Irresponsibilities, inequalities and injustice for autonomous vehicles

Liu, Hin-Yan

Published in:
Ethics and Information Technology

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Liu, H-Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. Ethics and Information Technology, 19(3), 193-207.

Ethics Inf Technol (2017) 19:193–207
DOI 10.1007/s10676-017-9436-2

ORIGINAL PAPER

Irresponsibilities, inequalities and injustice for autonomous vehicles

Hin-Yan Liu[1]

Published online: 22 August 2017

Abstract With their prospect for causing both novel and known forms of damage, harm and injury, the issue of responsibility has been a recurring theme in the debate concerning autonomous vehicles. Yet, the discussion of responsibility has obscured the finer details both between the underlying concepts of responsibility, and their application to the interaction between human beings and artificial decision-making entities. By developing meaningful distinctions and examining their ramifications, this article contributes to this debate by refining the underlying concepts that together inform the idea of responsibility. Two different approaches are offered to the question of responsibility and autonomous vehicles: targeting and risk distribution. The article then introduces a thought experiment which situates autonomous vehicles within the context of crash optimisation impulses and coordinated or networked decision-making. It argues that guiding ethical frameworks overlook compound or aggregated effects which may arise, and which can lead to subtle forms of structural discrimination. Insofar as such effects remain unrecognised by the legal systems relied upon to remedy them, the potential for societal inequalities is increased and entrenched, situations of injustice and impunity may be unwittingly maintained. This second set of concerns may represent a hitherto overlooked type of responsibility gap arising from inadequate accountability processes capable of challenging systemic risk displacement.

Keywords Autonomous vehicles · Structural discrimination · Responsibility · Network-effects · Inequality · Trolley-problem ethics · Risk allocation · Impunity

Introduction

Autonomous vehicles appear poised to enter society in the near future, considering the rapid and accelerating advancement of the technologies this millennium. While the technology has forged ahead relentlessly, our legal and ethical frameworks remain mired in doubt and seemingly paralysed in face of the challenges wrought forth by the prospect of vehicles functioning without a direct human operator. Central to this conundrum is the question of responsibility for the negative outcomes of autonomous vehicle operation, which is now verging upon statistical inevitability (Robbins 2016): the beleaguered 'who is responsible' question. Yet, the persistence of the responsibility issue suggests that the question has not been adequately refined, leading to proposals resting on assumptions and answers being couched in generalities. This article aims to infuse nuance into the overarching responsibility question: by questioning the necessity for ascribing responsibility at all; by differentiating between the disparate concepts that together inform existing notions of responsibility; and by approaching the issue from the polar perspectives of targeting and risk distribution. In defining these constituent issues, this article narrows the responsibility issues and identifies more precisely the nature of the persistent problems which plague attempts to define models of responsibility for autonomous vehicles.

The article then discusses moves that circumvent or marginalise the responsibility question which are made possible by the prospect for autonomous vehicles to have

✉ Hin-Yan Liu
hin-yan.liu@jur.ku.dk

[1] Centre for International Law, Conflict and Crisis, Faculty of Law, University of Copenhagen, Copenhagen, Denmark

pre-set preferences for accident scenarios. These preferences essentially allocate risks in advance of accidents in order to optimise crashes by seeking to minimise objective levels of damage. While lauded as an important improvement that promises to save lives and protect property, the possibility that risk allocation preferences for autonomous vehicles forecloses important aspects of the responsibility discussion. Furthermore, such risk allocation preferences can be coordinated and consolidated by designers and manufacturers across a fleet of autonomous vehicles, exerting systemic pressures that consistently structure the allocation of the risks posed by those autonomous vehicles. Such cumulative effects mask the potential for subtle, yet powerful, structural inequalities to be entrenched within the algorithms that govern their behaviour, but which remain unrecognised by contemporary legal concepts and doctrines.

This subtle scenario is explored through a thought-experiment which takes the logic of crash optimisation to speculatively radical extremes. The quest to minimise objective levels of damage might involve, indeed mandate, qualitative factors to be entered into the calculus. Optimisation might no longer be restricted to minimising damage, but transform into preventing the greatest forms of harm in more proactive ways. These perspectives suggest the limits of contemporary conceptions of responsibility in relation to the introduction of autonomous vehicles, suggesting that resolving the debates on responsibility and legal liability is necessary but insufficient in governing their introduction onto public roads.

Why responsibility?

Autonomous vehicles have been projected to fundamentally alter the nature of driving as an activity by substituting and rendering obsolete the human driver. Architectures for liability and compensation need to be constructed where such activities result in the production of harm to avert allegations of impunity.

The etymology of responsibility anchors the notion in the idea of responding or answering to, or being accountable for, a given action which is usually a wrong. The need for responsibility is rooted in relation to a perceived transgression. With autonomous vehicles, subject to the caveats that the programming was undertaken without malice and that their software has not been otherwise compromised, it may be difficult to categorise the harms they generate as moral wrongs or legal injuries.

Taking moral wrongs first, despite the obvious and necessary application of ethics to the operation of autonomous vehicles (Lin 2015) it is neither immediately obvious nor necessarily the case that ethical codes would be violated by specifying one ethical framework or set of morally-relevant priorities over another. This conclusion can be deduced from the ethical debates delving beneath this surface to discuss the finer points of applying ethical theories (Goodall 2014; Gerdes and Thornton 2015; Bonnefon et al. 2015) which identify problematics and preferences from particular moral perspectives. The point is that there is no sign of agreement as to what course of action is morally correct in any specific setting; rather a range of possibilities remain open that may be more or less morally defensible. Absent a clear prospect for unambiguous moral wrongs that are triggered by the prospect of autonomous vehicles, it may be that questions of responsibility in this realm can be deferred or suspended.

Moving to the legal categorisation, the threshold concern is how a particular harm is seen through the jurisprudential lens (Liu 2015). Neil MacCormick articulates this legal bias succinctly:

> The special sort of [jurisprudential] reasoning is one which *leaves aside* any general and abstract deliberation on what in a given context it would be best or would all things considered right to do or not to do. *Where law is appealed to, all things are not considered*. (MacCormick 1995. Emphasis added.)

Scott Veitch builds upon this selective admission of acceptable considerations to illustrate the blinkered and tautological nature of legal processes:

> What registers in law as a wrong? This question has a deceptively simple answer: *what registers in law as a wrong is a breach of the law*. Strictly speaking, then, it is not any particular loss—physical suffering, economic harm, or whatever—that registers as such, *but rather only that suffering or harm that is given legal cognisance*. (Veitch 2007. Emphasis added.)

Woven together, these arguments ground the legal myopia that may be capable of overlooking injuries. At this point, the distinctions drawn in this article in relation to the terminology need to be properly introduced: harm, damage, wrong and injury. Both harm and damage connote negative outcomes that can be devoid of ethical or legal consequences, and relate to the concept of casual responsibility elaborated below. The notion of wrong implies an infraction with potential moral, and possibly legal, consequences. As such, wrongs begin to assert the need for responsibility, yet at the same time conflate role and causal concepts of responsibility discussed below. Finally, injury as understood in its root sense of *injuria*, meaning an invasion of another's rights or conversely a legally actionable wrong, demonstrate the distance between the harm and its (legal) recognition. This gap obscures the prospect of damage occurring without injury being recognised (Veitch 2007), leading to the possibility that the liability conception of responsibility rendering obsolete the broader question of responsibility.

To unpack the last argument, two responsibility-related questions can be articulated in the present context: should the harms engendered by autonomous vehicles be categorised as damage, wrongs or injuries; and what are the broader implications of this taxonomy? The irrelevance of responsibility to the operation of autonomous vehicles can be asserted where any harm or damage these cause fail to be recognised as moral wrongs or legal injuries. While the autonomous vehicle might be responsible for bringing about a particular harm in a causal sense (Hart 2008), it is plausible that no infraction will be registered such that the harm is treated analogously to natural phenomena.[1] In such instances, restitution may become the driving principle to restore as closely as possible the situation the victim was in had the harm not occurred.[2] The broader implications of this legally-dominated classification scheme, therefore, include the possibility that wrongs and injuries may be foreclosed through moral and legal interpretations.

Flowing from these conclusions are consequences for the forms of compensation that victims of autonomous vehicle harms may take. Where such harm remains in the realm of damage, a restitutionary mechanism of compensation should suffice to undo as far as possible the detriment suffered by the victim, and can take the form of an insurance or other collectivised risk-mitigation scheme (including models of strict liability such as that proposed by Hevelke and Nida-Rümelin 2015). From this perspective, legalistic questions of responsibility-ascription are both unnecessary and convoluted: the victim can have ready access to adequate and appropriate compensation without the need to resort to the legal mechanisms, and the courts will not be required to overstretch existing modalities of responsibility to ensure that victims are not left without redress. Thus, the advantages to leaving responsibility out of the discussion are that practical avenues can be smoothed without undermining legal doctrine.

Despite good reasons to circumvent the thorny issues of responsibility entirely, however, the ascription of harms caused by autonomous vehicles may not appropriately remain within the sphere of harms and damage, for a plethora of reasons revolving around the possibility for the autonomous vehicle to exercise discretion. In other words, moral and legal questions arise as the direct result of the ability of the car to function autonomously: to commit itself to one course of action among many, and to do so based upon a set of parameters. Thus, discretion is the lynchpin of autonomy and introduces concepts of responsibility (Liu 2016) through the programming which establish the parameters and the actual behaviour of the autonomous vehicle.

Furthermore, unlike the refined responsibility questions raised in the context of autonomous weapons systems and the use of lethal force, a distinction need not be drawn between functional and discretional autonomy (Liu 2016).[3] Unlike the autonomous weapons system where the source of the hazard is in the explicit targeting function, with the autonomous vehicle the hazard inheres within its ordinary operation. Functional and discretional autonomy are collapsed because the autonomous vehicle allocates the potential harms it generates as a by-product of getting from one point to another, unlike an autonomous weapons system which actively selects which target to engage.

Taken together, these perspectives provide the basis for policies which can distribute the harms created by autonomous vehicles through a compensatory model that circumvents the responsibility conundrum beyond establishing causation.[4] Alternatively, policies may have to mandate detailed examinations of the precise contours of responsibility and liability, to which we now turn.

The rope metaphor of responsibility

To introduce and illustrate the notion of responsibility that is being unravelled in this article, it is useful to draw upon an analogy with the rope in order to highlight several key features. At the functional level, the rope is connective and relational: for it to be useful its two ends must be connected and in doing so it joins two particular objects together. Put differently, the true utility of the rope would not be properly appreciated if it was not deployed to attach two discrete points together. Furthermore, aside from abstractions, the two points or objects that are linked together by the rope are clearly identifiable and cannot be readily interchanged without affecting the relationship between both the end points or between the object and the rope.

Inspecting the rope for its intrinsic characteristics it becomes readily apparent that despite appearing, and being labelled, as a singular entity a rope in reality braids together several independent strands in order to deploy and reinforce them towards the same ends. Considered in this light, the apparent unity of the rope is subtly misleading for three

[1] HLA Hart, in proposing the typology of responsibility, suggested that 'it is clear that in this causal sense not only human beings but also their actions or omissions, and things, conditions, and events, may be said to be responsible for outcomes', p. 214.

[2] Whereas wrongs and injuries import blame, the neutral status of harm and damage may be remedied by reinstating the position prior to the event triggering the harm or damage.

[3] Functional autonomy describes systems which are capable of undertaking only predetermined or strictly limited forms of independent action, while systems possessing discretional autonomy substitute human decision-making processes in its domain, p. 327.

[4] Existing models for this include modalities of strict liability, as well as mandatory insurance scehemes.

reasons. The first contextualises this unity in relation to the functioning of the rope. Thus, while the notion of the rope exists in the abstract, what is really doing the binding work in the real world are the interwoven strands that are pulling in the same direction. This view also reveals that the strength of the rope is contingent upon that composition of, and interaction between, the component strands and also that each strand is capable of fulfilling analogous functions independently. Second, the labelling convenience reveals the rope as a heuristic labelling device, it is a short cut for describing a generic object used to bind things together. This highlights the thoughtlessness, usually unproblematic, of calling the rope 'the rope' even though such treatment obscures in reality many nuanced details. The third and perhaps most important, albeit seemingly trite, observation the rope is very often used, but rarely contemplated with such detail. Like with labelling, this treatment is largely justified because the utility of the rope is in its function, which remains the case only until the point where its composition becomes relevant to its usefulness.

A final set of observations that can be made about the rope is that it has a tensile strength, and it may break if stretched beyond its limits. But the precise breaking point is governed by two main factors, the composition of the rope and what objects it tied together. The nature of the braided strands largely determines the ultimate strength of the rope, but its real world performance also depends upon what it is deployed to join together.

The purpose of this digression is to draw out precise analogies with the notion of responsibility which reflect the detailed description of the rope above point for point. Thus, the first observation is that responsibility is necessarily relational: this observation can also be discerned from the shared etymological root of the 'responsibility' with 'response'. In this sense, the notion of responsibility is a connector that links together an actor with an action or consequence, such that the notion cannot stand by itself. As a result, the notion of responsibility is determined to a large extent by its relational context. The context in which responsibility is applied sets the tone for both the content and limitations of the notion.

Turning next to the intrinsic characteristics, the notion of responsibility interlaces several distinct and independent, albeit interrelated, concepts that work together towards shared goals. In this respect, however, it should be clear that the responsibility 'rope' need not be comprised of the same strands across situations, but rather that different strands may be plaited together as the specific situation dictates (these conceptual strands are introduced properly in the following section). In this light, the notion of responsibility can be seen as an overarching framework that is populated by appropriate and relevant concepts dictated by the situation. Following from this, and mirroring the first reason that a blanket approach to responsibility is misleading, the precise meaning of 'responsibility' becomes contingent upon actual responsibility concepts that are in fact deployed. As we will see below, different meanings, limitations and consequences attach to the different concepts of responsibility which feedback to determine both the content and the contours of 'responsibility' in any given situation. Thus, recourse to these constituent concepts can calibrate our expectations with respect to responsibility issues and reveal the boundaries of possibility inherent in deploying those concepts.

The second reason that 'responsibility' can be misleading then is that this is bandied about as a generic and blanket notion, readily discernible in claims that there will always be ultimate responsibility for autonomous vehicles (Marchant and Lindor 2012).[5] But it matters a great deal what form of responsibility is applied: crudely, it is clear that manufacturer responsibility is governed by omission rather than commission (this is discussed in greater detail below). Put differently, claims that a human being will be ultimately responsible for autonomous vehicles may be simultaneously true while missing the point entirely: true in the sense that a human being may bear some form of omission responsibility in relation to the autonomous vehicle, while remaining tangential to societal concerns which gravitate towards establishing responsibility for active commission of a particular harm.

Building from this, the under-examined nature of 'responsibility' is the third reason that appeals to the notion can be misleading. Absent a differentiated and calibrated model of responsibility assertions feel empty and unsatisfying, doing little to assuage the broad concerns flowing from emerging technologies and the ensuing questions of 'responsibility'. Yet, because 'responsibility' remains relatively unexplored, discussions can but skim the surface and neither the crux of the problem nor possible avenues leading towards solutions can be identified.

Finally, the cursory treatment of 'responsibility' thus risks over-extending the notion, creating increasingly tenuous connections and artificial assertions with relation to autonomous vehicles and human beings. The metaphor of tensile strength that was alluded to above applies in the conceptual context to indicate the limits to applying 'responsibility' to connect novel things together. In this sense, over-extension risks undermining the associated legal concepts and delegitimating the very idea that responsibilities, in whatever form, are an important considerations.

[5] Marchant and Lindor in this case stipulate the caveat that the manufacturer is ultimately responsible from a doctrinal perspective, which suggests that the flaw in reasoning is with the legal doctrine upon which they comment.

Role and capacity responsibility; causal responsibility

Having sketched the contours of responsibility as a broad notion, and having begun its dissection, we are now in a position to define the constituent concepts (or strands, in the rope metaphor) that combine under the notion of responsibility.

The core question of responsibility raised by autonomous vehicles is the replacement of the human driver with that of an artificial entity. While artificial entities, such as corporations, can possess legal personality and enjoy legal rights (in the United States, see Santa Clara County v Southern Pacific Railroad 1886, for the United Kingdom, see Salomon v A Salomon 1897), ascribing corporate responsibility has remained a persistent problem (Coffee 1981; Bakan 2005; Seck 2011) and may be a harbinger of the obstacles in the road ahead for autonomous vehicles. In the context of autonomous vehicles, however, the direct substitution of the human being as the causal agent is clear. To bridge the ensuing void of causal responsibility created by the autonomous vehicle (see generally, Matthias 2004), recourse has been made to omission-based forms of role responsibility, characterised in particular by the duty to intervene (Hevelke and Nida-Rümelin 2015).

In order to identify the contours and content of the various concepts that together comprise the notion of responsibility, it is useful to explicate the influential four-part typology developed by HLA Hart, which distinguished between role-responsibility, causal responsibility, liability-responsibility, and capacity-responsibility (Hart 2008). Role responsibility is defined by the performance of prescribed duties and can be understood as sufficient efforts towards fulfilling a set of defined obligations attaching to one's role or position. Causal responsibility connects causes with consequences, but neither needs to include value judgments nor necessarily involve legal ramifications. Liability responsibility, as discussed above, involve the circuitous definition that centres upon legally-recognised injuries. Finally, capacity responsibility predicates responsibility ascription upon the possession and exercise of certain traits and abilities: 'the ability to understand what conduct legal rules or morality require, to deliberate and reach decisions concerning these requirements, and to conform to decisions when made' (Hart 2008).

Having established this responsibility framework, we are now in a position to articulate accurately the responsibility issues introduced by the autonomous vehicle. It would be uncontroversial to assert that any harm caused by such a vehicle would mean that the vehicle would be responsible for the harm in a causal sense. Yet, because there are no moral or legal consequences that necessarily flow from such a determination, such a characterisation effectively amounts to labelling the harm as mere damage. Such a course of action may be unsatisfactory to the extent that it limits the victim to restitution, and where it fails to recognise any wrong or injury that is actually suffered as a result of the discretionary potential possessed by the autonomous vehicle. Taken to the extreme, where the harm engendered by the autonomous vehicle affects the life or physical integrity of human beings, this treatment may amount to the negation of human dignity and a violation of human rights.[6] The human being is treated in an instrumental fashion where the autonomous vehicle had committed itself to a course of action resulting in injury to human beings, which remains partially or incompletely recognised by the legal systems as an injury.

The human occupant[7] of the autonomous vehicle is placed in a position where she bears responsibility for the actions of the vehicle over which she has limited or no effective control. This statement immediately collapses role responsibility together with capacity responsibility, highlighting the intractable problems of borne by the human being in such a position. The duty to intervene is one such proposal (Hevelke and Nida-Rümelin 2015, an alternative characterisation is that of the auto-pilot; Douma and Palodichuk 2012) and involves human oversight of the autonomous vehicle on the road. Thus, the obligation is placed upon the human occupant to remain vigilant for, and to anticipate reasonably foreseeable accidents, and to intervene accordingly. Yet, on a practical note as the authors concede, 'it seems implausible that the otherwise idle user will be able to stay focused and searching for a possible risk of an accident which might occur on average once every 2 million kilometres or so' (Hevelke and Nida-Rümelin 2015). Research involving unmanned aerial vehicles in the military domain suggests that human boredom degrades both reaction times and the ability to maintain directed attention in oversight tasks (Cummings et al. 2013) which corroborate this claim. Taken together, this undermines the prospect for capacity responsibility for the occupant human beings for the harms caused by their autonomous vehicles.

Locating the duty to intervene within the realm of role responsibility and capacity responsibility underscore the limitations of this approach to bridging the responsibility gap. Taking the issue of role responsibility first, it is clear that this conception of responsibility is both grounded in and

[6] Christof Heyns articulates the human dignity argument: 'Death by algorithm means that people are treated simply as targets and not as complete and unique human beings' (Heyns 2016).

[7] The human beings physically in the vehicle have been characterised in the literature primarily as the owners. Obviously, a proprietary relationship with the vehicle need not be the determining variable in the responsibility calculus. Instead, a more encompassing and consistent criterion could be characterised on the basis a beneficiary status in relation to that vehicle.

curtailed by the scope of obligations borne by the individual. Thus, for role responsibility to function as a substitute for causal responsibility, both substantive content as well as the demarcated boundaries of the obligation must be specified clearly in advance. While this is not an absolute barrier to establishing role responsibility for autonomous vehicles, much work remains before a consensus can be reached.

The limitation inherent within the concept of role responsibility, however, is that it can be distinguished by its formulation as an omission. Because role responsibilities are functions of specific obligations, these are articulated as failures to fulfil a pre-existing duty. This has two implications for the current discussion. First, there is an unbridgeable conceptual gulf between the separate notions of role-omission responsibility and causal-commission responsibility (Liu 2016), such that negative outcomes can transpire *despite* the fulfilment of role-responsibilities. These are in effect two separate strands of the responsibility rope which have been braided together so often that these are no longer treated as distinct and discrete concepts. Their independence can be illustrated by situations where an undesirable outcome arises in spite of relevant role responsibilities being satisfactorily discharged: a situation of simultaneous responsibility and irresponsibility. The example with autonomous vehicles is where a vehicle was causally responsible for an accident despite the fact that the programmers, manufacturer and occupants had all fulfilled their due diligence obligations.

The second implication is that a role responsibility would be satisfied, or answered, by the fulfilment of the set of duties in question (depending on the precise formulation of the duty, an appropriate effort towards the satisfaction of the duty may be sufficient to offset role responsibility). In the context of autonomous vehicles, the occupant may have remained vigilant and fulfilled her (role-responsibility) duty to intervene in relation to an anticipated accident, but the collision may nevertheless ensue. In such an instance, the autonomous vehicle would have caused a negative consequence, despite the occupant both bearing and satisfying her duty to intervene. In such an instance, meaningful responsibility remains elusive.

Turning to capacity responsibility, there are signs that the occupants of autonomous vehicles may not be able to intervene in anticipated accidents. From a practical perspective, simulation experiments report that human drivers may need up to forty seconds to regain situational awareness, which is significantly longer than the responsive time-frame for typical accident scenarios (Lin 2015). Others suggest that the requisite capacities may not be forthcoming for reasons including age, physical or psychological impairment or disability, inebriation, and tiredness. Those advocating a duty to intervene also overlook the prospect for self-handicapping: to avoid responsibility ascription for accidents, those occupying autonomous vehicles may intentionally drink enough so that their blood alcohol concentration is over the legal limit, for example, or adopt clear strategies for distraction such as reading or sleeping that erode their capacity to intervene.[8] Because capacity responsibility is hinged on the characteristics and behaviour of individuals, there are clear avenues for volitional circumvention.

Further complications arise in relation to capacity responsibility where the autonomous vehicle replicates or exceeds human functioning on the road. In other words, where the human is the weakest link would it be 'ethical for an autonomous vehicle to return control to the human driver if the vehicle predicts that a collision with the potential for damage or injury in imminent?' (Gerdes and Thornton 2015). Alternatively, would it be ethical for an autonomous vehicle to wrest control from the human driver to avoid an accident or minimise the damage or injury of a collision? (Lin 2015). While these appear to be practical questions calibrated against the precise capabilities of both the human driver and the autonomous vehicle, the normative slant can be seen insofar as certain outcomes appear compelling.

Complicating the calculus even further is the prospect of an accident occurring where the human occupant takes over control from a properly functioning autonomous vehicle (Douma and Palodichuk 2012), resting on the presumption that the autonomous vehicle performs at least as well as a human driver. This fine balance suggests that the human occupant of an autonomous vehicle inhabits the proverbial space between a rock and a hard place. She may be compelled to take action to avert an anticipated disaster, but she will be penalised if it turns out that her intervention is inappropriate to the situation: a dilemma that in practice needs to be processed within the compact time frame necessary to avoid an accident.

Programmer and manufacturer responsibilities

Together, the configurations of role responsibility and capacity responsibility suggest that they will be poor candidates to substitute for the direct causal responsibility traditionally borne by the human driver in direct control of a vehicle.[9]

[8] Classically, responsibility doctrines trace intentional self-handicapping back to the initial intentional action of self-handicapping, thus treating it as recklessness with regard to risks of foreseeable damage. The clarity of this doctrine might be muddied by the omission nature of the occupant's role responsibility that is akin to oversight—the occupant only needs to be capable of intervening in an accident scenario and not of actively operating the vehicle. Moreover, where an accident occurs, the responsibility of the occupant will be limited to her failure to intervene, rather than for having caused the accident (thereby significantly curbing the scope of her liability).

[9] While negligence is the typical form of responsibility on the road at present, such conflating conceptual distinctions presented here

Moreover, it is precisely the concept of causal responsibility, when coupled with mechanisms of censure and sanction, that reflects the societal concern when the notion of responsibility is relied upon in debates that discuss emerging technology regulation.

With autonomous vehicles, however, issues of role and capacity forms of responsibility can be revisited from the perspective of both the programmer and the manufacturer.[10] It is uncontroversial to note that the discretional autonomy of the vehicle by definition undermines claims that either the programmer or the manufacturer enjoy direct control, thereby subverting the possibility for either to bear causal responsibility in lieu of the human driver. Again, and along similar lines to the occupant, the question of causal responsibility is instead displaced to obligation-omission forms of responsibility that are similarly curtailed by the content and character of the obligations. Distilled into different terms, the question of programmer or manufacturer responsibility concern issues of malicious intent or negligent oversight. While each bear clear role responsibilities, these are defined by duties which, if discharged, provide a satisfactory answer. Under this model, the programmer is expected to write code that leads to adequate performance, and the manufacturer is relied upon to create a product without significant defects. While the creation of both the code and the car set parameters that constrain the full range of autonomous vehicle behaviour to some extent, these amount at most to influences that fall short of determining the outcome. Under such circumstances, the role and capacity discussion above in relation to the occupant is mirrored for both the programmer and the manufacturer. Their role responsibility would be discharged provided that they had done their jobs well; at any rate, their capacities to influence the actual outcome may be curtailed by their remoteness to the final behaviour of the vehicle and the ensuing consequences.

It is curious to note, in this conjunction, the moral imperative advanced by some commentators to lessen the impediments to developing autonomous vehicles that includes structuring the assertion of responsibility in a manner that does not impede their development and introduction (Marchant and Lindor 2012; Hevelke and Nida-Rümelin 2015). This line of argumentation is grounded in the assumed social benefit of autonomous vehicles being safer in the long run, and that regulation should be designed purposively. Yet, because of the non-identification problem—that different lives will be spared and sacrificed—those making this claim need ground the specifics to allow the debate over the distribution of risks and vulnerabilities to progress (Lin 2013a). If we are not cautious, the asserted moral imperative not to impede autonomous vehicle development runs the risk of diluting the already ambiguous responsibility and regulatory structures we rely upon to govern these technologies. Overlooked in such calls are both the significant commercial incentives aligned to those who market autonomous vehicles as well as the corporate resources which might go some distance to overcoming obstacles imposed by robust regulation. Furthermore, it should be emphasised that corporations operate upon the legal imperative to generate profits for its shareholders, which predispose corporate activities to be measured along cost-benefit lines that has been documented to perversely maximise the bottom line, at least from society's perspective (Bakan 2005).

Analogies with autonomous weapons systems?

The responsibility question raised by autonomous vehicles is in many ways a more benign but complex iteration of the responsibility debate on the introduction of autonomous weapons systems (Jain 2016; Liu 2016). In that context, the responsibility lacuna left by the substitution of a causally responsible human agent for an artificial decision-making weapons system has led to proposals of overlapping omission responsibilities to bridge the gap (Schulzke 2013). With autonomous vehicles as with autonomous weapons systems, however, asserting occupant responsibility in situations either where she has discharged her duties, and thus satisfied the requirements of role responsibility, or where she is incapable, for whatever reason, of intervening, remains an unsolved problem.

The zealous pursuit to hold somebody, anybody, responsible amounts to more than moral defamation (as claimed by Hevelke and Nida-Rümelin 2015) but enters the perilous realm of imposing over-responsibilities: it is interesting to note in this context that impunity does not have a ready antonym, the nearest being that of the scapegoat (Liu 2015, 2016). Thus, proposing similar solutions revolving around overlapping role-responsibilities that highlight the obligations of designers, developers, manufacturers and maintainers will not only fail to bridge the responsibility gap, but may also introduce perilous new challenges of asserting

Footnote 9 (continued)

become increasingly problematic when additional actors are involved in producing the outcome.

[10] Both the programmer and the manufacturer are deployed here in their prototypical, singular, form. While the complexity involved in these processes suggest that these roles will be played by a multitude of persons and corporations, the point here is that significant responsibility issues remain even in this rudimentary caricature of reality. Furthermore, the consideration here excludes other identifiable parties who may bear at least a portion of the responsibility for an accident, such as the manufacturer of a component used in the autonomous system and the road designer where an intelligent road system is deployed to assist control of the autonomous vehicle (see Marchant and Lindor 2012).

accountability that is unhinged to control, foresight and due diligence.

Beyond responsibility and individualistic considerations?

Rather than obsessing over the responsibility void, if ready solutions are not forthcoming, it may be that we can circumvent it altogether by formulating policies that negate reliance upon the notion. In this sense, asking the question of responsibility railroads considerations towards individualistic notions dominated by concepts of individual autonomy, decision-making and discretion and directs attention towards accounting for one's actions in a field of options and choices.

Instead, a countervailing perspective might be offered that foregrounds the idea of (pre-) determinacy and that stipulated policies might bypass the entrenched responsibility debate. It could be argued that individualistic notions of responsibility are outmoded in an operational context where discretion in action can be replaced by formulae which dictate how the autonomous vehicle should behave in any given situation, and especially where such behaviour can be distributed broadly to a large number of units and can be updated regularly.

It is this very prospect for responsibility to become relegated to the side-lines through prescribed policies which introduces the peril of entrenched inequalities. Underlying the problematic question of responsibility remains the core of individuated discretion and autonomy that implies a variated and flexible approach to action. In a sense, the relational dimension of responsibility demands an account for justifications pertaining to a particular course of action in a given context. Yet, with widely replicated and prescribed policies, the variation of behaviours and outcomes have the potential to become severely constricted. Both the curtailed range of actions and the centralisation of discretion pose subtle structural amendments that threaten to erode equality and perpetuate injustice. The following sections examine the underexplored issues of centralisation and obfuscation associated with the prospect for private companies to determine behavioural policies for autonomous vehicles.

Before proceeding, however, a brief outline of the ethical dilemma at the core of this debate is necessary. This is the much-discussed 'trolley problem' thought experiment in which variations of a restrained choice is presented: undertake an action which will result in quantitatively less harm, or passively allow events to unfold which will result in greater objective harm (for an overview, see Edmonds 2013). Recent applications of the trolley problem to the introduction of autonomous vehicles have sparked a debate about how to appropriately programme them for precisely such exigencies (see for example, Lin 2013b; Doctorow 2015; Davis 2015; Bonnefon et al. 2015. On the problems in equating trolley problems and autonomous vehicle algorithms, see also; Nyholm and Smids 2016).

The foundational oversight of applying trolley problem ethics to the introduction of autonomous vehicles is, however, that such an approach overlooks the networked and coordinated effects which are likely to become implemented through programming processes. In other words, there is a gross mismatch between the individuated ethical framework expressed by the trolley problem and the emerging technological realities which have the potential at least to be in constant intercommunication and thereby act in coordinated and preordained fashion. Viewing the introduction of autonomous vehicles through the lens of trolley problem ethics thus neglects the issues of aggregation: that is, whether certain actions or inactions remain justified when the consequences are compounded or cumulative. A course of action or inaction that can be justified, or even virtuous, at the individual level may lose such grounding if implemented as a policy that is biased towards generating certain types of outcomes. Importantly, this oversight converges with the responsibility discussion above to create a significant blind spot because causal responsibility dimensions are considered through singular case-by-case analysis and because the focus of role responsibility is upon individual obligation.

We will leave aside the issues of coordination, prescription and aggregation that are overlooked by individuated trolley problem ethics for the moment in order to engage more directly in that debate, but will revisit these problems afterwards in the context of systematising inequalities through the operation of algorithms in autonomous vehicles.

Targeting

Invoking the trolley problem in the context of autonomous vehicles is often aligned with arguments for increasing road safety and decreasing accidents and fatalities. This has spurred discussion of crash optimisation (Lin 2015, see also, 2014a, b) which in turn transformed into questions of targeting individuals where crashes are unavoidable.[11] Empirical studies have begun to explore what essentially amount to targeting preferences (Bonnefon et al. 2015).

The question remains, however, whether results of rule-based and sequentially cumulative decisions can amount to

[11] This targeting discussion is written under the assumption that this is an expression of 'crash optimisation': as actions and inactions within accident scenarios that minimise the objective total harm. These comments do not encompass the possibility for autonomous vehicles to be directed towards placing the risk burden on certain groups or individuals which would invoke a different set of issues beyond the scope of this paper.

targeting individuals for harm per se, for three reasons. First, targeting implies intentional action against a pre-identifiable target, while trolley problem style accident scenarios involve coerced, time-restricted choices that curtail the scope of volition, and furthermore are not necessarily directed towards harming particular victims (although this can, troublingly, be the case as discussed below). Second, targeting is to a large extent decontextualised and preordained while trolley problem style accident scenarios set strong situational contexts which constrain decisions by placing parameters upon the range of possible actions. Third, autonomous vehicle programming does not determine particular future outcomes, but instead establish probabilistic courses of action in given contexts.

Taken together, there is significant conceptual distance between the intentional, unconstrained and directly causal act of targeting and the restrained time-pressured dilemma under which a decision in an accident scenario takes place. Adopting the targeting paradigm for autonomous vehicles also runs the risk of engraining the individualistic perspective because it isolates considerations to the impact of individual actions upon isolated victims. Furthermore, discussions of targeting foreground the considerations facing the agent in the relationship with the victims of the accident relationship. In doing so, adopting a targeting approach may become blind to cumulative impacts and systemic skews engendered by algorithmic policies aimed at accident optimisation.

Risk distribution: towards algorithmic risk allocation

The individuated and de-contextualised accident scenario, epitomised by the trolley-problem thought experiment, obscure three subtle yet pervasive factors that underlie the unease with which autonomous vehicle decision-making is discussed. First, stipulating preferences for action or inaction will be unavoidable, and a potential consequence will be that these pre-set values will remain consistent in practice over time. This tendency may then develop into structural biases where certain kinds of behaviour will be exhibited under certain conditions.[12] While vague, this leads to the second concern, that decision-making for autonomous vehicles will no longer be individualistically atomised and responsive, but instead will become centralised, coordinated and proactive. It is precisely because of the possibility for these effects that issues revolving around responsibility can become marginalised as accountability mechanisms are necessary only where possibilities exist for discretion and deviance. Compounding the centralisation and coordination tendencies in setting preferences for the behaviour of autonomous vehicles is the fact that these processes can be determined in a detached and prescriptive manner. This signals a significant departure from collective understandings of appropriate behaviour in traffic accident scenarios anchored in reactions within constrained scenarios. Unspoken assumptions may be grounded in the responsive paradigm which obfuscate the innovation that is introduced by autonomous vehicles which are able to replicate a particular set of prescribed preferences. Finally, when these effects are taken together, the preferences that are inscribed in the algorithms that govern the behaviour of autonomous vehicles have a harmonising or consolidating effect on the range of possible behaviours that may be exhibited. These exert pressures capable of curbing diversity and variation; yet because it is difficult to observe and resist such opaque policies they can also unwittingly entrench themselves.

The distribution of the risks of operating autonomous vehicles (see also, Goodall 2016) may fall short of the targeting paradigm specific individuals, yet the centralised and coordinated nature of algorithmic risk transforms this process to one of risk allocation rather than mere risk distribution. This is because allocation infers direction, edging towards intentionality, and emphasises the perspective of the patient-victim in the accident relationship by connoting the imposition of risk. Combining these influences together indicate the propensity for systematic harms and benefits to be generated from prescribing preferences.[13] The centralised and coordinated nature of algorithmic risk allocation thus has the potential to consistently structure both burdens and benefits.

[12] The suggestion here goes beyond the biased computer systems described by Friedman and Nissenbaum (1996): 'A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate' (at 332).This suggests that discrimination from technological artefacts bear human fingerprints, and responses such as value-sensitive design are able to go a long way towards meeting those concerns. Unlike the effects of biased computer systems discussed by Friedman and Nissenbaum, however, the structural form of discrimination countenanced here is difficult to fathom in a direct sense because the system neither "assigns" benefits or burdens, and it does not do so upon objectionable grounds. Rather, the system "optimises", offering results that look a lot like discrimination. Furthermore, these "discriminatory" effects are difficult to encapsulate because they are both cumulative and emergent, and not directly and causally connected in traditional manifestations of discrimination. These differences suggest that it would be difficult to effectively apply the solutions that have been designed against computer bias to the present problems posed by structural discrimination.

[13] At a basic level, systematic outcomes are endemic in policies that stipulate preferences because of the probabilistic likelihood that one type of result will occur more often as a result of the preference settings.

This hypothetical differs fundamentally from the situation that human drivers encounter today in accident scenarios because these are atomised, unconnected and contextual. Human decisions, in other words, are individual choices that need not be affected by how others have acted in similar situations nor must they have impact upon others facing similar situations in the future. It is this independent and volitional character of human decision-making that grounds the need for responsibility: with discretionary leeway comes the need to justify one's actions in a given situation. In this light, assertions of individual responsibility may indicate a lack of directing policy and thereby decrease the likelihood that structural inequalities are systematically disadvantaging specific groups.

Beyond responsibility? Structural inequalities and the non-identity problem

The final piece of the structural inequality puzzle is found in the non-identity problem. Essentially, this problem underscores the fact that different lives are saved and sacrificed, a fact that is hidden in assertions that there will be a net saving of lives (Lin 2013a). The non-identity problem infuses individuality into otherwise utilitarian considerations which accord equal weighting to faceless 'human units' presented in philosophical thought-experiments. While some variants of the trolley problem introduce certain personal characteristics, such as the variant that is designed to isolate factors of physical proximity and direct action by contemplating pushing a fat man off a bridge to stop the trolley (Edmonds 2013), the individuals in the thought experiment usually remain generic, indistinguishable and interchangeable human beings.[14] In effect, trolley problems focus upon quantitative considerations; the non-identity problem factors in qualitative considerations into crash optimisation calculations.

The prospect of recognising individual characteristics, however threaten to open the flood gates concerning which features can and will be recognised and what weighting these should be accorded. To some extent this has already taken place where gestures were made towards a dilemma between diverting an autonomous vehicle to hit either a motorcyclist wearing a helmet and another who does not (Lin 2015). As Patrick Lin observes, adherence to a principle aimed at minimising harm would result in the autonomous vehicle hitting the helmeted motorcyclist because her odds of surviving the accident is higher, yet such an outcome places burdens upon exactly those who adopted prudential measures to minimise their exposure to risk. This example neatly encapsulates the structural concerns of centralised preferences and risk allocation discussed above, because it illustrates the potential for certain groups to consistently bear a greater burden because of how their characteristics factor into the utilitarian crash optimisation framework.

The non-identity problem in the context of autonomous vehicles is not so much that individual lives may either be saved or lost as a direct consequence of implementing autonomous vehicles, but rather that the same centralised and coordinated rules govern their decisions. The possibility for identical responses that are governed by the same rule-structure creates a systemic and collective dimension whereby the generated outcomes will be systematically skewed. The crucial differentiator, and the source of dystopian concerns therefore, should be the removal of individual or independent variation from range of available responses. Accumulating these skews together results in systematic bias that can ground allegations of pernicious, and pervasive, discrimination.

Drawing these threads together, the introduction of autonomous vehicles raises the prospect for systemic, rule-based disadvantages which may constitute a particularly odious kind of discrimination. Trolley problem ethics are rooted in isolated and decontextualized thought-experiments that overlook real-world ramifications that arise from aggregating preferences discerned from individualised, action-based perspectives. The trolley problem's myopic focus upon the causes and consequences of a single scenario overlooks the possibility that the results will yield patterns and trends that cannot be predicted by studying the ethics in isolation. As a result of this ethical lens, accumulated effects cannot be recognised, thereby potentially sterilising what would otherwise be legal wrongs in conventional driving accident scenarios to mere harm and damage. Such structural pressures effectively immunise the system against the prospect for legal liability.

A rough analogy can be drawn with the concept of emergence to illustrate these effects, whereby complex and unforeseen consequences can arise through accretion from mere adherence to simple rules (Gleick 1997). The focus of trolley problem ethics remains at the singular occurrence and fails to account for cumulative effects. Adopting this perspective shows the difficulties that lie ahead for those seeking to challenge potentially discriminatory outcomes. Algorithmic preferences can be readily defended on the basis that each resultant decision, considered in isolation, is justifiable. Furthermore, that the overall policies are designed to minimise overall harm reinforce this defence, such that any discriminatory effect can only be tangential because of a lack of intention (and possibly also unforeseeable) so that cumulative impacts remain unrecognised.

[14] The choice of the fat man in the thought-experiment, for subliminal or pragmatic reasons, may hint towards discriminatory tendencies, when considered in light of research suggesting that systematic discriminatory biases disadvantage obese individuals (Puhl and Brownell 2001).

Buffering this burgeoning defence is the likelihood that empirical evidence will reflect the individualistic agent-centred perspectives espoused crash optimisation frameworks. For example, early studies show signs of such slant through the focus upon existing human preferences (Bonnefon et al. 2015), thereby obscuring the opinion of those who may have to bear the consequences without enjoying the benefits of autonomous vehicle use. As we are currently at the early stages of development, however, it is not too late to implement complementary research that collates the preferences of members of the public who are not involved with car purchases or car use in order to rectify this bias.[15] On this note, it is also worth registering the possibility that cultural and regional variations will exist which also need to be explored and documented, considering the globalised nature of the automotive industry. Indeed, it may also be the case that such preferences need to be accounted for by manufactures for specific export markets in order to remain consistent with the priorities of the importing societies.

Speculative inequalities? The immunity device thought experiment

Up until this point, we have been discussing only the prospect for implicit, perhaps even unintentional, structural inequalities that result from crash optimisation preferences designed to minimise objective levels of harm borne at a collective societal level. While speculative, a much more pernicious form of structural discrimination could, however, be developed that entails an elaboration of precisely the type of utilitarian calculus aimed at maximising collective happiness.

Taking departure in the above discussion on zero-sum structural risk displacement that accounts for the non-identity problem, what if the decision as to who the autonomous vehicle were to hit included considerations of an individual's innate talent, cultured ability, or latent potential? In other words, might the obverse approach to crash-optimisation impulses suggest spreading a protective aegis over individuals offering special contributions to society? Thus, an alternative approach to crash-optimisation would be to prevent the largest qualitative losses rather than seeking to incrementally minimise harms: a strategy preventing the worst outcome rather than aiming for the least bad. Such an approach could justify safeguarding individuals with a rare skillset, as Joseph Louis Lagrange quipped after Antoine-Laurent de Lavoisier, widely acknowledged as the father of modern chemistry, was guillotined, "It took them only an instant to cut off this head, and one hundred years might not suffice to reproduce its like". Would not society have an interest, albeit a potentially perverse one, in programming its autonomous vehicles to preserve the lives of its scientific and cultural elite in pursuit of the public benefit? After all, establishing protective preferences for certain categories of persons is implicit in the emotional appeal when school buses are inserted into the trolley dilemma. The presence of children invokes our intuitive responses that the lives of the young and innocent are intrinsically worth protecting. And, if any doubt remains, it appears quite clear that our political leaders and diplomatic representatives enjoy high levels of personal protection not afforded to the ordinary citizen.

Following this logical track, it would not be unreasonable for the manufacturers of autonomous vehicles to issue what I would call here an 'immunity device': the bearer of such a communicative device would become immune to collisions with autonomous vehicles. Such an amulet would protect its owner in situations where an autonomous vehicle finds itself careening towards him or her, and would have the effect of deflecting the car away from that individual and thereby force the car to engage in a new trolley problem style dilemma elsewhere. If the justifications for the bearer of the immunity device are sufficiently strong, and their numbers suitably restricted, this might be a practical response to the new quandaries introduced by autonomous vehicles.

But this appears to be the thin end of a very large wedge. The scenario introduced above is binary and absolute: the immunity device offers complete protection. Yet, if the enabling technology becomes available it is difficult to see how it would not expand due to market demand. Indeed, because such a device would prevent a victim from suffering an accident, rather than compensating a victim in the aftermath of an accident, it could conceivably become a preventative alternative to purchasing insurance policies. Developing the immunity device would introduce the ability for cars to communicate with their potential victims' devices in the event of an "accident" (now a metaphorical term since eventualities are calculated), and a range of pressures would push these capabilities into desperate or greedy hands. For obvious reasons, a widespread system of immunity devices would be impracticable and self-defeating, so hierarchical nuances would have to be introduced: essentially a ranking system that eases the trolley problem calculus for any unfortunate autonomous vehicle. The autonomous vehicle will essentially be playing a game of trump cards in the event of a trolley problem: whoever bears the highest status aversion device would be spared at the expense of those who possess lower status ones. This is a very uncomfortable outcome for what initially appeared to be a satisfactory configuration of

[15] Here, the methodologies deployed by value-sensitive design and especially those pertaining to the identification of direct and indirect stakeholders as well as the benefits and harms they might incur goes some way to addressing these issues (Friedman et al. 2006).

benefits and burdens imposed by the autonomous vehicle of the future.

This then asks the question as to how to allocate the particular status an individual should have and therefore the concomitant level of risk that she should bear in relation to autonomous vehicles. But all its variants run against the bold proclamation in the Universal Declaration of Human Rights that "[a]ll human beings are born free and equal in dignity and rights". Were a meritocratic system to be implemented, we might end up with the situation alluded to above, but perhaps more likely the distribution will be made economically especially given the commercial opportunities for profit-making that such a system introduces. A particularly interesting scenario would be a closed market where a zero-sum game is implemented where individuals seeking high status would be required to purchase the differential from other members of society, thus making risk displacement through material wealth explicit. Furthermore, recklessness incentives can be introduced where those with the resources are able to purchase their way out of harms' way.[16] It might even be possible to allocate risk as a form of punishment: those who have acted to the detriment of society might be forced to bear a burden of increased risk as a form of compensation. Yet, however the final structure may be, its hierarchical effects are evident, as are the tendencies towards heuristics, categorisation and entrenchment.

Leaving aside the practical details associated with the future market for autonomous vehicles, such as whether manufacturers will offer such systems, whether they would be preferential towards their own clientele and whether they would coordinate and centralise such a scheme, the ramification of such a system is that risk of injury shifts towards prevention rather than cure. Rather than enforce accident insurance upon drivers to cover for their future faults, this system would place a large share of the burden upon the ordinary citizen to avoid or minimise the likelihood of injury arising from autonomous vehicles. This fundamentally shifts the burden of risk from the agent-driver to the patient-pedestrian. In pre-determining the allocation of risks and costs in advance of any accidents, the remaining fault that must be covered by the driver (in reality the manufacturer of the autonomous vehicle) would be for departures from the course of action that has been promised in advance, and not the actual outcome caused by the autonomous vehicle. This would substantially narrow the range of liability for the agent-driver—the party who introduces the risk in the first place (see also, de Sio 2017).

To make matters even more complex is the fact that autonomous vehicles are not developed from any semblance of a neutral situation, but rather by corporate entities seeking to make material profits from their efforts. Even if such motives do not taint the product directly, preferences that increase profits are likely to become embedded in the decisional architecture of the autonomous vehicle that a company produces. It is neither uncommon nor unreasonable for a car manufacturer today to emphasise the safety features of its models that protect its customers and passengers. A driver who takes the decision to hurt others in order to protect his or her own interests similarly acts within the boundaries of social acceptability. Yet, when these two characteristics are united in the autonomous vehicle, however, the perceptions may shift. The vehicle subtly transitions from an artefact towards being an agent: from something that is inanimate and subject to human control to something that observes, orients, decides and acts (Suchman and Weber 2016). Crossing this line implicates programming that may systematically elevate the safety of its customers and occupants over all others, not least because an autonomous vehicle would not be able to convincingly make account of its decision-making processes. In an important sense the outcome, where the prospect of harm is unavoidable, has been automated and pre-determined: because discretionary autonomy is introduced without clear channels of responsibility, this has the effect of immunising the effects from responsibility.

The aim with the advent of autonomous vehicles should not merely be to integrate them into contemporary society and simply continue with business as usual, but rather to consider how such fantastic technologies may improve the daily lives of both those who own and access them as well as those who stand to be harmed through their introduction. Taking a glimpse into possible, dystopian, futures should provide us with a reality check that fuels the debate on how we want autonomous vehicles to behave, from all perspectives and standpoints. To abdicate such inclusive processes risks the introduction of a technology that favours the few and which exacerbates inequalities.

Maintaining injustices: two levels of responsibility gaps?

To draw these analyses together, there may be two different levels at which autonomous vehicles can introduce injustices for the victims of accidents which these will cause. While difficult to articulate precisely, they may be loosely framed as responsibility gaps. The first level involves the individualised division of responsibility in both circumstantial (the pragmatic situations where autonomous vehicles will be used) and conceptual contexts (primarily arising from the conflation between role and outcome responsibilities) (Liu 2016). As with the autonomous weapons systems discussion, while circumstantial responsibility questions are capable of

[16] I owe this idea to John Danaher, in conversation on Algocracy, https://algocracy.wordpress.com/.

being resolved through improved implementation of autonomous vehicles, the conceptual responsibility question will persist until the theoretical gaps between the separate notions of responsibility are harmonised (Liu 2016). Exacerbating responsibility gaps at this level are the uncertainties that hang over the distribution of liability between the users, programmers, and manufacturers: impunity is likely to arise in instances where responsibility is a possibility, but which is neither necessary nor certain (Liu 2015). Insofar as the responsibility questions at this level are not adequately addressed, our regulatory framework will remain misaligned with evolving practical realities.

Beyond the question of impunity for direct harms caused by autonomous vehicles are the structural pressures and systemic biases which widespread autonomous vehicle operation might precipitate. The responsibility concerns at this second level have hitherto remained imperceptible, raising the question whether conventional concepts of responsibility can even be meaningfully deployed in such diffused contexts. This would suggest that the probabilistic burdens borne by different groups will remain unrecognised and unaddressed. Insofar these effects are not raised as questions of responsibility, the systematic harm and damage caused by autonomous vehicle implementation will therefore remain unrecognised as injuries, understood in the above discussion meaning legal wrongs.

Delving deeper, there may be two sub-level issues conflated within this second level of injustices: responsibility for the direct impact of autonomous vehicle usage, and responsibility for the discriminatory pressures exerted by the system of autonomous vehicles operating as a whole. The former involve assertions of liability for what appear to amount as redistributions or heightened imposition of risks: do reallocations of risk imply or mandate reapportioning responsibilities? The latter concerns the application of responsibility to processes that might look like discrimination, but which do not accord benefits or burdens upon impermissible grounds which define classical discrimination: can responsibility be applied to account for processes which mimic discriminatory outcomes?

Interposing the prospect for structural discrimination arising from optimisation processes into questions of the responsibility gap underscores the under-preparedness of our juridical system in relation to autonomous vehicle usage. Regulatory discussions typically engage with the first level responsibility gap, seeking to apportion responsibility between the proximate human beings (Schulzke 2013). Thus, second-level responsibility questions relating to both the outcome and the processes of systematically redistributing risk in pursuit of crash optimisation creates an overlooked responsibility gap. Insofar as increased burdens and risks are imposed upon certain groups without the legal system developing concomitant concepts and processes to challenge these effects, these responsibility gaps at the systemic level will be the foundation for subtle forms of injustice that autonomous vehicles will unwittingly introduce.

Concluding thoughts

Possible shifts to ameliorate some of these challenges include refocussing responsibility doctrines away from competence, capacity, control and causation and towards the beneficiaries of the activities instead. Thus, an risk-based accountability duty can be forged between on the one hand the occupiers of the autonomous vehicles who benefit from usage, and the programmers and manufacturers who profit commercially, and on the other hand those third parties to whom are allocated greater burdens as a result of these activities. Instead of hinging upon autonomy and discretion, such an accountability duty would instead be connected to the introduction or redistribution of risks posed to other parties. While such an accountability duty looks very much like causal responsibility, the conceptual differences are two-fold. As a duty it is a procedural relationship which circumvents the direct causal requirement; its basis upon the imposition of risk may allow it to navigate the second level of responsibility gap involving structural or systemic displacements of risk. This should have the effect of shifting the onus to the agents to justify their decisions and account for the consequences ensuing from their activities. In doing so, third parties who are poised to be burdened with additional risks and imposed with subtle new vulnerabilities will be foregrounded in the responsibility calculus concerning autonomous vehicles and have more direct access to those who benefit from their operation.

To do this, the question of responsibility should also shift from the traditional 'who-dunnit' approach popular in crime fictions to focus instead on the broader structural issues captured by asking why it happened. Doing so encourages a re-evaluation of whether traditional legal notions of responsibility, both criminal and tortious, should be extended to the autonomous vehicle, and to define the limitations of these legal doctrines. As we have seen, autonomous vehicles replace the human driver who is both in direct control and bear the full gamut of responsibilities. As ascriptions of responsibility become increasingly tenuous and artificial, questions arise as to whether the doctrine remains tenable in face of contemporary challenges.

The purpose, then, is to question whether existing rules which govern human driven traffic should be unthinkingly applied to frame the regulation of autonomous vehicles. It may be, for example, that tort law can grapple and govern autonomous vehicles (Graham 2012), but the criminal law with its focus on individual mental states can usefully be excluded. In order to assess the continuing relevance of

responsibility concepts in relation to autonomous vehicles, the overarching purpose for their deployment needs to be articulated. This must then be set against eroding the plausibility or legitimacy of contriving or contorting those concepts to adapt them to autonomous vehicles. As a result, the natural malleability of the concepts must be respected, as to stretch them to breaking point would be a pyrrhic victory that ultimately undermines legal concepts more broadly.

Moving to crash optimisation impulses, bright regulatory lines may be necessary to curb some of the excesses envisaged here, and to militate against the possibility of indirect systematic discrimination through algorithmic policies. To establish impermissible types of development requires broad public engagement well in advance of technological maturation, and ideally at an earlier stage such that regulation can influence design and implementation. The thought-experiment outlined in this article that underscores the possibility for active (and perhaps not incidentally, profitable) forms of inequality may arise from an alignment of commercial and individual incentives that can go unchecked, or at least remain under-regulated. More prosaically, soliciting a broader range of opinion could lessen the prospect for regulatory blind-spots and increase the confidence of both those driving the technology forward as well as potential users and victims of autonomous vehicles.

References

Bakan, J. (2005). *The corporation: The pathological pursuit of profit and power*. London: Constable & Robinson.

Bonnefon, J. F., & Sharif, A., Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *arXiv*. Retrieved October 12, 2015, from http://arxiv.org/abs/1510.03346.

Coffee, J. C. (1981). 'No soul to damn: no body to kick': An unscandalized inquiry into the problem of corporate punishment. *Michigan Law Review, 79*(3), 386–459.

Cummings, M. L., Mastracchio, C., Thornburg, K. M., & Mkrtchyan, A. (2013). Boredom and distraction in multiple unmanned vehicle supervisory control. *Interacting with Computers, 25*(1), 34–47.

Davis, L. C. (2015). Would you pull the trolley switch? Does it matter? *The Atlantic*. Retrieved October 9, 2015, from http://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/.

de Sio, F. S. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice, 20*(2), 411–429.

Doctorow, C. (2015). The problem with self-driving cars: Who controls the code? The Guardian. Retrieved December 23, 2015, from http://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code.

Douma, F., & Palodichuk, S. A. (2012). Criminal liability issues created by autonomous vehicles. *Santa Clara Law Review 52*(4), 1157–1169.

Edmonds, D. (2013). *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*. Princeton: Princeton University Press.

Friedman, B., Kahn, P. H., & Borning, A. (2006). Value sensitive design and information systems. In P. Zhang & D. F. Galletta (Eds.), *Human–computer interaction and management information systems: Foundations* (pp. 348–372). London: Taylor and Francis.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems 14*(3), 330–347.

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren* (pp. 87–102). Berlin: Springer.

Gleick, J. (1997). *Chaos: Making a new science*. New York: Vintage.

Goodall, N. J. (2014). Machine Ethics and Automated Vehicles. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 93–102). Berlin: Springer.

Goodall, N. J. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence 30*(8), 810–821.

Graham, K. (2012). Of frightened horses and autonomous vehicles: Tort law and its assimilation of innovations. *Santa Clara Law Review, 52*(4), 1241.

Hart, H. L. A. (2008). Punishment and responsibility: Essays in the philosophy of law. Oxford: Oxford University Press.

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

Heyns, C. (2016). Autonomous weapons systems: Living a dignified life and dying a dignified death. In N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kress (Eds.), *Autonomous weapons systems* (pp. 3–20). Cambridge: Cambridge University Press.

Jain, N. (2016). Autonomous weapons systems: New frameworks for individual responsibility. In N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kress (Eds.), *Autonomous weapons systems—Law, ethics policy* (pp. 303–324). Cambridge: Cambridge University Press.

Lin, P. (2013a). The ethics of saving lives with autonomous cars is far murkier than you think. *WIRED*. Retrieved July 30, 2013, from http://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/.

Lin, P. (2013b). The ethics of autonomous cars. *The Atlantic*. Retrieved October 8, 2013, from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

Lin, P. (2014a). The robot car of tomorrow may just be programmed to hit you. *WIRED*. Retrieved May 6, 2014, from http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/.

Lin, P. (2014b). Here's a terrible idea: Robot cars with adjustable ethics settings. *WIRED*. Retrieved August 18, 2014, from http://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/.

Lin, P. (2015). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren* (pp. 69–85). Berlin: Springer.

Liu, H.-Y. (2015). *Law's impunity: Responsibility and the modern private military company*. Oxford: Hart Publishing.

Liu, H.-Y. (2016). Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems. In N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kress (Eds.), *Autonomous weapons systems—Law, ethics policy* (pp. 325–344). Cambridge: Cambridge University Press.

MacCormick, N. (1995). Argumentation and interpretation in law. *Argumentation, 9*(3), 467–480.

Marchant, G. E., & Lindor, R. A. (2012). The coming collision between autonomous vehicles and the liability system. *Santa Clara Law Review, 52*(4), 1321.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

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

Puhl, R., & Brownell, K. D. (2001). Bias, discrimination, and obesity. *Obesity Research, 9*(12), 788–805.

Robbins, M. (2016). Statistically, self-driving cars are about to kill someone. What happens next? *The Guardian*. Retrieved June 14, 2016, from https://www.theguardian.com/science/2016/jun/14/statistically-self-driving-cars-are-about-to-kill-someone-what-happens-next.

Salomon v A Salomon. (1897). [1897] AC 22. U.K. House of Lords.

Santa Clara County v Southern Pacific Railroad. (1886). 118 U.S. 394. U.S. Supreme Court.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*(2), 203–219.

Seck, S. L. (2011). Collective responsibility and transnational corporate conduct. In T. Isaacs & R. Vernon (Eds.), *Accountability for collective wrongdoing* (pp. 140–168). Cambridge: Cambridge University Press.

Suchman, L., & Weber, J. (2016). Human–machine autonomies. In N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kress (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 75–102). Cambridge: Cambridge University Press.

Veitch, S. (2007). *Law and irresponsibility: On the legitimation of human suffering*. Oxford: Routledge-Cavendish.