn et al. (2018)	M	M	5	NA	NA	NA	Any	I	NA	T	SP	both	NA	NA
5	Salay et al. (2020)	D/M/TO	M	6	DSY-S-3	NA	RE	5	C	N	T	NA	R	HC	C
5	Pek et al. (2017)	AV	AV	6	SY-S-1	NA	NA	3,4,5	COL	N	T	L	R	MC	C
5	Liu et al. (2021)	AV/M	AV	6	SY-S-1	NA	NA	3,4,5	COL	NA	NA	L	R	MC	C
5	Mattas et al. (2020)	AV	AV	6	SY-S-1	NA	NA	3,4,5	COL/DI	NA	NA	NA	NA	C	C
5	Tran and Bae (2021)	AV/M	AV/CIT	6	SY-S-1	NA	LSA	3,4,5	COL/DI	NA	NA	NA	R	HC	NA
5	Xu et al. (2021)	AV	AV	6	NA	NA	NA	Any	COL/D	NA	NA	L	R	HC	C
6	Aniculaesei et al. (2018)	M	D/M	6	SY-S	NA	RE	Any	C	NA	T	SP	both	HC	C

Author biographies

Frantz Rowe is Professor of Management of Information Systems at Institut d'Administration des Entreprises (IAE), Nantes University, and a member of LEMNA research lab. Frantz is a Fellow of the Association for Information Systems and a Senior Member of the Institut Universitaire de France where he holds a Fundamental Chair on digital entrapment and digital transformation, with related projects (e.g. critical social theory, technical debt, generative AI). His last papers focus on false consciousness and digital entrapment (in *Information Systems Research*), on the opportunities and limitations of types of AI for literature reviews, and on theoretical diversity and rivals on a given topic (both in the *Journal of the Association for Information Systems*). He has a PhD from the University of Paris, an ME from ENTPE, Lyons, France and an MS from UC Berkeley, USA.

Maximiliano Jeanneret Medina is a researcher at the Institute of Organization Digitalisation of the Haute école de gestion Arc, part of the University of Applied Sciences and Arts Western Switzerland, and a PhD student at the Human-IST Institute of the University of Fribourg. His research focuses on the design, development, and evaluation of interactive and accessible computer systems, as well as bibliometric tools and techniques to support literature reviews.

Benoit Journé is Professor of Strategic Management and Organization Sciences at Nantes University (Institut d'Administration des Entreprises (IAE)). He is a member of LEMNA research lab where he carries out qualititative researches in the domain of organizational reliability and resilience, with a focus on nuclear safety. He founded a research Chair in partnership with French nuclear industry and led a major post-Fukushima research project. His research has been published in various journals. https://orcid.org/0000-0001-9207-6317.

Emmanuel Coëtard is a PhD student in Human Resources Management at Nantes University. Researcher at LEMNA, he is a member of the 'Regulations and Transformations of Work' research group. His research focuses on the relationship between work and careers in socially devalued occupations, and our social representations of these occupations. He is a member of the French-speaking Human Resources Management Association.

Michael D. Myers is Professor of Information Systems in the Department of Information Systems and Operations Management at the University of Auckland Business School, New Zealand. His research interests are in the areas of digital transformation, the social, organizational, and cultural aspects of digital technologies, and qualitative research methods in information systems. Michael currently serves as Editor-in-Chief of *European Journal of Information Systems*. Michael is a Fellow of the Association for Information Systems and was awarded The LEO Award for Lifetime Exceptional Achievement by the Association for Information Systems in 2019. He also received the Silver Core award by the International Federation for Information Processing in 2007. https://orcid.org/0000-0001-8525-6395.