Safety of Human–Artificial Intelligence Systems: Applying Safety Science to Analyze Loopholes in Interactions between Human Organizations, Artificial Intelligence, and Individual People

Stephen Fox 1,* and Juan G. Victores 2

1 VTT Technical Research Centre of Finland, FI-02150 Espoo, Finland
2 RoboticsLab, University Carlos III of Madrid, 28911 Leganés, Spain; jcgvicto@ing.uc3m.es
* Correspondence: stephen.fox@vtt.fi

Abstract: Loopholes involve misalignments between rules about what should be done and what is actually done in practice. The focus of this paper is loopholes in interactions between human organizations’ implementations of task-specific artificial intelligence and individual people. The importance of identifying and addressing loopholes is recognized in safety science and in applications of AI. Here, an examination is provided of loophole sources in interactions between human organizations and individual people. Then, it is explained how the introduction of task-specific AI applications can introduce new sources of loopholes. Next, an analytical framework, which is well-established in safety science, is applied to analyses of loopholes in interactions between human organizations, artificial intelligence, and individual people. The example used in the analysis is human–artificial intelligence systems in gig economy delivery driving work.

Keywords: algebraic machine learning; driving; gig economy; human–artificial intelligence (HAI) systems; loopholes; narrow AI; quality management systems; Swiss Cheese Model; Theory of Active and Latent Failures

1. Introduction

The safety of human interaction with artificial intelligence (AI) is a topic of wide-spread debate [1,2]. While there is ongoing scholastic debate and scientific research concerning alignment between humans and AI [3,4], many individual people are already working in alignment with human organizations’ AI implementations as they become increasingly prevalent [5,6].

Here, an established safety science method is related to the safety of human–artificial intelligence (HAI) systems. In particular, the importance of identifying and addressing loopholes is recognized in safety science [7] and in applications of artificial intelligence (AI) [8]. Loopholes involve misalignments between rules about what should be done and what is actually done in practice [9]. Loopholes can emerge when rules have insufficient breadth and/or detail, and may be exploited when there is insufficient observation and/or enforcement of rules. The fundamental means to exploit loopholes has been human ingenuity [10]. Now, there are concerns that so-called bad actors could find ways around rules about what should be done by AI [11]. However, motivation to exploit loopholes can arise from common pressures experienced by many people such as organizational pressure [12] and/or time pressure [13].

In the remaining four sections of this paper, an analysis is provided of loopholes in interactions between human organizations, artificial intelligence, and individual people. In Section 2, established sources of loopholes are examined. In Section 3, new sources of loopholes from AI are explained. In Section 4, the Theory of Active and Latent Failures [14],
which is often referred to as the Swiss Cheese Model (SCM) [15], is related to human-artificial intelligence (HAI) systems. In conclusion, practical implications are discussed and directions for further research are proposed in Section 5. In the paper, the SCM is related to HAI for gig economy driving of delivery vehicles. This is the example because legal regulation of such work has loopholes that can be advantageous for human organizations but disadvantageous for individual people [16,17]. Also, this is the example because increasing automation technology is being introduced for vehicle driving: for example, Automated Lane-Keeping Systems [18] and Intelligent Speed Assistance Technologies that provide drivers with in-vehicle warnings if they are driving above the speed limit [19].

2. Sources of Loopholes between Human Organizations and Individual People

First, in this section, loopholes from changing culture-bound social choices are described. Next, loopholes due to prioritizing personal relationships above laws and regulations are discussed. Then, the potential for technologies to amplify loopholes is described. Subsequently, the potential for loopholes in HAI systems is discussed.

2.1. Emerging Loopholes from Changing Culture-Bound Social Choices

Laws, regulations, and related standard documents are seldom fixed in perpetuity. For example, laws related to the driving of motor vehicles have changed over time. At the beginning of the 20th century, there were few motor vehicles, which were driven in environments that did not have speed signs and traffic lights. Subsequently, the number of motor vehicles increased and the driving environment was changed with the introduction of signs, lights, etc., together with legal regulations to enforce them [20]. Legal regulations change over time through political processes [21,22] as culture-bound social choices change [23,24]. For example, driving motor vehicles while under the influence of alcohol was not considered a serious crime at the beginning of the 20th century [20]. Subsequently, there have been increasing criminal penalties for doing so [25]. However, as well as being increased over time, legal regulations can be reduced over time. For example, recent changes to some road environments have involved removing some warning signs and some traffic lights in order to improve the efficiency and safety of road traffic [26–29]. Thus, legal regulations may not be in place for long enough for all possible loopholes to become apparent and to be closed. Rather, loopholes can emerge as slow legislative processes struggle to address changing culture-bound social choices in sufficient breadth and detail in the context of rapid technological innovation. Nonetheless, rules in laws, regulations, and standards can provide some of the bases for human organizations’ quality management systems and corporate social responsibility statements [30]. These can guide human organizations’ implementations of task-specific AI, which have a narrow scope of implementation [31]. For example, standards are beginning to be introduced for intelligent transport systems [32]. Unlike possible future artificial general intelligence (AGI), narrow task-specific AI implementations may perform only one specific application. Unlike AGI, narrow task-specific AI already exists and is being introduced in a wide variety of everyday applications, such as food-delivery platforms [5].

2.2. Exploiting Loopholes to Conform with Group Affiliations

Within moral foundations theory [33], it is posited that the morality of loyalty, authority, and sanctity can be related to interactions with groups. For example, it can be argued that loyalty to groups is related to our long tribal history. Also, it can be argued that deference to authority is related to our long history of hierarchical social interactions. In addition, it can be argued that the sanctity of groups has been shaped by notions of striving to live in a more noble way. However, people who identify strongly with a group can conform to less than noble organizational pressure to take unethical action [34,35]. Here, it is important to note that alongside organizations’ quality management manuals and corporate social responsibility statements, which need to comply with legal regulations and related standards, there can be groups who may not manage organizations in accordance
with their organizations’ documented procedures [36]. In such cases, rather than corporate social responsibility, there can be corporate social irresponsibility [37]. Consequently, an individual person may have the dilemma of whether to align with legal regulations or whether to align with groups in the workplace who can have more immediate influence over that person [38,39]. Initially, a person may prefer to align with legal regulations. However, stress can change people’s deepest preferences [40]. Apropos, it is important to note that because human perception has evolved to enable human survival in changing environments, we can prefer to perceive the environment not exactly as it is, but rather as its usefulness for our survival to perceive it [41]. For example, one person can have several internal behavioral models to select from depending on situation-specific social interactions [42], and people can make a subjective estimation of utility maximization that could lead them to favor unethical action [43]. Thus, potential for the exploitation of loopholes predates the introduction of artificial intelligence into interactions between human organizations and individual people.

2.3. Loopholes Amplified by Technologies

Technologies can amplify human behavior [44–47]. Within this framing, “people have intent and capacity, while technology is merely a tool that multiplies human capacity in the direction of human intent” [44,45]. Today, many people can be in “technology-amplified environments” throughout their lives [46,47]. Fundamentally, human intent is to try to survive. Accordingly, it can be expected that many individual people and human organizations will deploy technologies, including AI, to increase their capacity to survive [48]. This can involve human organizations exploiting loopholes even if this is done at the expense of many individual people [49,50]. Overall, the introduction of AI can amplify the challenge discussed in Section 2.1, the challenge of slow legislative processes trying to keep up with changing culture-bound social choices in the context of rapid technological innovation, and the challenge discussed in Section 2.2, the challenge of people often attributing more importance to group affiliations than to legal regulations. Apropos, it has been reported that new legislation for the regulation of AI developments and applications has many loopholes, which could be exploited by human organizations at the expense of individual people [51,52].

2.4. Loopholes in HAI Systems

It has been argued that loopholes stem from the need to accommodate conflicting interests [53,54]. There can be large inequalities between the computer science resources of large organizations and individual people. Nonetheless, it can be possible for individual people who do not have any computer science skills to take action against what they may perceive to be unfair deployments of AI by large organizations. For example, vehicle drivers, who must make deliveries on time in order to get paid enough to survive, can seek to evade computer vision in speed cameras by covering speed camera lenses [55]. Often, there may be potential for exploitation of loopholes because of insufficient observation and/or enforcement of rules. For example, it may never be possible to put speed cameras everywhere that vehicles are driven, and it may never be possible to have every speed camera guarded at all times to prevent their lenses being covered. Even if it were possible, people could take some alternative action such as covering the number plates of their vehicles with reflective tape to prevent them being read by computer vision. This example illustrates there may never be technological solutions that eliminate the potential for cycles of loophole emergence and exploitation as organizations and people with different interests engage in tit-for-tat exchanges. Rather, technologies can be combined and used in new ways that were never envisaged by those who developed them [56], which could enable the circumvention of legal regulations and associated standards [51,52].
3. Sources of Loopholes from AI

In this section, sources of loopholes are explained from three types of AI: deterministic AI, probabilistic AI, and hybrid AI. For all types of AI, the recent introduction of powerful AI, such as Large Language Models, which is publicly available for use and can be applied quickly by non-computer scientists, introduces the increased potential for human organizations to automate organizational procedures that are consistent with their official organizational documentation, such as quality management system (QMS) manuals. However, it also introduces potential for human organizations to automate organizational procedures that are not consistent with their official documentation. As it has taken years for such misuse of computer science to be identified in the past, for example in the Dieselgate scandal [57], the widespread use of new more powerful publicly available AI could introduce even more widespread deliberate development of loopholes enabled by AI [58].

3.1. Deterministic AI

Deterministic implementations can come within the scope of “hand-crafted” [59] expert systems, which depend upon expert human knowledge to identify and program each specific case that may appear within the scope of their operation. Rule-based AI, which may be referred to as Good Old-Fashioned AI (GOFAI), can be deployed within expert systems and can achieve better results than people when used in constrained worlds such as the game of chess. However, amidst the complexity of real-world dynamics, GOFAI within expert systems is of limited usefulness. This is because of the difficulty of identifying each case that may appear and programming each possible case as hard-coded behavior in rule-based implementations. In practice, this can result in a GOFAI expert system not being able to offer guidance to people when they most need it—in unusual situations [60]. Thus, rather than a person receiving a statement from AI about what should be done, a person may receive a statement from the AI that it does not know what should be done.

3.2. Probabilistic AI

By contrast, machine learning is a non-deterministic data-driven field, where probabilistic approaches are used to cope with real-world uncertainties. Compared to GOFAI, machine learning (ML) has provided some advances. These include image classification via deep neural networks for supervised learning and better-than-human results through deep neural networks for reinforcement learning in more complicated constrained worlds such as the game of Go and Atari videogames. Yet, ML can have limited potential to transfer what it has learned from training data to situations that differ from the training data. Moreover, ML implementations rely on functions, which result in goals needing to be expressed as a single numerical value that represents how well it is performing. However, unlike GOFAI expert systems, rather than offering no information to people in unusual situations, they may offer information that seems credible but is actually not useful because it has been learned from irrelevant training data and/or has an irrelevant reward function [61]. This is an aspect of ML from which loopholes may emerge in ML. There are two different aspects of training data that can cause machine learning to provide misaligned outputs [62]. The first aspect is bias, where high proportions of data indicating one fact can lead to the model believing that fact is true. In other words, the model is based on popularity rather than actual fact-checking [63]. The second aspect is the potential for distribution shifts, which can occur when the training data is not representative or does not cover scenarios that the machine learning model may encounter during real-world use cases [64]. Some machine learning models may present a certain degree of robustness to distribution shifts due to generalization capabilities. However, large deviations may forcibly imply an output that leads to misalignment [65]. For example, a robot may learn to advance by training only on a flat surface; but in a real-world scenario, it may encounter a cliff, and decide to stride off into the abyss. Furthermore, it has been argued that machine learning datasets are often used, shared, and reused without sufficient visibility about the processes involved in their creation [66]. Such problems can contribute to the characteristics and origins of biases in
machine learning datasets being difficult to prevent and to correct. Also, they can make it difficult to identify potential for distribution shifts. Hence, possible loopholes can be very difficult to define in advance.

3.3. Hybrid AI

Deterministic and probabilistic approaches can be used together in hybrid approaches. For example, ML outputs can be thresholded by expert system rules [67]. Moreover, in all ML implementations, expert knowledge is implicitly incorporated at the initial design stage. One new hybrid approach is algebraic machine learning (AML) [68]. This AML allows explicit embedding of rules directly (deterministic) while it simultaneously uses multiple data inputs (probabilistic). Furthermore, AML can allow uncertainties to be embedded explicitly as part of a specific application, but again this depends upon expert human knowledge [69]. It cannot be assumed that hybrid implementations will overcome the limitations of their deterministic and probabilistic predecessors. Rather, hybrid implementations could introduce loopholes from statements from the AI that it does not know what should be done or from statements that seem credible but are actually not useful. For example, loopholes could arise from insufficient information or incorrect information about the relevance of rules to a situation. However, AML is an example of a hybrid ML that has some functionality that goes beyond combining deterministic ML and probabilistic ML. The core AML algorithms check that the human-defined constraints are maintained in the subsequent relationships between inputs and outputs that are learned during training [69]. Human-defined rules are enacted through explicit instances of the “≤” inclusion relationship operator, which is the same operator used to aggregate large quantities of data to the model and can be used at any time. The basic building blocks are AML “constants”. They are called constants because they are always present in an AML model, i.e., they are present in the model constantly. Constants can be expressed in natural language so they can be understandable for model users. At the same time, constants are the primitives used by the algebra and can be combined within the model via the via the “⊙” merge operator to be sets of constants called “atoms”. Training algorithms lead to the “freest model” of an algebra, which accurately describes the system through the minimum set of rules (through the reduction of “cardinality”). This is achieved by the trace-invariant “crossing” of atoms derived from the original constants. The “full crossing” training algorithm involves internally checking all the combinations of atoms of a training batch to establish an algebra, while the “sparse crossing” algorithm includes heuristics for the efficiency of these operations without the loss of generality.

4. Applying Safety Science to HAI Systems

4.1. The Swiss Cheese Model

The Theory of Active and Latent Failures, which is often referred to as the Swiss Cheese Model (SCM), brings together what had previously been three separate foci in safety management: individual people, engineering, and human organizations [14,15,70]. Overall, the SCM formalizes that holes, i.e., failures, which are the metaphorical basis of the Swiss Cheese Model, can open and close in the layers of a system. Open holes in different layers can align with each other but then unalign when a hole closes in one layer. Within the SCM, active failures can be immediately apparent in relation to specific actions, whereas latent failures can lie dormant and go unrecognized as underlying causes of current active failures and causes of potential future active failures [71,72].

The SCM corresponds with fundamental processes in Quality Management Systems (QMSs). In particular, what work should be done by human organizations, and how that work should be done, can be formally documented in QMSs as work procedures. Work that does not conform to work procedures, which would be called an active failure in the SCM, is referred to as a non-conformance in QMSs. Within QMSs, non-conformances, i.e., active failures, can be addressed through corrective actions. Within QMSs, sources of non-conformances, i.e., latent failures, can be addressed through preventative actions following
the application of methods such as failure modes and effects analysis (FMEA) which can be applied with QMSs and with the SCM [73]. Human organizations, artificial intelligence, and individual people can be framed within the SCM as being three interrelated layers in a HAI system. Figure 1 illustrates the SCM framing of interactions between loopholes in HAI systems. Within the scope of this paper, the holes represent loopholes.

Figure 1. SCM framing of interactions between loopholes in HAI systems.

4.2. Applying SCM to HAI System for Delivery Driving

4.2.1. Human Organizations

For human organizations, what should be done can be formally documented in quality management systems (QMSs) as work procedures. Loopholes can emerge when work procedures have insufficient breadth and/or detail (i.e., inadequate scope), and may be exploited under insufficient observation and enforcement (i.e., inadequate control). The scope will not be adequate if QMS documentation does not encompass in sufficient detail all possible situations related to the HAI system. As explained in Section 2, loopholes in human organizations can arise when the scope of QMS documentation is not up-to-date with changes in legal regulations. Also, as discussed in Section 2, road traffic regulations and the means of their implementation can be changed frequently. Nonetheless, QMS work procedures can be adequate by stating that current road traffic regulations must be adhered to by delivery drivers. Control will not be sufficient if not all work procedures can be observed and enforced. It is already common practice for a wide range of driving information to be collected. For example, digital tachographs are installed in some delivery vehicles and can record a wide variety of driving information [74]. Also, there can be in-vehicle surveillance [75]. This can raise concerns about privacy and surveillance, but nonetheless, such technologies are well-established. Thus, there may be few loopholes in the human organization layer unless group affiliations within the human organization override adherence to legal regulations. An example is the managers of a delivery organization arranging deliveries by self-employed gig economy delivery drivers to have short time durations, which leads to delivery drivers frequently exceeding speed limits [76]. The loophole for the organization is that it is the self-employed delivery drivers who are fined for speeding, not the organization’s management group. Thus, the delivery organization is responsible for the latent failure of insufficient time allocations for deliveries. However, it is the self-employed delivery drivers who suffer the consequences of the active failure of driving faster than speed limits.

4.2.2. AI Implementation

For human organizations’ artificial intelligence implementations, scope will not be adequate unless the artificial intelligence implementation encompasses all relevant work procedures in sufficient detail related to the HAI system. Control will not be adequate
unless the implementation of work procedures can be observed and enforced. Currently, it can be extremely difficult for AI to identify and address loopholes [9,77], which can lead to active failures and latent failures of work procedures. Nonetheless, efforts to close loopholes can begin with the AI implementations of human organizations being set up in appropriate systems architectures, such as typical systems architectures for automated driving which include real-time risk estimation. As shown in Figure 2, such architectures can include modules for sensing, understanding, predicting, planning, and acting [78], and can be situated within nested sources of safety requirements including laws, regulations, and standards.

![Figure 2. Nested sources of safety requirements for a HAI system.](image)

In efforts to close loopholes, systems architecture can be designed to incentivize people to make decisions that prioritize accident avoidance. This can be carried out through reminders of the importance of adhering to driving laws and regulations [79]. Research indicates that reminders of ethical context can reduce unethical behavior [80]. It is already common practice for a wide range of data to be collected about driving behavior [74,81]. These could be repurposed to contribute to the prediction of pressure on drivers that could motivate the exploitation of loopholes.

Human drivers of road vehicles, as anybody else, can experience psychophysiological depletion through frequent exercise of self-regulatory control [82], for example, self-regulatory control when overriding the temptation to continue driving when traffic lights are amber and driving faster than the speed limit when late. Such psychophysiological fatigue is important because people can need self-regulatory resources in order to be able to resist unethical temptation [83–85]. Temptation to violate road traffic regulations, such as speed limits, could increase when there is a high ratio of red traffic light signals ( rtl ) to green traffic light signals ( gtl ) during increasing time pressure due to traffic congestion, in terms of the ratio of actual driving time to destination ( adt ) compared to standard driving time to destination ( sdt ), i.e., immediate pressure potential ( IPP ) = rtl/gtl + adt/sdt. This same IPP metric can be used in different ways, depending on the specific AI implementation. A simple rule-based system can be implemented by thresholding the sum of these ratios over an arbitrary number, e.g., two, indicating an excess of immediate pressure potential. Accordingly, propositional logic can be applied: IF IPP is greater than two, THEN send reminders of the unacceptability of not adhering to driving laws and regulations. Such deterministic implementations can come within the scope of expert systems, which can deploy handcrafted Good Old-Fashioned AI (GOFAI). However, it is difficult to identify and program as hard-coded behavior in a rule-based implementation each specific case that may appear in vehicle traffic. Efforts to overcome this limitation can be made with data-driven probabilistic machine learning and with hybrid approaches.

However, as described above in Sections 3.2 and 3.3, there are also loopholes in their operation. For example, there are machine learning training datasets available for traffic signs and lights, which can enable the recognition of green and red traffic lights [86,87]. Tests using deep neural networks with datasets display baselines above 95 percent for the precision and recall of traffic light detection [88]. Also, machine learning system
accuracy can be improved beyond performance on fixed datasets, for example, through the incorporation of prior maps, which is an option several human organizations have established [89]. However, in practice, the functioning of artificial intelligence systems can be affected negatively by occlusion, varying lighting conditions, and multiple traffic lights at intersections [90]. Accordingly, it cannot be assumed that there will not be loopholes in the automated recognition of traffic light signals.

A further option is to apply machine learning in the analysis of the human driver’s psychophysiological fatigue, which could reduce self-regulatory control when having to override temptation to violate traffic regulations. A human driver’s baseline level of fatigue can be measured through gait analyses enabled by machine learning [91]. Real-time measurement of driver fatigue could possibly be carried out using cameras, but this can require particular lighting conditions [92], which are not necessarily consistently available inside delivery vehicles at different times of the day and night. Real-time measurement of driver fatigue could be carried out with different types of devices, which can be worn in the ear, on the wrist, around the head, and/or around the chest. The datasets collected with such devices can be analyzed with machine learning [93]. However, real-time measurements of drivers’ psychophysiological states may not always be accurate, because of issues such as lagged responses, inadequate sampling rates, motion-related artifacts, poor signal-to-noise ratio, and signal aliasing [94].

To address potential loopholes in individual assessments, there could be multiple assessments of immediate pressure on vehicle drivers and assessments of their psychophysiological states [95]. Multiple interrelated assessments of immediate pressure should better enable control for the sending of automated safety messages. In particular, control should encompass both the level of drivers’ temptation to violate traffic regulations due to the immediate pressure on them and the level of drivers’ motivation to resist that temptation due to their immediate fatigue. Such assessments could be facilitated by hybrid machine learning that incorporates deterministic and probabilistic elements with algorithms to check that human-defined constraints are maintained in the subsequent relationships between inputs and outputs that are learned during training [69]. Hybrid machine learning algorithms, such as those of AML, could enable the embedding of specific high-level directives. For example, by guaranteeing the fulfillment of the IPP metric, while simultaneously enforcing rules based on information inferred from complex data, such as responses to sensorial input which indicates stress and other psychophysiological states of the drivers. Embedding high-level directives is essential for incorporating domain-specific expert knowledge, which is required as a distribution shift can theoretically only be fully avoided at the limit of covering all of the potential sample space (an infeasible assumption for real-world scenarios, in which case inference would not even be required, and rules could be solely based on memorization). Similarly, the actual process of inference/prediction is necessary for generalization throughout where possible. In addition to enabling embedding high-level directives, hybrid algorithms such as AML could generate rules that would otherwise be hidden in complex data.

4.2.3. Individual People

Loopholes can emerge when work procedures have insufficient breadth and/or detail (i.e., inadequate scope), and may be exploited under insufficient observation and enforcement (i.e., inadequate control). There are practical limits to how much observation and enforcement of procedures there can be in vehicle driving. A loophole for individual drivers is that they can seek to override AI functionality in human–AI systems, which can lead to worse risk reductions than intended in the introduction of new vehicle technologies [96]. For the automated reminder procedure outlined above, this potential loophole can be addressed by delivery drivers having input into the design of the system architecture’s planning module to determine what types of reminders will be sent as a result of what types of predictions. This is because the effectiveness of automated reminders could be facilitated by recipients having some freedom of choice in their design and/or use [97,98].
For example, lasting changes in personal preferences have been linked to making a personal choice in what is described as a choice-induced preference change. Moreover, allowing some freedom of choice in new rules can reduce the potential for stress to arise from their implementation [99]. In other words, involving delivery drivers in the engineering design of HAI systems could reduce latent failures, which could subsequently reduce active failures.

4.2.4. HAI System

This example illustrates loopholes in interactions between human organizations, AI, and individual people. Loopholes in AI implementation could be closed through triangulation. However, in this example, there is a loophole between delivery organizations and individual drivers. That is, many delivery organizations do not directly employ delivery drivers. Rather, delivery drivers are categorized as self-employed. Furthermore, delivery drivers could be fined by the delivery organization if they do not make deliveries on time. Hence, delivery drivers can be under organizational pressure as well as time pressure to drive faster than speed limits. As described above, such pressures could be addressed to some extent by involving delivery drivers in the engineering design of HAI systems that incorporate the automated sending of reminders about the unacceptability of not adhering to driving laws and regulations. However, such an innovation would not be consistent with delivery organizations classifying delivery drivers as being self-employed. This example highlights the importance of first identifying latent failures in interactions between human organizations and individual people, and then analyzing how the introduction of narrow task-specific AI could be used to address those latent failures without introducing new sources of latent failures.

As illustrated in Figure 3, the application of the SCM should consider all layers in HAI systems and the relative amount of influence of the different layers. In this case, the layer of government regulation over the employment of delivery drivers should be included, and the relative lack of influence of AI over outcomes should be recognized.

Figure 3. SCM framing of HAI system for gig economy driving with AI nullified.

In particular, as illustrated in Figure 3, loopholes in government regulations for employing drivers [16,17] can be exploited by delivery companies that use drivers as self-employed gig economy workers. Specifically, gig economy drivers can be under so much organizational pressure and time pressure that they drive faster than speed limits. The potential of AI to reduce the incidence of speeding through safety technologies such as sending automated reminder messages can be nullified by delivery companies not encouraging such safety technology or even the established Intelligent Speed Assistance Technologies (ISATs), which are present in new vehicles but not necessarily in older vehicles used by self-employed gig-economy drivers. Also, delivery drivers can ignore messages or ISATs. Furthermore, as discussed in Section 2.4, AI safety technology such as computer
vision in speed cameras can be nullified by delivery drivers who can put reflective tape over vehicle number plates or even put plastic bags over speed camera lenses [55]. Thus, Figure 3 illustrates that it is at least as important to examine the potential for loopholes other than AI loopholes in HAI systems. It is also relevant to note that a common manifestation of AI loopholes is that of inconsistencies in terms of data. Deterministic AI systems can be unable to cope with these inconsistencies and can lead to spurious behavior when near pre-established thresholds. Probabilistic AI systems can have a propensity to silently fail in the face of out-of-distribution samples. However, it is possible that a hybrid AI such as AML [69] could be prepared to be both robust to inconsistencies, as well as to explicitly announce to users that a potential loophole has been detected.

5. Conclusions

5.1. Principal Contributions

The focus of this paper is one example of safety risks arising in interactions between individual people and the narrow task-specific artificial intelligence (AI) implementations of human organizations. There are several contributions to debate about the safety of HAI systems. First, an examination of sources of loopholes in interactions between human organizations and individual people is provided, which encompasses culture-bound social choice in the changing definition of legal regulations and the prioritization of group affiliations over legal regulations. Second, sources of loopholes are described from three types of AI: deterministic AI, probabilistic AI, and hybrid AI. Third, the Swiss Cheese Model (SCM) is related to HAI systems in terms of loopholes in interactions between human organizations, AI, and individual people. Fourth, the SCM is applied to the example of the gig economy driving work, which relates to the SCM and quality management systems (QMSs).

5.2. Practical Implications

The main practical implication is that the design and operation of HAI systems should take into account potential loopholes in interactions between human organizations, AI, and individual people. As illustrated in Figure 1, human organizations, artificial intelligence, and individual people can be framed within the SCM as being three interrelated layers in an HAI system. As explained with the example in Section 4, the emergence of loopholes in all three layers can be considered in terms of scope and control.

A further practical implication is that it should not be assumed that motivation to exploit loopholes in HAI systems is limited to a few so-called bad actors. Rather, motivation to exploit loopholes can arise from common pressures experienced by many people in their working lives. Hence, the most important loopholes may not be in the technical details of AI implementations, but in the social contexts of AI implementations. For example, social contexts can lead to many speed cameras being vandalized. As discussed in Section 2.4, there can be unpredictable cycles of loophole emergence and exploitation as organizations and people with different interests engage in tit-for-tat exchanges. In particular, technologies can be combined and used in new ways, which were never envisaged by those who developed them to enable the circumvention of legal regulations and associated standards. The precarious environment of the so-called gig economy [100] can add to the stress and fatigue long associated with driving work [101]. The more precarious survival is in changing environments, the more potential there can be for preferring not to prioritize what should be done as defined in laws, regulations, and standards [102,103]. This can be because of the combination of increased focus on the need to fulfill basic physiological needs and the depletion of self-control resources. In particular, people may need self-regulatory resources in order to be able to resist unethical temptation. In gig economy driving work, depletion of self-control resources can occur within the context of pressure to violate traffic regulations. This everyday example illustrates the need to engineer loopholes out of all HAI systems, not only those that are considered to be potential targets for so-called bad actors.
5.3. Directions for Future Research

Whatever machine learning methods are applied, there are fundamental challenges in achieving equivalence between natural language definitions of regulations, logical formalizations of regulations, for example in algebra, and implementations of regulations with computational languages. Thus, it would be unrealistic to expect any type of machine learning to bring simple solutions for addressing loopholes in AI implementations of regulations. Furthermore, as there are few people who have all the necessary skills (legal, mathematics, computer science) to facilitate the achievement of equivalence between natural language, logical formulations, and computational languages, equivalence may only be possible through the application of collaborative harmonization methods that are used in established translation practices [104]. Accordingly, one direction for future research is to investigate what insights can be found for minimizing loopholes in AI implementations from such harmonization practices, and how these practices could be enhanced through combination with the SCM. In addition, future research can encompass more challenging applications for the SCM in the engineering design of HAI systems. There are many potential applications for HAI systems for which very detailed exploratory research will be required to define potential loopholes and how they can be addressed. For example, there are many challenging settings for human–machine teams, which can include the application of AI. However, there has been little, if any, previous consideration of loopholes in studies concerned with the safety of human–machine systems.

Author Contributions: Conceptualization, investigation, writing, S.F. and J.G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union (EU) Horizon 2020 project ALMA grant number 952091.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nindler, R. The United Nation’s capability to manage existential risks with a focus on artificial intelligence. Int. Commun. Law Rev. 2019, 21, 5–34. [CrossRef]
12. Baur, C.; Soucek, R.; Kühnen, U.; Baumeister, R.F. Unable to resist the temptation to tell the truth or to lie for the organization? Identification makes the difference. J. Bus. Ethics 2020, 167, 643–662. [CrossRef]
77. Knox, W.B.; Allievi, A.; Banzhaf, H.; Schmitt, F.; Stone, P. Reward (mis)design for autonomous driving. Artif. Intell. 2023, 316, 103829. [CrossRef]
90. Williams, J.; Francombe, J.; Murphy, D. Evaluating the Influence of Room Illumination on Camera-Based Physiological Measurements for the Assessment of Screen-Based Media. Appl. Sci. 2022, 12, 3083. [CrossRef]
92. Williams, J.; Francombe, J.; Murphy, D. Evaluating the Influence of Room Illumination on Camera-Based Physiological Measurements for the Assessment of Screen-Based Media. Appl. Sci. 2022, 13, 8482. [CrossRef]
93. Williams, J.; Francombe, J.; Murphy, D. Evaluating the Influence of Room Illumination on Camera-Based Physiological Measurements for the Assessment of Screen-Based Media. Appl. Sci. 2022, 13, 8482. [CrossRef]
95. Williams, J.; Francombe, J.; Murphy, D. Evaluating the Influence of Room Illumination on Camera-Based Physiological Measurements for the Assessment of Screen-Based Media. Appl. Sci. 2022, 13, 8482. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.