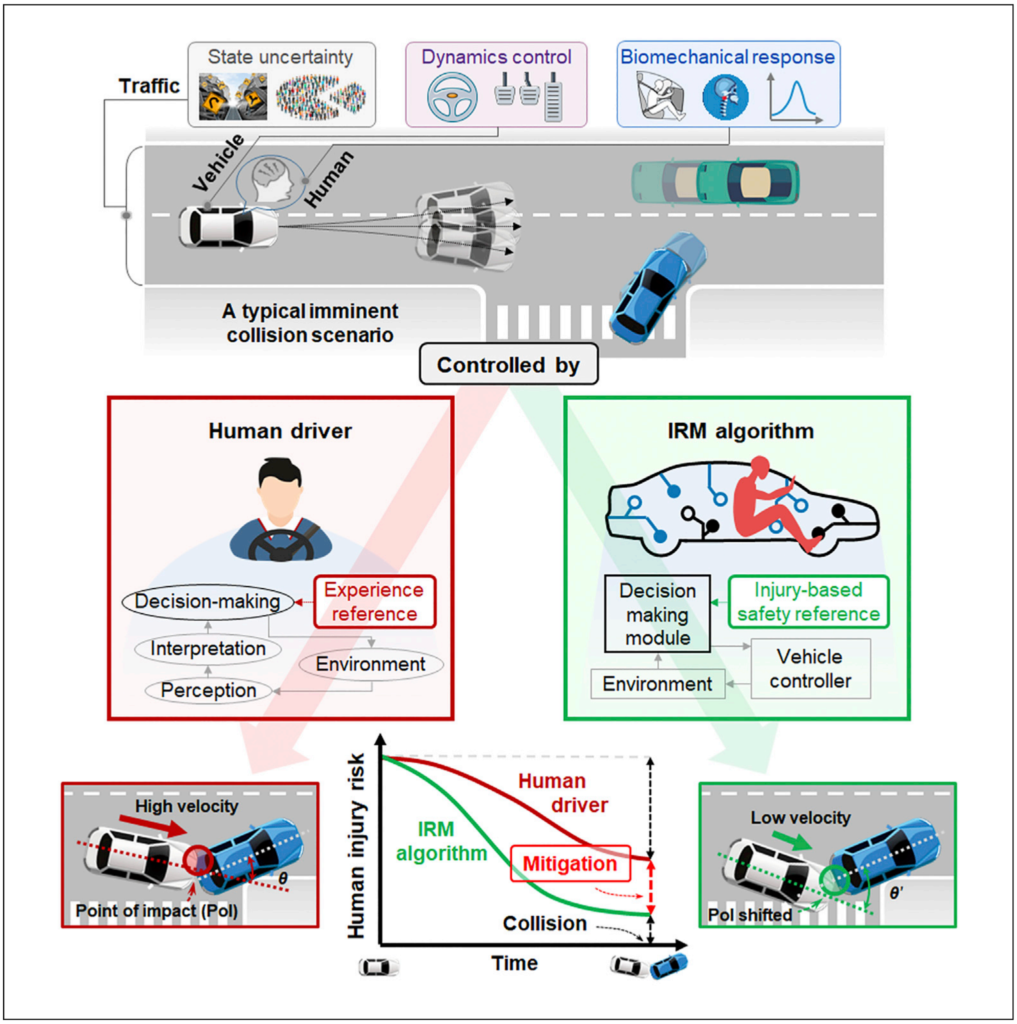
iScience

CellPress
OPEN ACCESS

Article

Human injury-based safety decision of automated vehicles



Qingfan Wang, Qing Zhou, Miao Lin, Bingbing Nie

nbb@tsinghua.edu.cn

Highlights

We propose an injury risk mitigation-based decision-making algorithm for AVs

A real-time, data-driven human injury prediction model was established

We applied mining tools to identify mechanisms embedded in accumulated crash data

We analyzed traffic ethical issues quantitatively, closer to feasible solutions

Wang et al., iScience *25*, 104703
August 19, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.104703

iScience

CellPress
OPEN ACCESS

Article

Human injury-based safety decision of automated vehicles

Qingfan Wang,[1] Qing Zhou,[1] Miao Lin,[2] and Bingbing Nie[1,3,*]

SUMMARY

Automated vehicles (AVs) are anticipated to improve road traffic safety. However, prevailing decision-making algorithms have largely neglected the potential to mitigate injuries when confronting inevitable obstacles. To explore whether, how, and to what extent AVs can enhance human protection, we propose an injury risk mitigation-based decision-making algorithm. The algorithm is guided by a real-time, data-driven human injury prediction model and is assessed using detailed first-hand information collected from real-world crashes. The results demonstrate that integrating injury prediction into decision-making is promising for reducing traffic casualties. Because safety decisions involve harm distribution for different participants, we further analyze the potential ethical issues quantitatively, providing a technically critical step closer to settling such dilemmas. This work demonstrates the feasibility of applying mining tools to identify the underlying mechanisms embedded in crash data accumulated over time and opens the way for future AVs to facilitate optimal road traffic safety.

INTRODUCTION

Road traffic crashes remain a serious public health and social problem, resulting in 1.35 million fatalities, tens of millions of injuries, and 3% GDP loss annually (World Health Organization, 2018). Automated vehicles (AVs) are expected to revolutionize road traffic safety and provide enhanced protection for occupants (Schwarting et al., 2018; Olaverri-Monreal et al., 2020). Such anticipated benefits of these intelligent machines largely rely on their capacity to make appropriate safety decisions. Decisions are usually achieved using rule-based (Campbell et al., 2010; Kala and Warwick, 2013), optimization-based (Liniger et al., 2015; Kolekar et al., 2020), and learning-based methods (Sallab et al., 2017; Lechner et al., 2020). Existing research efforts have been devoted to enhancing AV collision avoidance ability in a normal driving mode. However, the complete avoidance of crashes remains unattainable, confirmed by the hundreds of AV-involved crashes reported by the California Department of Motor Vehicles (California DMV, 2020). Based on these valuable AV crash reports, many studies focused on analyzing AV collision frequencies, crash types, and associated contributing factors (Xu et al., 2019; Boggs et al., 2020; Kutela et al., 2022), evidencing that traffic crashes can still occur in the future due to the perception failures of ego vehicles or the irrational actions of other traffic participants (Bonnefon et al., 2016; Wang et al., 2019; Franklin et al., 2021). Yet, when confronting inevitable obstacles, little attention has been given to AV decision-making systems that manage to reduce the severity of an imminent collision.

Conventional passive safety systems (e.g., vehicle body structure and occupant restraints) take the primary responsibility for injury mitigation and work only after a collision begins. However, since vehicle maneuvers during the critical time window from risk perception to collision occurrence will directly determine the collision severity, it is possible to establish a decision-making algorithm that makes safety decisions (i.e., vehicle control inputs) focused on minimal human injury prior to an impending collision.

A vital part of such decision-making algorithms is accurate injury prediction. Occupant injuries in traffic crashes are biomechanical consequences of the human body under transferred dynamic loads (e.g., deformation and failure of tissues and bones). Inherent highly nonlinear characteristics lead to a lack of accurate mathematical methods to interpret an injury process explicitly. A few simplified indicators have been used to represent crash severity (e.g., impact location or vehicle body deformation), which cannot characterize human injuries well due to the oversimplification of human-vehicle system interactions (Wang et al., 2019; Simon et al., 2019; Parseh et al., 2021). To solve the need for injury prediction, one promising solution is to

[1] State Key Laboratory of Automotive Safety and Energy, School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China

[2] China Automotive Technology & Research Center (CATARC), Tianjin 300399, China

[3] Lead contact

*Correspondence: nbb@tsinghua.edu.cn

https://doi.org/10.1016/j.isci.2022.104703

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

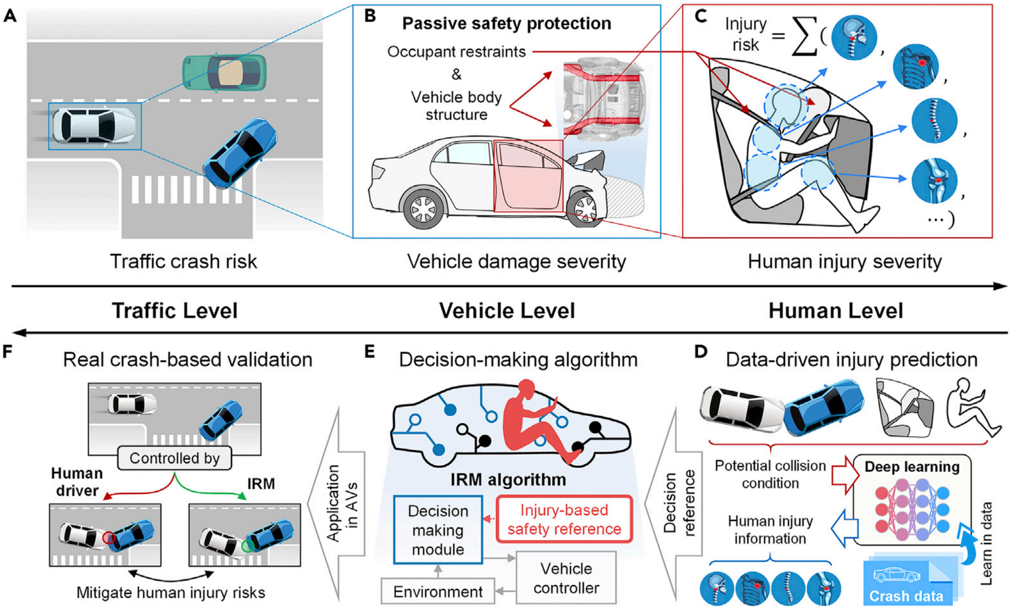


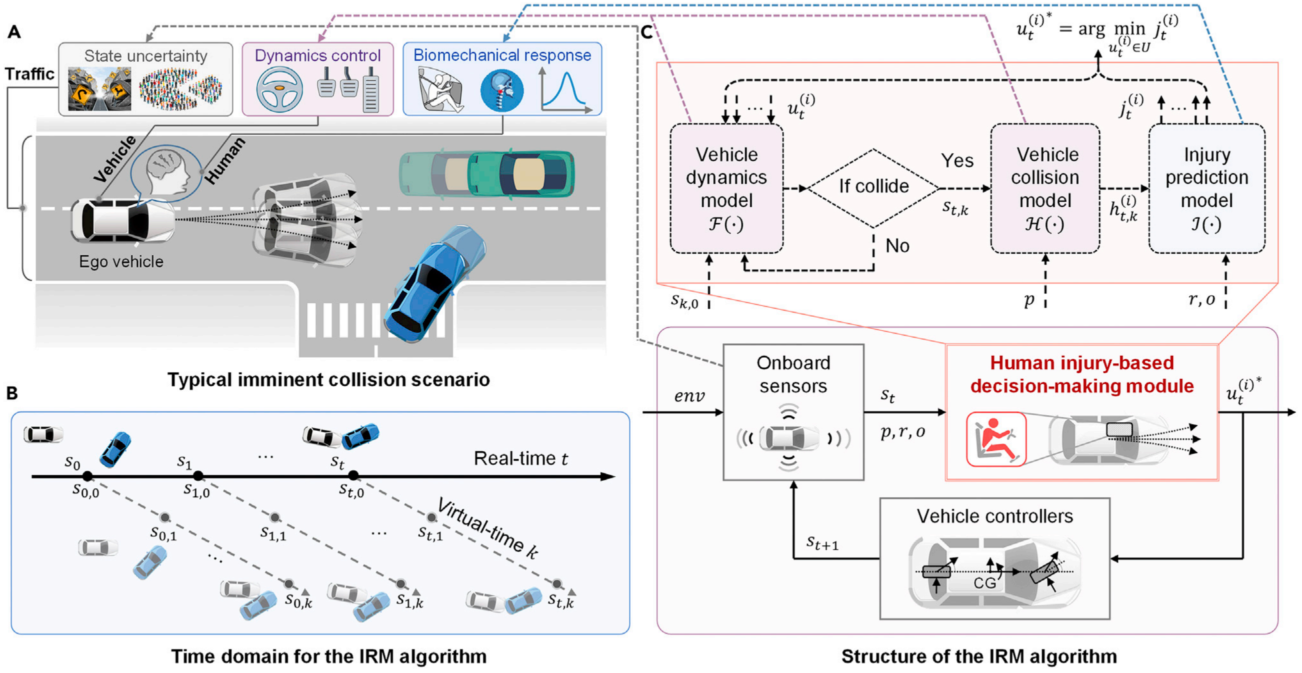
Figure 1. Multiple scales of the road traffic safety study and human injury-based decisions: the traffic, vehicle, and human levels

(A–F) We understand traffic safety problems as a ''top-down'' process: the external load from a vehicle collision is transferred and human injuries are generated via a traffic-vehicle-human interaction (A–C). Following this, we formulate a decision-making algorithm as a ''bottom-up'' approach: a data-driven injury prediction model is developed by applying deep learning algorithms to accumulated crash data (D) and subsequently used as an injury reference for safety decision-making (E) to mitigate injury risk for imminent collision scenarios (F).

combine existing traffic crash information as a databank with proper mining tools (Delen et al., 2017; Li et al., 2019; Wang et al., 2021). The long-term accumulated crash information collected from different countries and regions provides vast amounts of data (Pfeiffer and Schmidt, 2007; Radja, 2016). In parallel, the rapid development of data-driven methods creates new possibilities for identifying the complex but intrinsic interactions between crash characteristics and resulting human injuries (Silva et al., 2020; Lian et al., 2020).

When multiple traffic participants are involved and minimal loss is being pursued, vehicle safety decisions on distributing harm and making trade-offs will possibly go further into the moral domain (Awad et al., 2018). This leads to an increasing dispute about two primary decision principles with no conclusion until now: ''egoism'' pursues the maximization of individual self-interest (i.e., the ego vehicle's safety), while ''utilitarianism'' deems the common interests of all participants as the actual motive (Mordue et al., 2020). Such social dilemmas remain a growing concern for AVs' actual application and universal popularization, which strikes a large nerve for consumers, automakers, and policymakers (De Freitas et al., 2021; Wang et al., 2022). Existing literature on ethical issues remains on the qualitative description of human safety concerns, namely, sacrifice or survival (Bonnefon et al., 2016; Awad et al., 2018; Faulhaber et al., 2019). When a subsequent injury can be precisely predicted, there is a possibility of quantifying the risks to different participants, bringing insight for in-depth analysis of social dilemmas.

This paper explores whether, how, and to what extent such human injury-based decisions by AVs can enhance human protection in improved traffic safety. Understanding safety protection in road traffic follows a ''top-down'' process involving traffic, vehicle, and human levels (Figures 1A–1C). Correspondingly, we propose a framework of an injury risk mitigation (IRM)-based decision-making algorithm that makes safety decisions toward minimizing injury risk, which incorporates a ''bottom-up'' process (Figures 1D–1F). First, an occupant injury prediction model was established using deep learning algorithms to mine accumulated traffic crash data (Figure 1D). We then formulated a decision-making algorithm that generates vehicle control inputs according to the quantified injury information, which, in essence, aims to



Typical imminent collision scenario

Time domain for the IRM algorithm

Structure of the IRM algorithm

Figure 2. Formulation of the IRM algorithm

Confounding factors underlying the determination of the optimal safety decisions are first diagnosed at three levels: the state uncertainty of surrounding vehicles at the traffic level, dramatic dynamic changes at the vehicle level, and complex biomechanical responses at the human level. We then propose three corresponding strategies when establishing the IRM algorithm: introducing state-of-the-art sensor models and vehicle-to-vehicle communication (see STAR methods), combining vehicle dynamics model $\mathcal{F}(\cdot)$ and vehicle collision model $\mathcal{H}(\cdot)$, and establishing a data-driven human injury prediction model $\mathcal{I}(\cdot)$.

(A) A typical imminent collision scenario. A vehicle (white) is driving through a T-intersection at high speed and cannot stop in time to avoid collision with the surrounding vehicles (green or blue).

(B) Time domains for the IRM algorithm. When transferring vehicle states from the real-time domain (indicated by solid lines) into the virtual-time domain (indicated by dashed lines), we have $s_t = s_{t,0}$.

(C) Structure of the IRM algorithm. The optimal vehicle safety decision $u_t^{(i)*}$ at time step t is determined based on sensed information and updates the vehicle dynamics s_{t+1} in the real-time domain. The zoomed-in decision-making block uses an enumeration method to solve the optimization problem $u_t^{(i)*} = \arg\min_{u_t^{(i)}\in U} j_t^{(i)}$. (*env*: environment information).

find the optimal collision condition that results in minimal risk (Figure 1E). The IRM algorithm was comprehensively tested in a simulation platform using real-world crashes. The results showed that integrating injury prediction into AV decision-making is promising for reducing traffic casualties (Figure 1F). Furthermore, potential ethical dilemmas were analyzed and discussed with quantified injury information via case studies. To the best of our knowledge, this study is the first to integrate real-time injury prediction into AV decision-making systems and quantify the potential benefits of human injury-based decisions for imminent collision scenarios.

RESULTS

Formulation of the IRM algorithm

Consider an imminent collision scenario of n AVs (Figure 2A). $s_t^{(i)}$ and $u_t^{(i)}$ are the i-th AV's states (e.g., position and velocity) and control inputs (i.e., accelerating, braking, and steering) at time t, respectively. Vehicle states are updated as follows according to the dynamics model $\mathcal{F}(\cdot)$:

$$s_{t+1} = \mathcal{F}(s_t, u_t) = \left[\mathcal{F}^{(1)}\left(s_t^{(1)}, u_t^{(1)}\right)^T, \cdots, \mathcal{F}^{(n)}\left(s_t^{(n)}, u_t^{(n)}\right)^T\right]^T \qquad \text{(Equation 1)}$$

The vehicle control input $u_t^{(i)} \in U$ consists of longitudinal and lateral control variables. We discretize decision space U into N decisions, i.e., $U = \{u_{(1)}, \cdots, u_{(N)}\}$, considering a trade-off between optimization accuracy and computational efficiency (see STAR methods).

The IRM algorithm performs decision optimization in the predictive horizon (i.e., virtual-time domain) and executes the optimal decision in the physical world (i.e., real-time domain) (Figure 2B). When confronting inevitable obstacles in real time t, using the vehicle collision model $\mathcal{H}(\cdot)$ (see STAR methods), the i-th AV estimates the simulated potential collision conditions $h_{t,k}^{(i)}$ (including the collision velocity, angle, and impact location of the two vehicles) as follows:

$$h_{t,k}^{(i)} = \mathcal{H}\left(s_{t,k}, u_t^{(i)}, \widehat{u}_t^{(\neg i)}, p\right) \qquad \text{(Equation 2)}$$

where $s_{t,\ k}$ denotes the vehicle states with collision occurring at virtual time k, p is vehicle structural parameters, and $\widehat{u}_t^{(\neg i)}$ represents the i-th AV's estimation of the control inputs of the surroundings $u_t^{(\neg i)}$. $\widehat{u}_t^{(\neg i)}$ equals $u_t^{(\neg i)}$ if vehicle-to-vehicle communication (V2VC) is available for the AVs. Otherwise, the surrounding vehicles are assumed to maintain their kinematics, i.e., $\widehat{u}_t^{(\neg i)}$ is estimated to be $u_{t-1}^{(\neg i)}$ based on the observed control inputs at time $t-1$.

Without loss of generality, this study focuses on driver injuries and assumes no other passengers for brevity. The occupant injury severity score (OISS) $j_t^{(i)}$, predicted by the injury prediction model $\mathcal{I}(\cdot)$ (see STAR methods) and serving as the optimization objective, is determined as follows:

$$j_t^{(i)} = \alpha^{\mathrm{T}} \cdot inj = \alpha^{\mathrm{T}} \cdot \mathcal{I}\left(h_{t,k}^{(i)}, r, o\right) \qquad \text{(Equation 3)}$$

where $inj \in \mathbb{R}^4$ represents the predicted probability vector of four injury severity levels and satisfies $inj = 1$; $\alpha = [\alpha_{\mathrm{I}}, \alpha_{\mathrm{II}}, \alpha_{\mathrm{III}}, \alpha_{\mathrm{IV}}] \in \mathbb{R}^4$ denotes a coefficient vector that assigns weights for each level; α is determined by the distribution of real-world injuries in the NASS/CDS dataset (2004–2015) (Table S1); and r, o are vehicle restraint system states and occupant characteristics identified from sensing information, respectively.

In summary, the i-th AV receives real-time predicted, periodically updated injury information and maintains the optimal safety decision with minimal OISS, i.e., $u_t^{(i)*}$, until the collision either occurs or is avoided:

$$u_t^{(i)^*} = \underset{u_t^{(i)} \in U}{\arg\min}\, j_t^{(i)} = \underset{u_t^{(i)} \in U}{\arg\min}\, \alpha^{\mathrm{T}} \cdot \mathcal{I}\left(\mathcal{H}\left(s_{t,k}, u_t^{(i)}, \widehat{u}_t^{(\neg i)}, p\right), r, o\right) \qquad \text{(Equation 4)}$$

For other AVs, the optimal decision can be determined in the same way, forming a distributed algorithm for multiple AVs, where each AV is assigned an independent decision-making problem. We consider different types of errors and time delays to enhance the robustness (see STAR methods).

Establishment of the validation scenarios

We constructed imminent collision scenarios using real-world vehicle dynamic responses on our in-house-developed simulation platform (Figure 3). We first collected and screened 200 real-world traffic crashes (human drivers; from 2012 to 2019) with detailed on-site data, including vehicle information (e.g., mass, size, and damage) and dynamics (trajectories and control inputs), occupant information (e.g., age and gender), and restraint system information (e.g., belt and airbag usage). The filter criterion was vehicle-to-vehicle crashes, middle-size passenger vehicle type, and no rollovers, occupant ejections, or multiple impacts.

The collisions were reconstructed in the simulation platform as the validation scenarios. The IRM algorithm is designed to be activated when danger emerges. If a collision eventually occurs, the critical time (i.e., from detecting an emerging risk to collision occurrence) varies among traffic scenarios. Thus, it is not feasible to set a fixed activation time for the IRM algorithm (t_{act}) in different cases. To comprehensively test the performance when implemented in AVs, we set a 1-s time window with 100-ms intervals before the actual collision occurrence and assumed that the IRM algorithm could be activated at any of the intervals. AV dynamic responses were simulated following two rules: (1) before t_{act}, the vehicle follows the real-world trajectory in the original crash, and (2) after t_{act}, the IRM algorithm exerts complete control on the vehicle's maneuver.

Injury mitigation mechanisms of the IRM algorithm

Quantitative injury prediction is provided to guide decision-making toward minimal injury in imminent collision scenarios. The IRM algorithm by itself identified and extracted the underlying mechanisms for injury mitigation from the large-scale training dataset. We screened the protection measures taken by

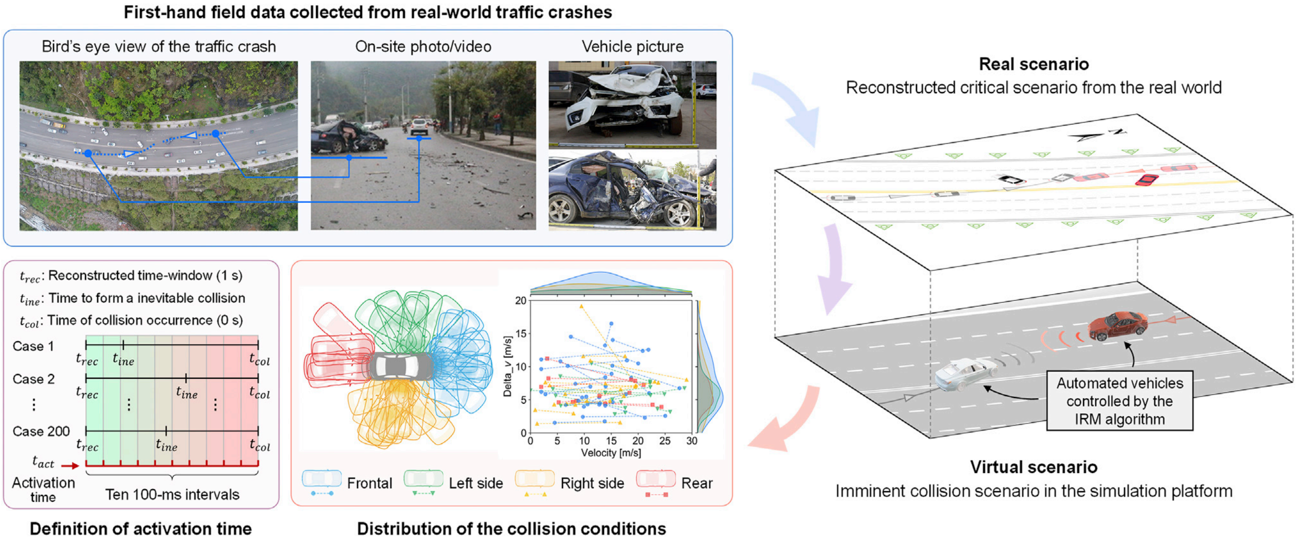

First-hand field data collected from real-world traffic crashes
Bird's eye view of the traffic crash
On-site photo/video
Vehicle picture
Real scenario
Reconstructed critical scenario from the real world
t_{rec}: Reconstructed time-window (1 s)
t_{ine}: Time to form a inevitable collision
t_{col}: Time of collision occurrence (0 s)
Case 1
Case 2
Case 200
t_{rec}
t_{ine}
t_{col}
t_{act}
Activation time
Ten 100-ms intervals
Delta_v [m/s]
Velocity [m/s]
0
5
10
15
20
25
30
Frontal
Left side
Right side
Rear
Automated vehicles controlled by the IRM algorithm
Virtual scenario
Imminent collision scenario in the simulation platform
Definition of activation time
Distribution of the collision conditions

Figure 3. Reconstructed critical scenarios for assessing the performance of the IRM algorithm

Our well-trained specialist teams collected first-hand on-site evidence (e.g., skid marks, broken glass, and bent guard rails), including locating the vehicles involved, measuring the crash damage, and identifying the state of occupant restraint systems. The distribution of collision position and velocity indicates that these collected scenarios cover a wide range of collision conditions, enhancing the comprehensiveness of the assessment. Detailed information of the crashes is provided in Table S9. The inset on the lower left illustrates the IRM algorithm's activation time covering a 1-s window before a collision.

the IRM algorithm from all the simulated scenarios and grouped them into two categories and four subcategories.

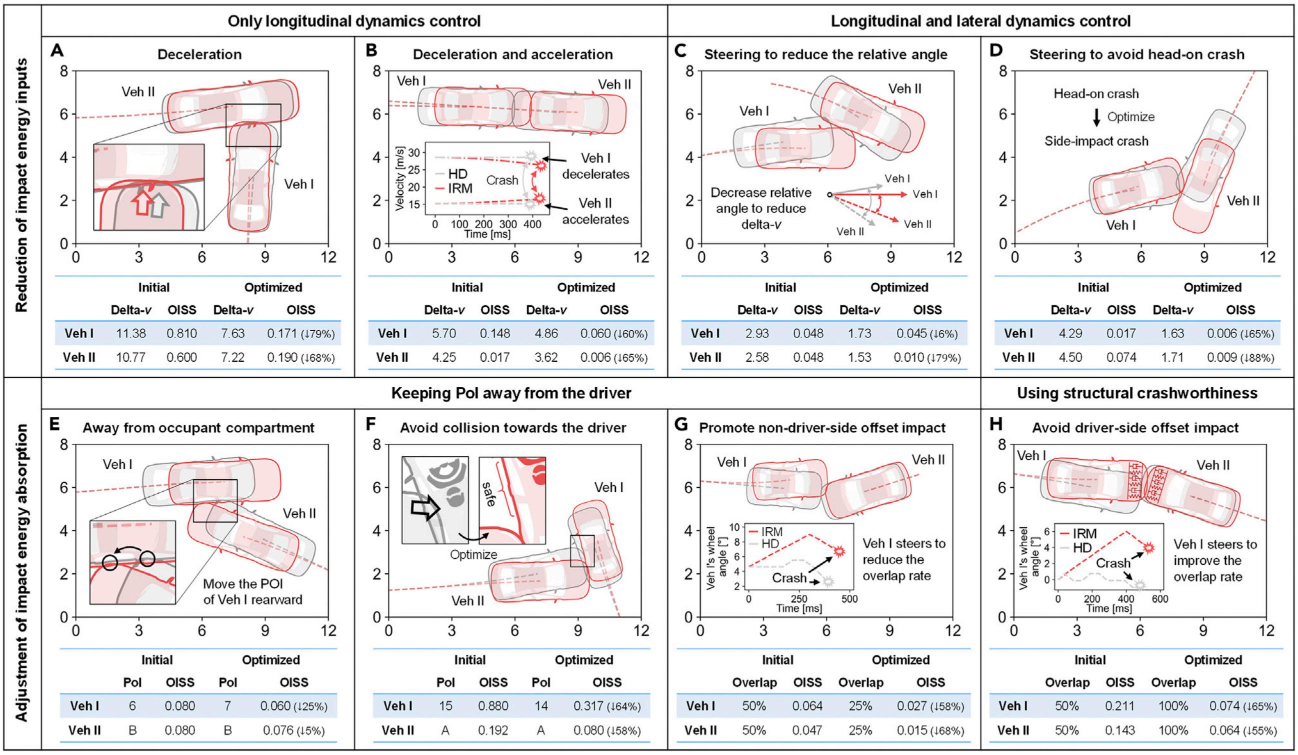
The first reduces impact energy inputs, which is the most intuitive way to mitigate collision severity (Figures 4A–4D). The impact energy inputs are usually quantified by delta-*v*, i.e., the equivalent maximum velocity change sustained by the vehicle during a crash. The IRM algorithm managed to reduce delta-*v* through 1) longitudinal dynamics control, i.e., deceleration or acceleration to reduce the relative velocity, and 2) lateral dynamics control, i.e., steering to reduce the relative angle and optimize the collision pattern (e.g., switching a head-on crash to a side-impact impact with significantly lower delta-*v*).

The second adjusts the impact energy absorption (Figures 4E–4H). For a given level of impact energy inputs, automotive engineers have been devoted to optimizing energy-absorption structures and avoiding passing energy to humans inside. One practice by the IRM algorithm keeps the point of impact (PoI) away from the occupant compartment and avoids direct collision toward the driver to guarantee the living space. The other way utilizes vehicle body structural crashworthiness (e.g., crash box and longitudinal rail). For example, the IRM algorithm improved the overlap rate to enlarge the energy-absorption structures involved for driver-side offset impact, reducing a large intrusion on the driver side.

Performance of the IRM algorithm

We present a performance assessment based on three defined application scenarios distinguished by the AV development stage from low to high (i.e., S1, S2, and S3) (see STAR methods). We also introduced emergency braking (EB) technology as a baseline strategy to represent existing active safety technology in imminent collision scenarios.

Via the combination of the above injury mitigation mechanisms, the AVs with the IRM algorithm exhibited a statistically lower OISS than human drivers in the same scenarios (Figure 5A). Taking t_{act} at −500 ms as an example, the green bar indicates that the S1-IRM algorithm reduced OISS by 2.1%–76.0% (median: 29.4%) compared with human drivers among all the simulated cases. The best injury reductions in the S2 and S3 scenarios both reach 100% (i.e., collision avoidance). In terms of time sensitivity, the mitigation of injury severity is positively related to t_{act} for a given application scenario (i.e., the OISS reduction drops gradually from 100% to 0% when t_{act} changes from −1000 ms to −100 ms). The earlier the IRM algorithm takes over


Reduction of impact energy inputs
Only longitudinal dynamics control
Longitudinal and lateral dynamics control
A
Deceleration
Veh II
Veh I
B
Deceleration and acceleration
Veh I
Veh II
Velocity [m/s]
HD
IRM
Crash
Time [ms]
Veh I decelerates
Veh II accelerates
C
Steering to reduce the relative angle
Veh I
Veh II
Decrease relative angle to reduce delta-v
D
Steering to avoid head-on crash
Head-on crash
Optimize
Side-impact crash
Veh I
Veh II
Adjustment of impact energy absorption
Keeping PoI away from the driver
Using structural crashworthiness
E
Away from occupant compartment
Veh I
Veh II
Move the POI of Veh I rearward
F
Avoid collision towards the driver
safe
Optimize
Veh I
Veh II
G
Promote non-driver-side offset impact
Veh I
Veh II
Veh I's wheel angle [°]
IRM
HD
Crash
Time [ms]
Veh I steers to reduce the overlap rate
H
Avoid driver-side offset impact
Veh I
Veh II
Veh I's wheel angle [°]
IRM
HD
Crash
Time [ms]
Veh I steers to improve the overlap rate

A

	Initial		Optimized	
	Delta-*v*	OISS	Delta-*v*	OISS
Veh I	11.38	0.810	7.63	0.171 (↓79%)
Veh II	10.77	0.600	7.22	0.190 (↓68%)

B

	Initial		Optimized	
	Delta-*v*	OISS	Delta-*v*	OISS
Veh I	5.70	0.148	4.86	0.060 (↓60%)
Veh II	4.25	0.017	3.62	0.006 (↓65%)

C

	Initial		Optimized	
	Delta-*v*	OISS	Delta-*v*	OISS
Veh I	2.93	0.048	1.73	0.045 (↓6%)
Veh II	2.58	0.048	1.53	0.010 (↓79%)

D

	Initial		Optimized	
	Delta-*v*	OISS	Delta-*v*	OISS
Veh I	4.29	0.017	1.63	0.006 (↓65%)
Veh II	4.50	0.074	1.71	0.009 (↓88%)

E

	Initial		Optimized	
	PoI	OISS	PoI	OISS
Veh I	6	0.080	7	0.060 (↓25%)
Veh II	B	0.080	B	0.076 (↓5%)

F

	Initial		Optimized	
	PoI	OISS	PoI	OISS
Veh I	15	0.880	14	0.317 (↓64%)
Veh II	A	0.192	A	0.080 (↓58%)

G

	Initial		Optimized	
	Overlap	OISS	Overlap	OISS
Veh I	50%	0.064	25%	0.027 (↓58%)
Veh II	50%	0.047	25%	0.015 (↓68%)

H

	Initial		Optimized	
	Overlap	OISS	Overlap	OISS
Veh I	50%	0.211	100%	0.074 (↓65%)
Veh II	50%	0.143	100%	0.064 (↓55%)

Figure 4. Representative scenes reflecting the protection measures taken by the IRM algorithm

In each inset, the gray and red denote the original human-driven vehicles and the IRM-controlled AVs as its counterpart; the vehicle diagram reflects the actual length measured from the collisions. (Veh: vehicle; delta-*v* measured in m/s).

(A and B) With longitudinal control only, Vehicle I reduced the relative velocity through braking (delta-*v*: 11.38 → 7.63 m/s), with the PoI nearly unchanged. A similar law remains for rear-end collisions. The rear vehicle's deceleration and the front vehicle's acceleration jointly contributed to OISS reduction for both drivers.

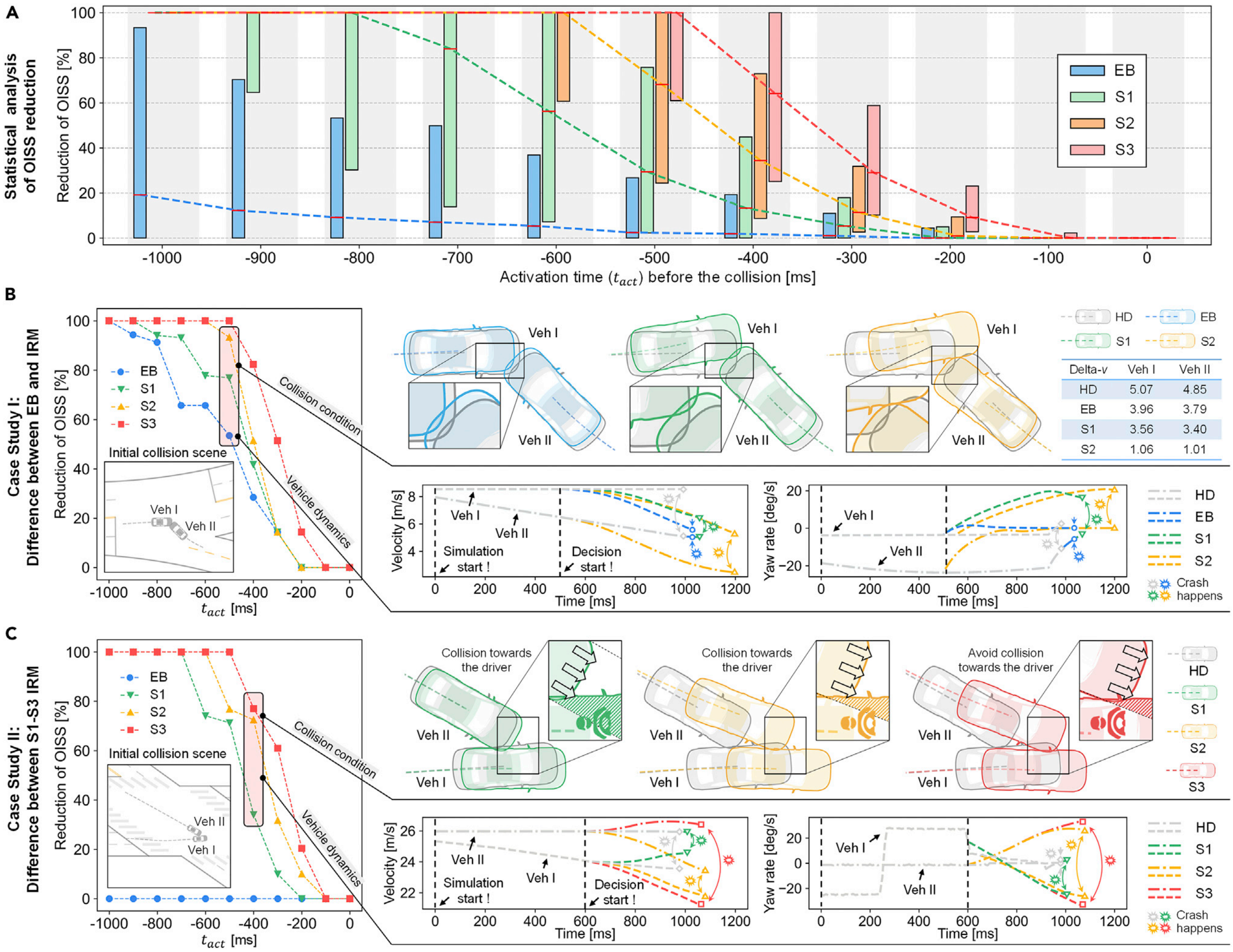
(C and D) With dynamics control in both longitudinal and lateral directions, AVs can brake and steer to either reduce the relative angle (41.3° → 21.2°) or switch a head-on crash to a side-impact impact with lower delta-*v* (4.29 → 1.63 m/s).

(E and F) AVs moved the PoI rearward (PoI: 6 → 7) or kept the PoI away from the occupant compartment (PoI: 15 → 14) to avoid passing impact energy to the driver inside.

(H) For a driver-side offset impact, the IRM algorithm managed to utilize the energy-absorption structures by increasing the overlap rate from 50% to 100%. Here, we referred to another case (G) for an interesting comparison: in the case of a non-driver-side offset impact, on the contrary, the IRM algorithm preferred to reduce the overlap rate (50% → 25%) to shift the impact further away from the driver side.

control, the more time AVs have to make appropriate maneuvers and lower injury severity. In terms of application scenarios, the S3-IRM algorithm provides the maximum benefit in injury risk reduction; S2-IRM, S1-IRM, and EB follow in decreasing order, which aligns well with the expectation of AVs' progression. The S3-IRM algorithm avoids more than half of the crashes (i.e., 55.5%) at t_{act} of −500 ms and mitigates injury severity by 64.1%, 29.0%, and 9.1% at t_{act} of −400, −300, and −200 ms, respectively. The EB strategy exhibits the worst performance in injury mitigation, demonstrating that existing safety protection remains to be further enhanced. We selected two representative crash cases that detailed the difference in injury mitigation among different application scenarios of the IRM algorithm.

Case I: IRM outperformed EB via more flexible control. The initial real-world crash occurred at a Y-crossing road without traffic lights, where ego (I) and surrounding (II) vehicles encountered frontal collisions (Figure 5B). Compared to human-driven counterparts, Vehicle I with the EB strategy reduced impact energy inputs using braking along the traveling direction (delta-*v*: 5.07 → 3.96 m/s). Furthermore, because of the more flexible control of vehicle dynamics, Vehicle I with the S1-IRM algorithm combined appropriate steering and braking, reduced the overlap rate, and obtained a lower delta-*v* (3.56 m/s). The S2-IRM algorithm simultaneously optimized the safety decisions on both vehicles. The initial head-on crash was turned into a side impact, achieving the lowest delta-*v* (1.06 m/s) and human injury severity (i.e., OISS reduction of 93.3%).



Delta-v	Veh I	Veh II
HD	5.07	4.85
EB	3.96	3.79
S1	3.56	3.40
S2	1.06	1.01

Figure 5. Performance of the IRM algorithm

(A) Statistical analysis of OISS reduction. We calculate OISS reduction by comparing the IRM algorithm's optimization results with the counterpart of human drivers. The median injury mitigation at each t_{act} is connected with dotted lines. The whiskers are omitted for the sake of simplicity.

(B) Case Study I: IRM outperformed EB via more flexible control.

(C) Case Study II: IRM optimized energy absorption from S1 to S3. The initial collision scene and OISS reduction by the IRM algorithm. The collision condition and vehicle dynamics optimized by the IRM algorithm at t_{act} of −500 ms (B) or −400 ms (C). (Veh: vehicle; HD: the original human-driven vehicles; delta-*v* measured in m/s).

Case II: IRM optimized energy absorption from S1 to S3. In the initial crash, the surrounding vehicle (II) rushed through a crossing with a speed of approximately 90 km/h at the intersection and collided with the ego vehicle (I) (Figure 5C). The EB strategy was not activated as the collision occurred from the rear side. The S1-S3 IRM algorithms avoided the collision with t_{act} earlier than −600 ms and mitigated human injuries with t_{act} later than −600 ms. Compared with S1-IRM, the S2-IRM algorithm slowed Vehicle II and reduced the relative angle to minimize the impact energy inputs, reducing delta-*v* from 4.60 m/s to 2.76 m/s. In comparison, benefiting from V2VC, the S3-IRM algorithm achieved better cooperation to position and orient two vehicles, avoiding direct collision toward the driver. Despite suffering a slightly higher delta-*v* (2.93 m/s), the S3-IRM algorithm obtained the most significant OISS reduction of 77.4%. As can be summarized from this typical case, three factors contributed to S3's best protection performance: (1) in comparison with EB, more flexible control of vehicle dynamics (i.e., both longitudinal and lateral); (2) in comparison with S1-IRM, simultaneous optimization of decisions for multiple vehicles; and (3) in comparison with S2-IRM, the availability of V2VC, which empower AVs to accurately perceive the dynamics and intentions of surrounding vehicles.

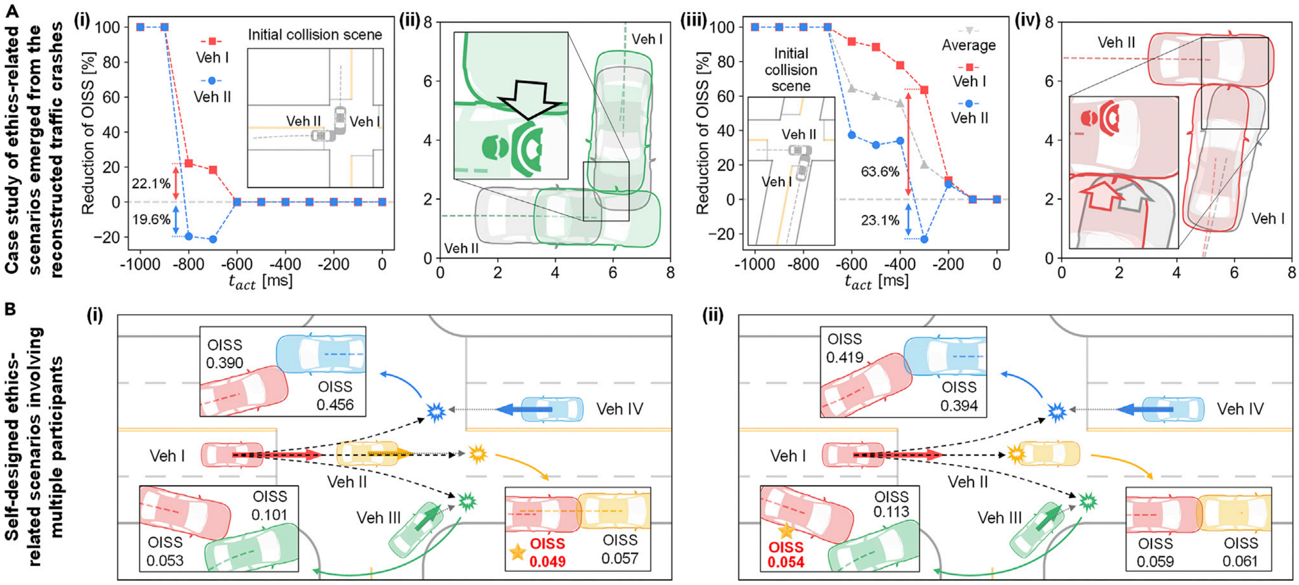


Figure 6. Possible ethical issues encountered by AVs when making safety decisions

(A) Case study of two ethics-related scenarios that emerged from reconstructed crashes. Case III (“Egoism” decision) (i-ii). Injury information and collision conditions with the S1-IRM algorithm. Case IV (“Utilitarianism” decision) (iii-iv). Injury information and collision conditions with the S3-IRM algorithm.

(B) Self-designed ethics-related scenarios involving multiple participants. The arrow length with vehicle diagrams represents the magnitude of the initial driving velocity. (Veh: vehicle).

Intervention of ethical issues

Ethical issues can emerge when AVs try to minimize human injury risks in critical scenarios, although the machines themselves do not realize this. Another two representative scenarios precisely reflect such potential social dilemmas.

Case III: An AV followed the “Egoism” decision over a human-driven vehicle (Figure 6A (i-ii)). The initial frontal collision occurred at an intersection where the surrounding vehicle (II; human-driven vehicle) rushed through due to distraction and crashed into the ego vehicle (I; AV with the S1-IRM algorithm). Facing an inevitable collision, the AV followed an “egoism” principle and made safety decisions toward minimizing its own injury risk. Vehicle I slowed down and turned the initially possible frontal collision into a side impact to reduce crash severity, decreasing its own OISS by 22.1%. However, Vehicle II suffered a direct impact on the driver side, and the resultant intrusion into the occupant compartment increased Vehicle II’s OISS by 19.6%.

Case IV: AVs followed the “Utilitarianism” decision and made a trade-off (Figure 6A (iii-iv)). The initial real-world crash occurred at an intersection at night in poor visual condition (i.e., neither of the vehicles noticed the other). Vehicle I traveled at approximately 80 km/h and collided with Vehicle II. The S3-IRM algorithm aimed to reduce the injuries of all traffic participants, categorized as the “utilitarianism” principle from a global perspective. Vehicle I slowed down and turned left slightly simultaneously, increasing the overlap of the two vehicles to utilize the energy-absorbing structures, managing to mitigate its OISS significantly by 63.6%. However, the direct collision with the occupant compartment of Vehicle II caused severe deformation. Despite the overall injury risk declining by 40.6% when considering both vehicles, the driver in Vehicle II was actually “sacrificed” with an even higher OISS (i.e., aggravation of 23.1%).

Furthermore, we intentionally extended the ethical issues to a multiple-vehicle scenario by constructing two inevitable collision scenarios at an intersection (Figure 6B). Vehicles I, II, and IV go straight through the intersection, while Vehicle III turns right. We set the ego vehicle (I) as the only IRM-controlled AV. The only difference between the two scenarios is the position and velocity of Vehicle II. For both scenarios, Vehicle I cannot stop in time to avoid collisions due to a high initial speed and had to make an ethical decision (driving straight, turning left, or turning right) that would directly affect the injury risks to the surrounding vehicles (II, III, and IV). It demonstrated that the “optimal” decision is driving straight and colliding into Vehicle II (minimal OISS of 0.049) if Vehicle II is moving (Figure 6B (i)) and that it should turn right and collide into Vehicle III (minimal OISS of 0.054) if Vehicle II is stationary (Figure 6B (ii)).

The question is, Vehicle I acts as an intelligent machine and directly decides whether the initially innocent Vehicle III will, or will not, be involved in the crash. Specifically, the egoism principle may receive public criticism for "murder-like" decisions in cases where vulnerable road users are sacrificed to protect vehicle occupants. For example, if we replace Vehicle III in Figure 6B (i) with pedestrians or cyclists, the "optimal" egoism decision for Vehicle I would be making a right turn, leading to the sacrifice of the vulnerable road users with no injury to its own occupants. Such social dilemmas have triggered a heated debate, yet there is no widely accepted solution (Bonnefon et al., 2016). As one step further than the simple qualitative descriptions (such as sacrifice or survival) used in existing studies, this study models such problems and uses quantitative injury information. Technically, this provides a safety reference to find a feasible solution to such dilemmas.

DISCUSSION

AVs are expected to drive better than human drivers in any situation. Such expectations largely rely on their decision capability and directly concern human safety, especially when confronting inevitable collisions (Simon et al., 2019). However, recent studies have suggested that AVs drive less well in scarce edge cases (i.e., near-crash scenarios) than in regular traffic scenarios (De Freitas et al., 2021). Furthermore, most existing studies on autonomous driving technology only focused on the collision avoidance capacity in dangerous conditions yet neglected injury mitigation in imminent collision scenarios (Liniger et al., 2015; Sallab et al., 2017). Thus, it is necessary to prepare a backup decision-making algorithm that activates and substitutes for the normal algorithm. To this end, we integrate a quantified evaluation of human safety into the decision-making module, formulating a human injury-based algorithm (i.e., IRM) that generates optimal vehicle control inputs for injury minimization. The IRM algorithm was thoroughly assessed using first-hand data collected from real-world traffic crashes and demonstrated a strong capability for injury mitigation. Therefore, we propose to apply such human injury-based safety decision-making algorithms to future safer AVs as the last defense line for occupant protection when a crash is inevitable and about to happen.

Predicting human injury from accumulated data

Real-time human injury prediction is a compulsory module that serves the IRM algorithm's decision optimization. Data from motor vehicle collisions have accumulated over time and can provide a possible resource for prediction needs. One difficulty lies in finding an efficient mining tool to identify the highly nonlinear relationship among human-, vehicle-, and traffic-relevant factors and injury outcomes. The other is extracting the physical features from redundant information and then turning these features into quantitative indicators that machines can understand (e.g., OISS for quantifying injury severity). Keeping this in mind, we make a pre-crash injury prediction by combining long-term accumulated traffic crash data (i.e., the NASS/CDS, 2004–2015 dataset) with emerging data-driven methods (i.e., deep learning algorithms). Compared with traditional statistical regressions, the superior performance of deep learning algorithms in handling problems involving complex nonlinearity and strong coupling can be harnessed in modeling the multi-scale traffic-vehicle-human system. Concretely, the proposed injury prediction model significantly outperformed existing regression-based methods, e.g., accuracy: 81.9% vs. 52.3%–62.5% (Stitzel et al., 2016; Fiorentini and Losa, 2020).

Identifying injury mitigation mechanisms

The IRM algorithm is essentially a "mixed" form of learning-based (i.e., data-driven injury prediction) and optimization-based (i.e., safety decisions toward minimal injury risk) methods. Such a framework differs from other mainstream AV decision-making algorithms, i.e., rule-based methods (Campbell et al., 2010; Kala and Warwick, 2013). It is usually not feasible to establish "if-else" decision-making rules to cover all possible critical scenarios due to the inherent diversity of real-world collision conditions. Meanwhile, although researchers have identified mature occupant protection mechanisms, the translation from the natural language understood by humans to the programming language understood by intelligent machines (i.e., rules) requires considerable manpower and resources, and sometimes it is complicated and ambiguous. For example, how should AVs act when two rules are contradictory in a specific scenario, e.g., reducing impact energy inputs and changing the PoI, cannot be satisfied simultaneously (Figure 5C). In contrast, by learning from objective crash data, the IRM algorithm can automatically determine the optimal safety decisions without suffering the above problems.

The learning-based model has been deemed a "black-box" and questioned regarding its interpretability, i.e., humans cannot understand the decision-making process. However, the validation results of the IRM algorithm demonstrated that its measures for reducing injuries were consistent with our perspectives on safety protection and supported by a sound mechanics basis (Figure 4). Nevertheless, not all the safety decisions it makes can be comprehensively explained. It may stem from either the inherent limitations of the prediction approach (e.g., insufficient data amount, skewed data distribution, and mining capacity) or be attributed to the fact that the IRM algorithm identified some injury mechanisms that are too complex to be observed and understood by humans directly. For example, the injury risk under a particular collision angle is significantly higher. We assume that the body structure at this angle is generally weaker, or the impact at this angle will lead to vehicle instability (e.g., out of control) and a secondary collision. We do not know if other physical laws dominate such a decision discovered by the IRM algorithm. If these uncovered underlying mechanisms can be interpreted in the future, they may provide some insights that guide a safer design of next-generation vehicles. This iteration triggered another interesting topic that deserves a thorough investigation.

Elaborating on the ethical issues with predicted injuries

Beyond technical advancements, a major parallel challenge before AVs are widely accepted lies in the ethical principles that guide machine behaviors. Such rules are in urgent need, and any details must be approached thoughtfully both in technology and morals (Shariff et al., 2017). We elaborate on the ethical issues caused by the decision principle of either egoism or utilitarianism via case studies screened from reconstructed real-world collisions (Figure 5). An ethical dilemma occurs: does a vehicle have the right to determine which surrounding vehicles to collide with, even if it can exactly know the resultant injury severity? If yes, on what basis should decisions be made? Egoism? Utilitarianism? Or a combination of the two? The first question will significantly impact the public's attitudes and the social acceptance of AVs. The second question will affect the purchase desire of potential consumers because studies show that most consumers prefer AVs with the egoism principle instead of utilitarianism (Bonnefon et al., 2016; Awad et al., 2018).

For AVs with IRM onboard, even when the resulting injuries of all the participants are precisely predicted on the technical side, it is still difficult to tell a machine in advance how to react on the moral side. The judgment on any decision principle has gone beyond the scope of this study and requires joint efforts from academia, industry, and policymakers. This study presents the first efforts to use quantitative injury descriptions of candidate safety decisions to model such social dilemmas, which, hopefully, provides a critical step closer to establishing rules to formulate AV behaviors.

Limitations of the study

The establishment of the IRM algorithm in this study has several simplifications and limitations. First, due to the size limits of currently available crash data, we decided to restrict the proposed injury prediction model to drivers. The scalability and transferability of the framework (i.e., establishing a crash dataset, training injury prediction models, and formulating safety decision algorithms) make it possible to extend to the prediction of other road traffic participant injuries, such as pedestrians and cyclists. In that way, we suggest taking all participant injury risks into account when making safety decisions. Second, we maintain the discussion on ethical issues among AVs in the present study (Figure 6) while believing that human-machine interactions can lead to complicated situations, as evidenced in recent studies (Whiting et al., 2021; Ramchurn et al., 2021). Thus, it is also necessary to evaluate the effects of the IRM algorithm when it is applied in a mixed traffic system with the progressive penetration of AVs on the road. Meanwhile, we collected real-world traffic crashes only involving conventional vehicles due to the restrictions on data access. It is suggested to further directly compare the injury mitigation performance between the existing AV decision-making algorithms and the IRM algorithm by reconstructing imminent collision scenarios using AV-involved crashes. Finally, injury mitigation is estimated based on predicted injury information, and the inevitable prediction error might be further reduced via state-of-the-art algorithms. In addition, since the IRM algorithm makes decisions by minimizing the predicted injury severity, the associated prediction error might overestimate the injury mitigation. Inspired by double Q-learning (van Hasselt et al., 2016), we suggest using two independent injury prediction models (one for decision-making and the other for injury mitigation estimation) to minimize such overestimations in future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104703.

ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of China (52072216) and by Tsinghua-Toyota Joint Research Fund. The authors would like to thank Ms. Heye Huang and Ziyu Lin for the thorough review and valuable discussion. Mr. Dongyang Xu contributed to the development of the Python GUI.

AUTHOR CONTRIBUTIONS

Q.W. designed the model, analyzed results, and wrote the paper. B.N. conceived the original concept, analyzed results, and supervised this project. Q.Z. analyzed the results and literature. M.L. collected and reconstructed the on-site crash cases. Q.W. and B.N. prepared the figures. Q.W., B.N., and Q.Z. contributed to discuss the results and helped review and edit the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 15, 2022
Revised: June 2, 2022
Accepted: June 27, 2022
Published: August 19, 2022

SUPPORTING CITATIONS

The following references appear in the Supplemental information: Alkheder et al., 2017, Assi, 2020, Chen et al., 2015, Chen et al., 2016, He et al., 2018, Ji and Levinson, 2020, Liu et al., 2020, Lubbe and Kiuchi, 2015, Mansoor et al., 2020, Rezapour et al., 2020, Rezapour and Ksaibati, 2020, Sameen and Pradhan, 2017, Tang et al., 2019, Wahab and Jiang, 2020, Wang and Kim, 2019, Zheng et al., 2019.

REFERENCES

Alkheder, S., Taamneh, M., and Taamneh, S. (2017). Severity prediction of traffic accident using an artificial neural network. J. Forecast. *36*, 100–108. https://doi.org/10.1002/for.2425.

Assi, K. (2020). Traffic crash severity prediction-A synergy by hybrid principal component analysis and machine learning models. Int. J. Environ. Res. Public Health *17*, 7598. https://doi.org/10.3390/ijerph17207598.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., and Rahwan, I. (2018). The moral machine experiment. Nature *563*, 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Boggs, A.M., Wali, B., and Khattak, A.J. (2020). Exploratory analysis of automated vehicle crashes in California: a text analytics & hierarchical Bayesian heterogeneity-based approach. Accid. Anal. Prev. *135*, 105354. https://doi.org/10.1016/j.aap.2019.105354.

Bonnefon, J.F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science *352*, 1573–1576. https://doi.org/10.1126/science.aaf2654.

California DMV (2020). Autonomous Vehicle Collision Reports - California DMV. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/.

Campbell, M., Egerstedt, M., How, J.P., and Murray, R.M. (2010). Autonomous driving in urban environments: approaches, lessons and challenges. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. *368*, 4649–4672. https://doi.org/10.1098/rsta.2010.0110.

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., and Guan, H. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. *80*, 76–88. https://doi.org/10.1016/j.aap.2015.03.036.

Chen, C., Zhang, G., Yang, J., Milton, J.C., and Alcántara, A.D. (2016). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naive Bayes (DTNB) hybrid classifier. Accid. Anal. Prev. *90*, 95–107. https://doi.org/10.1016/j.aap.2016.02.002.

Choi, W.Y., Kim, D.J., Kang, C.M., Lee, S.H., and Chung, C.C. (2018). Autonomous vehicle lateral maneuvering by approximate explicit predictive control. In In 2018 Annual American Control Conference (ACC) (IEEE), pp. 4739–4744.

De Freitas, J., Censi, A., Walker Smith, B., Di Lillo, L., Anthony, S.E., and Frazzoli, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. Proc. Natl. Acad. Sci. USA *118*. e2010202118. https://doi.org/10.1073/pnas.2010202118.

Delen, D., Tomak, L., Topuz, K., and Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. J. Trans. Health *4*, 118–131. https://doi.org/10.1016/j.jth.2017.01.009.

Eppinger, R., Sun, E., Bandak, F., Haffner, M., Khaewpong, N., Maltese, M., Kuppa, S., Nguyen, T., Takhounts, E., Tannous, R., et al. (1999). Development of Improved Injury Criteria for the Assessment of Advanced Automotive Restraint Systems: II.

Faulhaber, A.K., Dittmer, A., Blind, F., Wächter, M.A., Timm, S., Sütfeld, L.R., Stephan, A., Pipa, G., and König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. Sci. Eng. Ethics *25*, 399–418. https://doi.org/10.1007/s11948-018-0020-x.

Fiorentini, N., and Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures *5*, 61. https://doi.org/10.3390/infrastructures5070061.

Franklin, M., Awad, E., and Lagnado, D. (2021). Blaming automated vehicles in difficult situations. iScience *24*, 102252. https://doi.org/10.1016/j.isci.2021.102252.

Hancock, P.A., Nourbakhsh, I., and Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. Proc. Natl. Acad. Sci. USA *116*, 7684–7691. https://doi.org/10.1073/pnas.1805770115.

He, Y., Wu, D., Beyazit, E., Sun, X., and Wu, X. (2018). Supervised data synthesizing and evolving–a framework for real-world traffic crash severity classification. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 163–170.

Ishikawa, H. (1994). Impact center and restitution coefficients for accident reconstruction. SAE Trans. *103*, 657–671.

Ji, A., and Levinson, D. (2020). Injury severity prediction from two-vehicle crash mechanisms with machine learning and ensemble models. IEEE Open J. Intell. Transp. Syst. *1*, 217–226. https://doi.org/10.1109/Ojits.2020.3033523.

Kala, R., and Warwick, K. (2013). Motion planning of autonomous vehicles in a non-autonomous vehicle environment without speed lanes. Eng. Appl. Artif. Intel. *26*, 1588–1601. https://doi.org/10.1016/j.engappai.2013.02.001.

Kolekar, S., de Winter, J., and Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. Nat. Commun. *11*, 4850. https://doi.org/10.1038/s41467-020-18353-4.

Kutela, B., Avelar, R.E., and Bansal, P. (2022). Modeling automated vehicle crashes with a focus on vehicle at-fault, collision type, and injury outcome. J. Transp. Eng. Pt. A Syst. *148*, 04022024. https://doi.org/10.1061/Jtepbs.0000680.

Lechner, M., Hasani, R., Amini, A., Henzinger, T.A., Rus, D., and Grosu, R. (2020). Neural circuit policies enabling auditable autonomy. Nat. Mach. Intell. *2*, 642–652. https://doi.org/10.1038/s42256-020-00237-3.

Li, W., Zhu, J., Xia, Y., Gorji, M.B., and Wierzbicki, T. (2019). Data-driven safety envelope of lithium-ion batteries for electric vehicles. Joule *3*, 2703–2715. https://doi.org/10.1016/j.joule.2019.07.026.

Lian, Y., Zhang, G., Lee, J., and Huang, H. (2020). Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles. Accid. Anal. Prev. *146*, 105711. https://doi.org/10.1016/j.aap.2020.105711.

Lin, N., Zong, C., and Shi, S. (2019). The method of mass estimation considering system error in vehicle longitudinal dynamics. Energies *12*, 52. https://doi.org/10.3390/en12010052.

Liniger, A., Domahidi, A., and Morari, M. (2015). Optimization-based autonomous racing of 1:43 scale RC cars. Optim. Contr. Appl. Met. *36*, 628–647. https://doi.org/10.1002/oca.2123.

Liu, L., Zhang, X., Liu, Y., Zhu, W., and Zhao, B. (2020). An ensemble of multiple boosting methods based on classifier-specific soft voting for intelligent vehicle crash injury severity prediction. In 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 17–24.

Lu, Y., Ma, H., Smart, E., and Yu, H. (2021). Real-time performance-focused localization techniques for autonomous vehicle: a review. IEEE Trans. Intell. Transp. Syst. 1–19. https://doi.org/10.1109/Tits.2021.3077800.

Lubbe, N., and Kiuchi, T. (2015). Injury estimation for advanced automatic collision notification (aacn) in Germany. In Proceedings of the Sixth International Conference on Expert Symposium on Accident Research (ESAR).

Mansoor, U., Ratrout, N.T., Rahman, S.M., and Assi, K. (2020). Crash severity prediction using two-layer ensemble machine learning model for proactive emergency management. IEEE Access *8*, 210750–210762. https://doi.org/10.1109/Access.2020.3040165.

Mordue, G., Yeung, A., and Wu, F. (2020). The looming challenges of regulating high level autonomous vehicles. Transp. Res. Pt. A Policy Pract. *132*, 174–187. https://doi.org/10.1016/j.tra.2019.11.007.

Olaverri-Monreal, C. (2020). Promoting trust in self-driving vehicles. Nat. Electron. *3*, 292–294. https://doi.org/10.1038/s41928-020-0434-8.

Parseh, M., Asplund, F., Svensson, L., Sinz, W., Tomasch, E., and Torngren, M. (2021). A data-driven method towards minimizing collision severity for highly automated vehicles. IEEE Trans. Intell. Veh. *6*, 723–735. https://doi.org/10.1109/Tiv.2021.3061907.

Pfeiffer, M., and Schmidt, J. (2007). Statistical and methodological foundations of the GIDAS accident survey system. In Proceedings of the 2nd International Conference on ESAR, pp. 81–87.

Radja, G.A. (2016). National Automotive Sampling System-Crashworthiness Data System, 2015 Analytical User's Manual.

Ramchurn, S.D., Stein, S., and Jennings, N.R. (2021). Trustworthy human-AI partnerships. iScience *24*, 102891. https://doi.org/10.1016/j.isci.2021.102891.

Rezapour, M., Nazneen, S., and Ksaibati, K. (2020). Application of deep learning techniques in predicting motorcycle crash severity. Eng. Rep. *2*, e12175. https://doi.org/10.1002/eng2.12175.

Rezapour, M., and Ksaibati, K. (2020). Application of various machine learning architectures for crash prediction, considering different depths and processing layers. Eng. Rep. *2*, e12215. https://doi.org/10.1002/eng2.12215.

Sallab, A.E., Abdou, M., Perot, E., and Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. Electron. Imag. *29*, 70–76. https://doi.org/10.2352/issn.2470-1173.2017.19.avm-023.

Sameen, M., and Pradhan, B. (2017). Severity prediction of traffic accidents with recurrent neural networks. Appl. Sci. Basel *7*, 476. https://doi.org/10.3390/app7060476.

Schwarting, W., Alonso-Mora, J., and Rus, D. (2018). Planning and decision-making for autonomous vehicles. Annu. Rev. Contr. Robot. *1*, 187–210. https://doi.org/10.1146/annurev-control-060117-105157.

Shariff, A., Bonnefon, J.F., and Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. Nat. Hum. Behav. *1*, 694–696. https://doi.org/10.1038/s41562-017-0202-6.

Silva, P.B., Andrade, M., and Ferreira, S. (2020). Machine learning applied to road safety modeling: a systematic literature review. J. Traffic Transp. Eng. Engl. Ed. *7*, 775–790. https://doi.org/10.1016/j.jtte.2020.07.004.

Simon, B., Franke, F., Riegl, P., and Gaull, A. (2019). Motion planning for collision mitigation via FEM-based crash severity maps. In 2019 30th IEEE Intelligent Vehicles Symposium, pp. 2187–2194.

Stitzel, J.D., Weaver, A.A., Talton, J.W., Barnard, R.T., Schoell, S.L., Doud, A.N., Martin, R.S., and Meredith, J.W. (2016). An injury severity-time sensitivity-and predictability-based advanced automatic crash notification algorithm improves motor vehicle crash occupant triage. J. Am. Coll. Surg. *222*, 1211–1219.e6. https://doi.org/10.1016/j.jamcollsurg.2016.03.028.

Tang, J., Liang, J., Han, C., Li, Z., and Huang, H. (2019). Crash injury severity analysis using a two-layer Stacking framework. Accid. Anal. Prev. *122*, 226–238. https://doi.org/10.1016/j.aap.2018.10.016.

van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double Q-learning. In Thirtieth Aaai Conference on Artificial Intelligence, pp. 2094–2100.

Wahab, L., and Jiang, H. (2020). Severity prediction of motorcycle crashes with machine learning methods. Int. J. Crashworthiness *25*, 485–492. https://doi.org/10.1080/13588265.2019.1616885.

Wang, H., Huang, Y., Khajepour, A., Zhang, Y., Rasekhipour, Y., and Cao, D. (2019). Crash mitigation in motion planning for autonomous vehicles. IEEE Trans. Intell. Transp. Syst. *20*, 3313–3323. https://doi.org/10.1109/Tits.2018.2873921.

Wang, H., Khajepour, A., Cao, D., and Liu, T. (2022). Ethical decision making in autonomous vehicles: challenges and research progress. IEEE Intell. Transp. Syst. Mag. *14*, 6–17. https://doi.org/10.1109/Mits.2019.2953556.

Wang, Q., Gan, S., Chen, W., Li, Q., and Nie, B. (2021). A data-driven, kinematic feature-based, near real-time algorithm for injury severity prediction of vehicle occupants. Accid. Anal. Prev. *156*, 106149. https://doi.org/10.1016/j.aap.2021.106149.

Wang, X., and Kim, S.H. (2019). Prediction and factor identification for crash severity: comparison of discrete choice and tree-based models. Transp. Res. Record *2673*, 640–653. https://doi.org/10.1177/0361198119844456.

Whiting, T., Gautam, A., Tye, J., Simmons, M., Henstrom, J., Oudah, M., and Crandall, J.W. (2021). Confronting barriers to human-robot cooperation: balancing efficiency and risk in machine behavior. iScience *24*, 101963. https://doi.org/10.1016/j.isci.2020.101963.

World Health Organization (2018). Global Status Report on Road Safety 2018: Summary (No. WHO/NMH/NVI/18.20).

Xu, C., Ding, Z., Wang, C., and Li, Z. (2019). Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. J. Safety. Res. *71*, 41–47. https://doi.org/10.1016/j.jsr.2019.09.001.

Yeong, D.J., Velasco-Hernandez, G., Barry, J., and Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: a review. Sensors *21*, 2140. https://doi.org/10.3390/s21062140.

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., and Wang, Z. (2019). Traffic accident's severity prediction: a deep-learning approach-based CNN network. IEEE Access *7*, 39897–39910. https://doi.org/10.1109/Access.2019.2903319.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw data	https://crashviewer.nhtsa.dot.gov/LegacyCDS/Search	N/A
Processed data	https://github.com/wangqf1997/Human-injury-based-safety-decision-of-automated-vehicles	N/A
Software and Algorithms		
Python	https://www.python.org/	Version 3.7
Matplotlib	https://matplotlib.org/	Version 3.1.3
PyTorch	https://pytorch.org/	Version 1.7.1
Numpy	https://numpy.org/	Version 1.16.0
Scikit-learn	https://scikit-learn.org/	Version 0.0
Code from this paper	https://github.com/wangqf1997/Human-injury-based-safety-decision-of-automated-vehicles	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be handled by the lead contact, Bingbing Nie.

Materials availability

This study did not generate new materials.

Data and code availability

- Dataset: The raw traffic crash data from NASS/CDS (2004–2015) is publicly available at https://crashviewer.nhtsa.dot.gov/LegacyCDS/Search. The processed traffic crash data and our reconstructed collision data are made available on GitHub at https://github.com/wangqf1997/Human-injury-based-safety-decision-of-automated-vehicles.
- Code: All original code has been deposited on GitHub at https://github.com/wangqf1997/Human-injury-based-safety-decision-of-automated-vehicles and is publicly available as of the date of publication.
- Additional information: Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Injury risk prediction

We combined the accumulated traffic crash data with proper data-driven methods to establish the injury prediction module.

We screened crash data from the National Automotive Sampling System/Crashworthiness Data System (NASS/CDS, 2004–2015). The NASS/CDS is one of the largest and most authoritative crash datasets that collected traffic crashes from 24 locations throughout the United States and has supported many in-depth crash investigations (Radja, 2016). After data pre-processing, the final dataset consisted of 5,942 crash cases focusing on single-collision sedan-to-sedan collisions (Figure S1). Nine crash elements were selected as the explanatory valuables, including four traffic-level variables (i.e., delta-*v*, collision angle, and PoIs of the ego vehicle and the opposing vehicle), three vehicle-level variables (i.e., vehicle mass ratio, belt, and airbag usage), and two human-level variables (i.e., occupant age and gender). Since injuries are sensitive to different PoIs, we manually labelled detailed PoI information for each crash case according to the reconstructed scene diagram and post-crash vehicle images (Figure S2). The dependent variable is occupant injury severity, represented by the injury severity score (ISS) (Eppinger et al., 1999). The risk of suffering severe or fatal injuries increases with increasing ISS (ranging from 0 to 75). We divided ISS into four levels (i.e., I: 0, II: 1-3, III: 4–14, and IV: 15–75) to describe injuries with different degrees of severity, i.e., the injury prediction is a four-classification task. The detailed descriptions of variables are provided in Table S1, Figures S3 and S4.

Most occupants in real-world collisions are non or slightly injured. Such data imbalance, where the severely injured remain underrepresented, can affect prediction performance in model training. We alleviated this issue through resampling methods to adjust the label distribution (i.e., the ratio between different injury severity levels). There are three categories of resampling methods: under-sampling (US), over-sampling (OS), and the combination of under- and over- sampling (CS). US methods balance data labels by discarding samples from the majority class, while OS methods add samples to the minority class. More specifically, we selected the edited nearest neighbour (ENN), the synthetic minority over-sampling technique (SMOTE), and their combination (SMOTE-ENN) for US, OS, and CS methods, respectively. Meanwhile, we randomly divided the dataset into three sets: training set (70%), validation set (15%), and test set (15%). The label distribution of the resampled dataset is displayed in Table S2.

The best-performance data-driven method usually varies with the data source and filter criteria. We established five machine learning algorithms (i.e., support vector machines, decision trees, k-nearest neighbours, naive Bayes, and adaptive boosting) and three deep learning algorithms (i.e., feedforward neural networks, recurrent neural networks, and convolutional neural networks) to determine the best prediction model. In terms of the loss function, the cross-entropy loss was used for the classification task. Adaptive moment estimation (ADAM) was adopted as the optimizer. The learning rate decreased with the training epochs. We adopted different regularization approaches to reduce the over-fitting problem: 1) the training process would early stop when validation loss increased in five consecutive epochs; 2) dropout layers were added to reduce complex co-adaptations on the training data; 3) normalization methods also improved the generalization ability. All eight data-driven algorithms were implemented in Python v3.7.0 (Python Software Foundation, United States) with an Intel Core i9-10900K 3.70 GHz CPU and an NVIDIA RTX3090 GPU. Five machine learning algorithms were implemented with the scikit-learn framework, while three deep learning algorithms were implemented using the PyTorch framework.

After parameter tuning, the model hyperparameters with the best performance on the validation set were determined (Tables S3 and S4). Among these data-driven algorithms, the stacked bidirectional RNN-based injury prediction model with an over-sampling imbalance treatment achieved the best prediction accuracy and a near real-time prediction ability (3.01 ± 0.44 ms) (Figure S5). Therefore, it was selected as the on-board injury prediction model $\mathcal{I}(\cdot)$ for the IRM algorithm, which was developed as follows:

$$inj = \mathcal{I}(h, r, o) \tag{Equation 5}$$

where $inj \in \mathbb{R}^4$ represents the predicted probability vector of the four injury levels and satisfies $inj_1 = 1$; h is the potential collision conditions, including the delta-*v*, collision angle, and PoIs of the two vehicles; and r and o are vehicle restraint system states and occupant characteristics, respectively.

To better evaluate the performance of the proposed model, we selected several papers from the literature on occupant injury severity prediction, including three major approaches: statistical regression (SR), machine earning (ML), and deep earning (DL) methods, and made a comprehensive comparison (Table S5). Since the prediction models were trained and tested on different crash datasets with varying characteristics (e.g., size, region, and variables), it is not suitable to directly compare the performance of different approaches. Combining the NASS/CDS data (one of the most authoritative crash databases) and deep learning algorithm (one of the most powerful data mining methods), the proposed RNN-based model demonstrated a satisfying prediction ability in three aspects: 1) obtaining a competitive prediction performance (G-mean: 0.726, accuracy: 81.9%); 2) using less and readily available input valuables (only nine critical valuables, excluding valuables not-directly related to injuries, e.g., alcohol test results and month of the year); 3) predicting occupant injuries with four severity levels (a more detailed prediction task in comparison with binary classification).

Collision condition estimation

Delta-*v* is the maximum velocity change sustained by the vehicle during the crash and indicates crash severity. We estimated delta-*v* before the collision considering its strong correlation with occupant injury severity (Equations 6, 7, 8 and 9). The estimation was based on the plane 2-DOF rigid-body collision model with momentum conservation (Figure S6).

$$P = \frac{(1+e)\left(v_2^n + h_2\omega_2 - v_1^n - h_1\omega_1\right)}{1/m_1 + 1/m_2 + h_1^2\Big/I_1 + h_2^2\Big/I_2} \qquad \text{(Equation 6)}$$

$$e = \frac{C}{v_2^n + h_2\omega_2 - v_1^n - h_1\omega_1} \qquad \text{(Equation 7)}$$

$$\Delta v_1 = P/m_1 \qquad \text{(Equation 8)}$$

$$\Delta v_2 = P/m_2 \qquad \text{(Equation 9)}$$

where *P* represents the collision impulse; *e* denotes the restitution coefficient; and *C* is a constant parameter that depends on the principal impact direction. For the two vehicles in a collision (i.e. $i = 1, 2$), Δv_i is the estimated delta-*v*; v_i and v_i^n represent the total velocity and the velocity component in the impulse direction; and m_i, I_i, ω_i, and h_i denote the mass, the moment of inertia, the yaw rate, and the distance from the CG (i.e., the centre of gravity) to the impulse line, respectively . We set $C = 2$ for frontal collision, $C = 1.5$ for side collision, and $C = 1$ for rear collision (Ishikawa, 1994).

Based on the planar 2-DOF rigid-body collision model, the onboard collision estimation model $\mathcal{H}(\cdot)$ for the IRM algorithm was developed as follows:

$$h = \mathcal{H}(s, u, p) \qquad \text{(Equation 10)}$$

where *h* represents the potential collision condition that includes the delta-*v*, collision angle, and PoIs of the two vehicles; *s* is the vehicle states at the collision moment; *u* denotes the vehicle control inputs; and *p* is vehicle structural parameters (e.g., vehicle size, mass, and moment of inertia).

Modelling of errors and time delays

The potential errors and time delays were also modelled on our in-house developed simulation platform to enhance the robustness. Two types of errors were introduced: perception and control errors. The perception error was defined as the position errors when determining the positions of the surrounding vehicles, which was developed as follows:

$$x_d = x_t + \varepsilon_x, \ \varepsilon_x \in \mathrm{N}(0, \ \sigma_x) \qquad \text{(Equation 11)}$$

$$y_d = y_t + \varepsilon_y, \ \varepsilon_y \in \mathrm{N}\left(0, \ \sigma_y\right) \qquad \text{(Equation 12)}$$

$$\sigma_x = \begin{cases} \sigma_{x\_0}, & |x_t| = 0 \\ k_{\sigma_x}|x_t| + \sigma_{x\_0}, & |x_t| > 0 \end{cases} \qquad \text{(Equation 13)}$$

$$\sigma_y = \begin{cases} \sigma_{y\_0}, & |y_t| = 0 \\ k_{\sigma_y}|y_t| + \sigma_{y\_0}, & |y_t| > 0 \end{cases} \qquad \text{(Equation 14)}$$

where x_d and y_d are the relative positions of vehicles with error disturbances; x_t and y_t denote the true relative positions; ε_x and ε_y are the random errors sampled from Gaussian distributions with zero mean error; σ_x and σ_y are the variances; and σ_{x_0}, σ_{y_0}, k_{σ_x}, and k_{σ_y} are parameters related to the relative position. The control error was developed as follows:

$$\ddot{v}_d = \ddot{v}_{x\max} + \varepsilon_{\ddot{v}}, \ \varepsilon_{\ddot{v}} \in \mathrm{N}\left(0, \ \sigma_{\ddot{v}}\right) \qquad \text{(Equation 15)}$$

$$\dot{\delta}_d = \dot{\delta}_{\max} + \varepsilon_{\dot{\delta}}, \ \varepsilon_{\dot{\delta}} \in \mathrm{N}\left(0, \ \sigma_{\varepsilon_{\dot{\delta}}}\right) \qquad \text{(Equation 16)}$$

where $\ddot{v}_d$ represents the change rate of gas or brake pedal position (i.e., the jerk) with disturbances; $\dot{\delta}_d$ is the change rate of the steering angle with disturbances; $\varepsilon_{\ddot{v}}$ and $\varepsilon_{\dot{\delta}}$ are the random errors sampled from Gaussian distributions with zero mean error; and $\sigma_{\ddot{v}}$ and $\sigma_{\varepsilon_{\dot{\delta}}}$ are the variances.

The time delay contained the perception time T_p and the decision time T_d. The former referred to the time to percept the positions and velocities of surrounding vehicles using onboard sensors, while the latter was defined as the time to make decisions based on onboard computing resources. We assumed that time delays with the S3-IRM algorithm are marginal and almost negligible in the future because of the availability of V2VC and cloud computing platforms. The above parameters regarding errors and time delays were determined according to the state-of-the-art technologies (Table S6) (Lu et al., 2021; Yeong et al., 2021; Lin et al., 2019; Choi et al., 2018).

Technical details of the IRM algorithm

The discretized vehicle safety decision space U comprises five acceleration decisions (i.e., acceleration a_{acc}, partial acceleration a_{p-acc}, constant-speed, partial deceleration a_{p-dec}, and deceleration a_{dec}) and five swerve decisions (i.e., left, partial left, straight, partial right, and right). The left or right decision satisfies the steering wheel angle $\delta = \pm\delta_{\max}$, while the partial left or partial right decision satisfies $\delta = \pm\delta_{\max}/2$. The maximum steering angle changes with the vehicle velocity to ensure the stability of the vehicle dynamics.

$$\delta_{\max} = \begin{cases} \delta_{thre} + k_{\delta_{\max}}(v_{thre} - v_x), \ if\ v_x < v_{thre} \\ \delta_{thre}, \ if\ v_x \geq v_{thre} \end{cases} \quad \text{(Equation 17)}$$

where v_x represents the vehicle's longitudinal velocity and v_{thre}, $k_{\delta_{\max}}$, and δ_{thre} are constant parameters related to the vehicle's steering system.

The pseudocode and parameters are summarized in Box S1 and Table S7.

To assess the IRM algorithm's performance, we classified three application scenarios (i.e., S1, S2, and S3) considering the development stage of AVs according to four indicators as follows (Table S8):

Optimizable vehicles

Optimizable vehicles represent AV penetration levels. Since the public's acceptance of AVs is believed to be progressive, a mixed traffic system will be retained over a considerable time, i.e., AVs share the roads with human-driven vehicles (Hancock et al., 2019).

Availability of V2VC

V2VC empowers AVs to accurately perceive surrounding vehicle dynamics and intentions in real time, facilitating the decision-making of the ego vehicle.

Computing resource

Due to limited onboard computing resources, nonnegligible computational time delays can significantly threaten the quality of AV decisions, especially for the critical time window (generally in sub-seconds) before a collision. However, utilizing large-scale cloud computing platforms based on real-time communications is a promising solution.

Optimization target

Regarding the IRM algorithm's optimization target, we define the minimization of the OISS for a specific vehicle or the average OISS for all vehicles to differentiate the decision principle of ''egoism'' or ''utilitarianism''.

QUANTIFICATION AND STATISTICAL ANALYSIS

The statistical analysis and plots were performed with python.