Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamic Fragility of Informative Signaling in Mixed Autonomous–Human Traffic: A Coupled Signaling–Evolutionary Safety Model

Mohammad Reza Omidi

mromidi_91@yahoo.com

University of Bonab

Research Article

Keywords: autonomous vehicles, human-driven vehicles, mixed traffic safety, Signaling game, evolutionary game theory, bounded rationality

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11682375/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dynamic Fragility of Informative Signaling in Mixed Autonomous–Human Traffic: A Coupled Signaling–Evolutionary Safety Model

Author: Mohammad Reza Omidi
Affiliation: Department of Industrial Engineering, University of Bonab, Bonab, Iran
Corresponding Author: Mohammad Reza Omidi, mromidi_91@yahoo.com

Abstract

Autonomous vehicles (AVs) are increasingly interacting with human-driven vehicles (HVs), creating safety problems in which private information, human adaptation and technological penetration interact. Existing transportation studies have examined Signaling games, evolutionary AV–HV interaction, vehicle-to-vehicle (V2V) information and bounded rationality largely as separate mechanisms. This study develops a coupled Signaling–evolutionary safety model in which an AV privately observes its safety type, chooses whether to transmit an authenticated safety signal, and interacts with human drivers whose responses depend on the observed signal and evolve over repeated encounters. The AV penetration rate is modelled separately from the prior probability of a high-safety AV type and enters the behavioral and Signaling incentives through a penetration-dependent collision-risk structure. The static analysis derives necessary and sufficient incentive conditions for a strict separating Perfect Bayesian Equilibrium. Informative separation requires a type-specific signal-cost window, $L_A\delta_L p < c < L_A\delta_H p$, together with receiver conditions determining whether a signal can induce defensive driving. These conditions yield a closed-form lower penetration boundary. The static model is then embedded in a logit-regularised evolutionary system describing signal-conditioned human responses and type-specific Signaling adaptation. Local stability is evaluated from the spectral abscissa of the resulting four-dimensional Jacobian. Under an illustrative parameterisation, the receiver threshold is $p_R = 0.6944$, and strict separation at full penetration is feasible for $0.20 < c < 0.72$. At $p = 0.90$and $c = 0.40$, the static separating conditions hold. The associated regularised dynamics exhibit both a stable near-separating state and an unstable stationary state, demonstrating that incentive compatibility does not by itself guarantee dynamic robustness. The results provide a tractable framework for evaluating AV communication design jointly with deployment conditions and human behavioral adaptation.

Keywords: autonomous vehicles; human-driven vehicles; mixed traffic safety; Signaling game; evolutionary game theory; bounded rationality.

1. Introduction

Autonomous vehicles (AVs) are increasingly entering traffic systems in which human-driven vehicles (HVs) remain present. During this transition, safety is determined not only by the technical performance of automated systems but also by the strategic behavior of the humans and machines interacting with them. Game-theoretic analyses of road-user behavior have long shown that observed behavior can depart from formal traffic rules because road users respond strategically to one another [1,2].

The introduction of private information creates an additional safety problem. An AV may possess information about its safety-relevant capability that is not directly observable to an interacting human driver. The resulting communication problem is a classical Signaling problem: an informed sender chooses an observable action and an uninformed receiver forms beliefs before choosing a response [3]. Safety regulation has already been analysed through Signaling games under asymmetric information [4,5], demonstrating that signal costs and the incentives associated with different types can produce separating or pooling outcomes.

The evolutionary dimension has an equally long theoretical history. Communication was formally studied through evolutionary dynamics before its application to automated driving, including feasibility of communication under selection–mutation dynamics and the evolutionary stability of Signaling systems [6–8]. Bounded-rational responses can additionally be represented using quantal-response models in which agents respond probabilistically to payoff differences rather than selecting deterministic best responses [9]. Evolutionary stability and game dynamics provide the mathematical basis for modelling how advantageous strategies increase in frequency over time [10,11].

Transportation research has already brought Signaling into mixed traffic. Shao et al. [12] developed a Signaling-game-based lane-changing mechanism for mixed traffic involving autonomous and regular vehicles. Their model incorporated driver heterogeneity, used Perfect Bayesian Equilibrium and was calibrated and validated against NGSIM observations. This study is therefore a direct precedent for the use of Signaling games in mixed AV–conventional traffic and must be distinguished explicitly from the present contribution. Bitar et al. [13] subsequently analysed discretionary communication among road users in a non-cooperative game-theoretic framework.

Recent behavioral research provides further evidence that human responses to automated vehicles are conditional rather than fixed. Zhang et al. [14] found that AV driving style can alter human manoeuvres in mixed traffic, including riskier responses to aggressive automated vehicles. Guo et al. [15] formulated a game-theoretic driver steering model incorporating individual risk perception, while Panahandeh et al. [16] used Bayesian and Stackelberg reasoning to address perception uncertainty and right-of-way constraints. Xie et al. [17] developed an adaptive non-cooperative game for shared-control and automated-vehicle interaction.

Safety consequences are also sensitive to vehicle type and traffic composition. A matched case-control study identified differences in crash circumstances between automated and human-driven vehicles [18]. Hou et al. [19] developed a surrogate safety measure and behavioral modelling framework for mixed traffic, while Rajia et al. [20] used Lyft and Waymo data, calibrated a Wiedemann 99 car-following model and analysed mixed-traffic crash risk with a dynamic Bayesian network. Li et al. [21] recently used evolutionary game theory to model long-run AV–HV strategic interaction and incorporated heterogeneous human preferences. Oikonomou and Yannis [22] found nonlinear changes in crash-hotspot probabilities as automation penetration increased. Hu and Lee [23] showed that vehicle-pair composition affects car-following behavior in mixed traffic, while Zhang et al. [24] demonstrated that degradation of connected automated vehicles can alter traffic safety.

Communication quality provides an additional behavioral mechanism. Lee and Pitts [25] experimentally showed that vehicle reliability, confidence information and repeated interaction affect driver trust and takeover behavior. Kim et al. [26] found that structured ADAS training affects driver understanding, confidence, trust and appropriate system use. Chen and Zhang [27] found that V2V communication can reduce driver confusion and aggressive responses toward defensive AVs. Ye et al. [28] showed experimentally that perfect, interrupted and lost C-V2X communication conditions generate different driver responses. Yu et al. [29] further demonstrated that information-assisted takeover modalities and information levels affect driver takeover performance.

These findings reveal an important modelling problem. Transportation Signaling studies have established that safety information can affect human responses [12,13], while evolutionary AV studies have established that interaction strategies can change over time [21]. However, these two perspectives are normally treated separately. Similarly, studies of evolutionary Signaling have primarily examined abstract sender–receiver systems rather than mixed autonomous–human traffic [6–8,30,37]. Recent V2V information-design research has demonstrated that information can have endogenous behavioral consequences [31,32], but does not use a private AV safety type and a type-specific separating equilibrium as the basis for a coupled evolutionary traffic system.

The present research addresses this intersection.

The objective is not to claim that Signaling games, evolutionary games or AV interaction games are individually new. Rather, the study investigates whether the following mechanisms can be represented within a single safety model:

private AV safety type → signal → human belief → signal-conditioned behavior
→ Signaling incentive.

AV penetration provides a second structural channel because it changes the safety consequence of behavioral choices.

The study addresses three research questions:

RQ1. Under what conditions can a high-safety AV signal its type while a lower-safety AV does not imitate the signal?

RQ2. How does AV penetration affect both receiver behavior and the incentive to signal?

RQ3. When informative Signaling is statically feasible, under what conditions can its dynamically represented behavioral continuation become unstable?

The central contribution is consequently a **coupled safety mechanism**, rather than a new isolated game-theoretic concept. In the literature reviewed for this study, no published transportation study was identified that simultaneously combines private AV safety type, type-dependent Signaling, signal-conditioned evolutionary human response, penetration-dependent Signaling incentives and local dynamic stability analysis in one model.

2. Historical positioning and theoretical foundations

2.1 From Signaling to evolutionary Signaling

The theoretical origin of costly Signaling can be traced to Spence's model of Signaling under asymmetric information [3]. Evolutionary game theory subsequently formalised the dynamics through which strategies can change in a population [10,11].

For Signaling specifically, Hofbauer and Huttegger [6] studied the feasibility of communication in binary Signaling games. Huttegger et al. [7] examined evolutionary and learning dynamics of Signaling games, while Hofbauer and Huttegger [8] developed selection–mutation dynamics for Signaling. Wagner [37] showed that adaptive Signaling systems can display persistent non-equilibrium behavior. De Silva and Sigmund [30] provide a modern synthesis of sender–receiver games under adaptive, replicator and best-response dynamics.

These studies establish an important point for the present work: static Signaling equilibria and dynamic Signaling behavior are conceptually distinct.

2.2 Evolutionary adaptation in safety systems

Evolutionary game theory has also been applied extensively to safety management. Li et al. [33] developed a tripartite evolutionary model of occupational safety and health governance. Liu et al. [34] used evolutionary games to explain periodic fluctuations in coal-mine safety supervision. Chen et al. [35] recently applied evolutionary game theory to emergency resource sharing in chemical industry parks. Xu et al. [36] combined evolutionary game theory with simulation to study dynamic safety regulation.

These studies support the use of evolutionary adaptation as a safety mechanism, but their strategic actors and information structures differ from the AV–HV communication problem considered here.

2.3 Signaling and mixed traffic

The closest prior transportation precedent is Shao et al. [12]. Their Signaling-game model addressed discretionary lane-changing under mixed traffic, included driver heterogeneity, used PBE and was calibrated using NGSIM data. The present model differs in its object and mechanism. It treats the AV as the informed sender, assigns it a private safety type, models Signaling as a safety-relevant information action, and then allows human response strategies and type-specific Signaling probabilities to evolve jointly.

2.4 Information design and behavioral feedback

Gould and Brown [31] analysed partial adoption of V2V hazard communication and demonstrated that information provision can produce undesirable equilibrium consequences. Their subsequent work [32] explicitly examined rationality and behavioral feedback in V2V information systems. The implication is important for the current study: the effect of information cannot be evaluated independently of the behavior it induces.

The present model extends this logic in a different direction by introducing private AV safety type, type-specific Signaling costs and an evolutionary adaptation process.

3. Model formulation

3.1 Players and information structure

The model consists of an AV sender and an HV receiver.

Each AV privately observes a safety type

$$\theta \in \{H, L\}, \qquad (1)$$

where Hdenotes a high-safety type and Ldenotes a lower-safety type.

The prior probability of the high-safety type is

$$q = \Pr(\theta = H),\ 0 < q < 1. \qquad (2)$$

The AV penetration rate is separately defined as

$$p \in [0, 1]. \qquad (3)$$

The distinction

$$q \neq p$$

is fundamental: $q$describes uncertainty about vehicle type, whereas $p$describes the prevalence of AVs in the traffic environment.

The AV chooses a binary signal

$$\sigma \in \{1, 0\}, \qquad (4)$$

where $\sigma = 1$denotes an authenticated safety-related signal and $\sigma = 0$denotes no signal.

The human driver chooses

$$s \in \{D, A\}, \qquad (5)$$

where Ddenotes defensive driving and Adenotes aggressive driving.

Because the driver's response may depend on whether a signal has been observed, define

$$x_1(t) = \Pr(D \mid \sigma = 1), \qquad (6)$$

and

$$x_0(t) = \Pr(D \mid \sigma = 0). \qquad (7)$$

For each AV type,

$$y_H(t) = \Pr(\sigma = 1 \mid \theta = H), \qquad (8)$$

and

$$y_L(t) = \Pr(\sigma = 1 \mid \theta = L). \qquad (9)$$

The complete dynamic state is therefore

$$z(t) = \begin{pmatrix} x_1(t) \\ x_0(t) \\ y_H(t) \\ y_L(t) \end{pmatrix}. \qquad (10)$$

The variables $x_1$and $x_0$are conditional behavioral shares. They should not be interpreted as two physically separate human populations; rather, they represent the population's response composition conditional on the information state.

3.2 Penetration-dependent collision risk

The model uses reduced-form safety functions to isolate the strategic component of collision risk.

Let

$$r_{DH} = \bar{r}_{DH},\ r_{DL} = \bar{r}_{DL}$$

denote baseline collision probabilities under defensive human behavior.

Aggressive-response risk is defined as

$$r_{AH}(p) = \bar{r}_{DH} + \delta_H p, \qquad (11)$$

and

$$r_{AL}(p) = \bar{r}_{DL} + \delta_L p, \qquad (12)$$

where

$$\delta_H > \delta_L > 0. \qquad (13)$$

The quantity $p$is therefore an environmental exposure term that captures how the prevalence of automated vehicles changes the strategic consequences of human behavior.

These functions are reduced-form theoretical representations, not empirical crash-rate estimates.

3.3 Human utility

Let $L_H > 0$denote the normalised human loss associated with a collision and let

$$\Delta E = E_A - E_D > 0 \qquad (14)$$

be the efficiency gain from aggressive driving.

Human utility is

$$U_H(s, \theta, p) = -L_H r_{s\theta}(p) + E_s. \qquad (15)$$

The payoff advantage of defensive behavior for type $\theta$is

$$\Delta u_\theta(p) = U_H(D, \theta, p) - U_H(A, \theta, p), \qquad (16)$$

which gives

$$\Delta u_\theta(p) = L_H \delta_\theta p - \Delta E. \qquad (17)$$

Hence,

$$\Delta u_H(p) = L_H \delta_H p - \Delta E, \qquad (18)$$

and

$$\Delta u_L(p) = L_H \delta_L p - \Delta E. \qquad (19)$$

For a posterior belief $m_b = \Pr(H \mid \sigma = b)$, the expected advantage of defensive driving is

$$G_b(p, m_b) = m_b \Delta u_H(p) + (1 - m_b)\Delta u_L(p). \qquad (20)$$

The receiver prefers defensive driving whenever

$$G_b(p, m_b) \geq 0. \qquad (21)$$

3.4 Bayesian beliefs

For interior Signaling probabilities,

$$m_1 = \frac{q y_H}{q y_H + (1 - q) y_L}, \qquad (22)$$

and

$$m_0 = \frac{q(1-y_H)}{q(1-y_H)+(1-q)(1-y_L)}. \qquad (23)$$

Under strict separation,

$$y_H = 1,\ y_L = 0,$$

and therefore

$$m_1 = 1,\ m_0 = 0. \qquad (24)$$

Both signal states remain on the equilibrium path when $0 < q < 1$.

3.5 AV utility and Signaling benefits

Let $L_A > 0$denote the normalised collision-loss weight for the AV.

Conditional on information state $b \in \{1, 0\}$, expected collision risk for type $\theta$is

$$R_\theta(b) = x_b r_{D\theta} + (1-x_b) r_{A\theta}(p). \qquad (25)$$

The collision-risk advantage of Signaling is therefore

$$R_\theta(0) - R_\theta(1) = \delta_\theta p(x_1 - x_0). \qquad (26)$$

The corresponding utility benefit of Signaling is

$$B_\theta(x_1, x_0, p) = L_A \delta_\theta p(x_1 - x_0). \qquad (27)$$

The net Signaling incentive is

$$\Gamma_\theta = B_\theta(x_1, x_0, p) - c, \qquad (28)$$

where $c > 0$denotes Signaling cost.

Equation (27) is central because it shows that a signal has value only when it changes driver behavior.

4. Static Signaling equilibrium

4.1 Candidate separating outcome

Consider the strategy profile

$$y_H = 1,\ y_L = 0, \qquad (29)$$

and

$$x_1 = 1,\ x_0 = 0. \qquad (30)$$

Thus, the high-safety AV signals, the lower-safety AV remains silent, the receiver becomes defensive after receiving a signal and aggressive after receiving no signal.

Proposition 1

A strict separating PBE exists if and only if

$$L_H \delta_H p > \Delta E, \qquad (31)$$
$$L_H \delta_L p < \Delta E, \qquad (32)$$

and

$$L_A \delta_L p < c < L_A \delta_H p. \qquad (33)$$

Proof

Under strict separation,

$$m_1 = 1$$

and

$$m_0 = 0.$$

Therefore,

$$G_1 = L_H \delta_H p - \Delta E. \qquad (34)$$

The receiver chooses Dafter a signal when $G_1 > 0$, which gives Eq. (31).

Similarly,

$$G_0 = L_H \delta_L p - \Delta E. \qquad (35)$$

The receiver chooses Aafter no signal when $G_0 < 0$, which yields Eq. (32).

For type H,

$$\Gamma_H = L_A \delta_H p - c, \qquad (36)$$

so Signaling requires

$$c < L_A \delta_H p.$$

For type L,

$$\Gamma_L = L_A \delta_L p - c, \qquad (37)$$

and non-Signaling requires

$$c > L_A \delta_L p.$$

Combining the four strict conditions gives Eqs. (31)–(33). □

4.2 Signal-cost window

Define

$$c_L(p) = L_A \delta_L p, \qquad (38)$$

and

$$c_H(p) = L_A \delta_H p. \qquad (39)$$

Then

$$c_L(p) < c < c_H(p) \qquad (40)$$

is the informative Signaling region.

The width of this region is

$$W_c(p) = c_H(p) - c_L(p) = L_A(\delta_H - \delta_L)p. \qquad (41)$$

Hence,

$$W_c(p) > 0$$

if and only if

$$\delta_H > \delta_L. \qquad (42)$$

This condition has a direct interpretation: the high-safety type must obtain a larger behavioral benefit from Signaling than the lower-safety type.

4.3 Receiver penetration condition

From Eq. (31),

$$p > \frac{\Delta E}{L_H \delta_H}.$$

Define

$$p_R = \frac{\Delta E}{L_H \delta_H}. \qquad (43)$$

This is the minimum penetration required for a perfectly identified high-safety AV to induce defensive driving.

4.4 Complete penetration interval

The high-type Signaling condition gives

$$p > \frac{c}{L_A \delta_H}, \qquad (44)$$

while lower-type non-imitation requires

$$p < \frac{c}{L_A \delta_L}. \qquad (45)$$

The low-type receiver condition gives

$$p < \frac{\Delta E}{L_H \delta_L}. \qquad (46)$$

Consequently, define

$$\underline{p}(c) = \max\left\{\frac{\Delta E}{L_H \delta_H}, \frac{c}{L_A \delta_H}\right\}, \qquad (47)$$

and

$$\bar{p}(c) = \min\left\{1, \frac{\Delta E}{L_H \delta_L}, \frac{c}{L_A \delta_L}\right\}. \qquad (48)$$

A strict separating regime is feasible if and only if

$$\boxed{\underline{p}\,(c) < p < \bar{p}\,(c)}. \qquad (49)$$

This result is stronger than a single "critical penetration rate": it identifies the entire penetration interval over which informative separation is simultaneously credible, behavioral ly effective and resistant to imitation.

4.5 Static safety-information margin

For diagnostic purposes, define the static separation margin as

$$M_S = \min\{L_H, \delta_H, p, -\Delta E\,\Delta E - L_H\delta_L p\; c - L_A\delta_L p\; L_A\delta_H p - c\}. \qquad (50)$$

Then

$$M_S > 0$$

is equivalent to strict separation.

A small positive $M_S$indicates a statically feasible but fragile Signaling regime, while

$$M_S \leq 0$$

implies failure of at least one strict equilibrium condition.

5. Evolutionary extension

5.1 Regularised human adaptation

To represent bounded rationality, define the log-odds function

$$\ell(z) = \ln\left(\frac{z}{1-z}\right). \qquad (51)$$

The conditional evolution of defensive behavior after a signal is

$$\dot{x}_1 = x_1(1-x_1)[\kappa_H G_1(p, m_1) - \tau_H \ell(x_1)], \qquad (52)$$

where $\kappa_H > 0$denotes behavioral responsiveness and $\tau_H > 0$controls bounded rationality.

Following no signal,

$$\dot{x}_0 = x_0(1-x_0)[\kappa_H G_0(p, m_0) - \tau_H \ell(x_0)]. \qquad (53)$$

The specification follows the logit-response logic associated with quantal response equilibrium [9] while retaining an evolutionary adjustment structure [10,30].

5.2 Evolution of Signaling

Type-specific Signaling probabilities evolve according to

$$\dot{y}_H = y_H(1 - y_H)[\kappa_S(B_H - c) - \tau_S \ell(y_H)], \qquad (54)$$

and

$$\dot{y}_L = y_L(1 - y_L)[\kappa_S(B_L - c) - \tau_S \ell(y_L)], \qquad (55)$$

where $\kappa_S > 0$measures Signaling responsiveness and $\tau_S > 0$measures stochasticity in Signaling behavior.

The complete feedback structure is

$$(y_H, y_L) \to (m_1, m_0) \to (x_1, x_0) \to (B_H, B_L) \to (y_H, y_L). \qquad (56)$$

This is the model's defining coupling mechanism.

5.3 Interior stationary conditions

At an interior stationary state,

$$\frac{x_b^*}{1 - x_b^*} = \exp\left[\frac{\kappa_H G_b(p, m_b^*)}{\tau_H}\right], \qquad (57)$$

and

$$\frac{y_\theta^*}{1 - y_\theta^*} = \exp\left[\frac{\kappa_S(B_\theta - c)}{\tau_S}\right]. \qquad (58)$$

Consequently,

$$x_b^* = \frac{\exp[\kappa_H G_b/\tau_H]}{1 + \exp[\kappa_H G_b/\tau_H]}, \qquad (59)$$

and

$$y_\theta^* = \frac{\exp[\kappa_S(B_\theta - c)/\tau_S]}{1 + \exp[\kappa_S(B_\theta - c)/\tau_S]}. \qquad (60)$$

As

$$\tau_H, \tau_S \to 0^+,$$

the regularised solutions approach the corresponding deterministic best responses when the static conditions are strict.

6. Dynamic stability

6.1 Jacobian

Let

$$F(z) = \begin{pmatrix} F_1 \\ F_0 \\ F_H \\ F_L \end{pmatrix}$$

denote the right-hand side of Eqs. (52)–(55). The Jacobian is

$$J(z) = \frac{\partial F}{\partial z}, \qquad (61)$$

or explicitly,

$$J = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_0} & \frac{\partial F_1}{\partial y_H} & \frac{\partial F_1}{\partial y_L} \\ \frac{\partial F_0}{\partial x_1} & \frac{\partial F_0}{\partial x_0} & \frac{\partial F_0}{\partial y_H} & \frac{\partial F_0}{\partial y_L} \\ \frac{\partial F_H}{\partial x_1} & \frac{\partial F_H}{\partial x_0} & \frac{\partial F_H}{\partial y_H} & \frac{\partial F_H}{\partial y_L} \\ \frac{\partial F_L}{\partial x_1} & \frac{\partial F_L}{\partial x_0} & \frac{\partial F_L}{\partial y_H} & \frac{\partial F_L}{\partial y_L} \end{pmatrix}. \qquad (62)$$

The feedback between human behavior and Signaling follows directly from

$$\frac{\partial B_\theta}{\partial x_1} = L_A \delta_\theta p > 0, \qquad (63)$$

and

$$\frac{\partial B_\theta}{\partial x_0} = -L_A \delta_\theta p < 0. \qquad (64)$$

The reverse channel follows because m_1 and m_0 depend on y_H and y_L.

6.2 Spectral stability

Let

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4$$

be the eigenvalues of J^*, evaluated at a stationary state z^*.

Define the spectral abscissa as

$$\alpha(J^*) = \max_i \mathrm{Re}[\lambda_i(J^*)]. \qquad (65)$$

Proposition 2

A regularised stationary state is locally asymptotically stable if

$$\alpha(J^*) < 0, \qquad (66)$$

and locally unstable if

$$\alpha(J^*) > 0. \qquad (67)$$

Proof

Linearisation around the stationary point gives

$$\dot{u} = J^* u + o(\| u \|).$$

The local behavior is determined by the eigenvalues of J^*. If every eigenvalue has a negative real part, all sufficiently small perturbations decay. If at least one eigenvalue has a positive real part, perturbations grow along the corresponding eigendirection. Therefore, the signs of the spectral abscissa determine local asymptotic stability and instability. □

6.3 Feasibility versus instability

The model makes a necessary conceptual distinction:

$$\boxed{\text{static equilibrium failure} \neq \text{dynamic instability}}. \qquad (68)$$

Static failure occurs when

$$M_S \leq 0,$$

whereas dynamic instability occurs when

$$\alpha(J^*) > 0.$$

Therefore, a statically feasible Signaling regime may still possess dynamically unstable stationary states.

7. Numerical analysis

7.1 Illustrative parameterisation

The numerical analysis is explicitly theoretical. The parameters do not represent empirical crash-rate calibration.

Table 1:Illustrative parameterisation

Parameter	Value	Interpretation
$\bar{r}_{DH}$	0.008	Baseline collision probability, high type, defensive response
$\bar{r}_{DL}$	0.040	Baseline collision probability, lower type, defensive response
δ_H	0.072	Incremental aggressive-response risk for type Hat $p = 1$
δ_L	0.020	Incremental aggressive-response risk for type Lat $p = 1$
L_H	10	Human collision-loss weight
L_A	10	AV collision-loss weight
ΔE	0.50	Efficiency gain from aggressive driving
q	0.80	Prior probability of high-safety AV
κ_H	0.80	Human behavioral responsiveness
κ_S	0.60	Signaling responsiveness
p	0–1	AV penetration
c	0–0.80	Signal cost

7.2 Receiver threshold

Using Eq. (43),

$$p_R = \frac{0.50}{10(0.072)} = 0.6944. \qquad (69)$$

Therefore,

$$p > 0.6944$$

is required for the high-safety signal to induce defensive driving under perfect type identification.

For the lower-safety type,

$$\Delta u_L(p) = 10(0.020)p - 0.50 = 0.20p - 0.50. \qquad (70)$$

Since

$$0.20p - 0.50 \leq -0.30$$

for all

$$p \leq 1,$$

the low-type receiver condition remains satisfied throughout the feasible penetration range.

7.3 Signal-cost window

At full penetration,

$$c_L(1) = 10(0.020) = 0.20, \qquad (71)$$

and

$$c_H(1) = 10(0.072) = 0.72. \qquad (72)$$

Hence,

$$\boxed{0.20 < c < 0.72}. \qquad (73)$$

Table 2:Static Signaling regimes at $p = 1$

Signal cost	Static implication
$0 < c < 0.20$	Both types have an incentive to signal; pooling tendency
$0.20 < c < 0.72$	Strict separation is incentive-compatible
$c > 0.72$	High-safety type does not find Signaling worthwhile

This result shows why communication cost has a non-monotonic relationship with information quality.

7.4 Penetration sensitivity

For

$$c = 0.40,$$

the high-type sender condition requires

$$p > \frac{0.40}{10(0.072)} = 0.5556. \qquad (74)$$

The receiver condition remains stronger:

$$p > 0.6944.$$

Thus,

$$p_{\min}(0.40) = 0.6944. \qquad (75)$$

Table 3.Penetration sensitivity for $c = 0.40$

p	High-type receiver advantage	$c_L(p)$	$c_H(p)$	Separation
0.50	−0.140	0.100	0.360	No
0.60	−0.068	0.120	0.432	No
0.694	≈0	0.139	0.500	Boundary
0.75	0.040	0.150	0.540	Yes
0.90	0.148	0.180	0.648	Yes
1.00	0.220	0.200	0.720	Yes

The numerical threshold is conditional on the model's parameterization and is not a universal AV deployment target.

7.5 Static feasibility at the main dynamic test point

Consider

$$q = 0.80,\ p = 0.90,\ c = 0.40. \qquad (76)$$

The receiver conditions are

$$10(0.072)(0.90) - 0.50 = 0.148 > 0, \qquad (77)$$

and

$$10(0.020)(0.90) - 0.50 = -0.32 < 0. \qquad (78)$$

The sender benefits are

$$B_H(1, 0, 0.90) = 10(0.072)(0.90) = 0.648, \qquad (79)$$

and

$$B_L(1, 0, 0.90) = 10(0.020)(0.90) = 0.180. \qquad (80)$$

Hence,

$$0.180 < 0.400 < 0.648. \qquad (81)$$

The strict separating outcome is therefore statically feasible.

The associated static margin from Eq. (50) is

$$M_S = \min\{0.148, \ , 0.320, \ 0.220 \ 0.248\} = 0.148. \qquad (82)$$

Thus, the equilibrium is not only feasible but has a positive analytical safety-information margin.

7.6 Stable near-separating state

For

$$\tau_H = 0.03, \ \tau_S = 0.025, \qquad (83)$$

the regularized system produces the stationary state

$$z_S^* = (0.980774, \ , 0.000234, \ 0.996493 \ 0.004660). \qquad (84)$$

The Jacobian eigenvalues are

$$\lambda(J_S^*) = \{-, 0.029921, \ , -0.029992 \ -0.025083 \ -0.025000\}. \qquad (85)$$

Therefore,

$$\alpha(J_S^*) = -0.025000 < 0, \qquad (86)$$

indicating local asymptotic stability.

The state is close to strict separation:

$$x_1^* \approx 0.981,$$
$$x_0^* \approx 0,$$
$$y_H^* \approx 0.996,$$

and

$$y_L^* \approx 0.005.$$

7.7 Coexisting unstable stationary state

Under the same structural parameters,

$$q = 0.80,\ p = 0.90,\ c = 0.40,\ \kappa_H = 0.80,\ \kappa_S = 0.60,$$

a second stationary state is obtained as

$$z_U^* = (0.980940,\ ,0.345522,\ 0.570040\ 0.001053). \qquad (87)$$

Its eigenvalues are approximately

$$\lambda(J_U^*) = \{0.038572,\ ,-,0.093572\ -0.030000\ -0.025000\}. \qquad (88)$$

Hence,

$$\alpha(J_U^*) = 0.038572 > 0. \qquad (89)$$

The stationary state is therefore locally unstable.

This state is not labelled a separating PBE. It is a regularized stationary point of the coupled dynamic system. This distinction is essential because the static equilibrium concept and the dynamic system answer different questions.

Table 4

Stable and unstable stationary states under the same structural environment

State	x_1	x_0	y_H	y_L	$\alpha(J)$	Classification
Near-separating	0.980774	0.000234	0.996493	0.004660	−0.025000	Stable
Interior	0.980940	0.345522	0.570040	0.001053	+0.038572	Unstable

The coexistence of these states shows that local dynamic robustness is not determined solely by the static Signaling inequalities.

8. Safety implications

The expected interaction-level collision probability under strict separation is

$$R_{\mathrm{sep}}(q,p) = q\bar{r}_{DH} + (1-q)[\bar{r}_{DL} + \delta_L p]. \qquad (90)$$

If all AVs pool on no signal and human drivers respond aggressively, the corresponding risk is

$$R_{\mathrm{pool},0}(q,p) = q[\bar{r}_{DH} + \delta_H p] + (1-q)[\bar{r}_{DL} + \delta_L p]. \qquad (91)$$

At

$$p = 1,$$

the model gives:

Table 5.Illustrative aggregate risk consequences

q	R_{sep}	$R_{\mathrm{pool},0}$	Relative reduction
0.20	0.0496	0.0640	22.5%

0.50	0.0340	0.0700	51.4%
0.80	0.0184	0.0760	75.8%

These values are generated by the theoretical model and are not observed crash rates.

The safety interpretation is straightforward. Informative Signaling can reduce the exposure generated by aggressive human responses because it allows drivers to distinguish vehicles for which defensive behavior has greater value. However, the signal is useful only if it remains discriminating.

9. Discussion

9.1 What is genuinely new

The analysis does not claim novelty for Signaling games themselves, evolutionary games themselves, or AV–HV game theory. Each has substantial historical precedent [3,6–13,21,30–32].

The contribution lies in the coupling

$$\boxed{\begin{array}{c}\text{private type} + \text{costly signal} + \text{conditional behavioral evolution} + \text{penetration-dependent risk} \\ + \text{dynamic stability}\end{array}}. \tag{92}$$

Shao et al. [12] provide the closest Signaling precedent in mixed traffic, but their model does not formulate the AV as a privately typed safety sender whose Signaling strategy co-evolves with signal-conditioned human behavior. Li et al. [21] provide a close evolutionary AV–HV precedent but do not introduce this Signaling-information structure. Gould and Brown [31,32] show that information can interact with driver behavior and partial communication adoption but use a different information-design formulation. Evolutionary Signaling studies [6–8,30,37] establish the dynamic theory but not the mixed AV–HV safety application developed here.

This positioning avoids an overly broad novelty claim while preserving a specific methodological contribution.

9.2 Why the Signaling threshold is an interval

The sender side generates two incentive constraints:

$$B_L < c$$

and

$$c < B_H.$$

These constraints are qualitatively different. The first prevents imitation by the lower-safety type. The second prevents abandonment by the high-safety type.

The result is

$$B_L < c < B_H.$$

The communication-cost problem is consequently not monotone.

This finding is particularly relevant to safety certification. Minimising communication cost may actually reduce informational value if low-safety systems can freely imitate the signal.

9.3 Why penetration is not a universal number

The model generates

$$p_R = \frac{\Delta E}{L_H \delta_H}.$$

The value depends on:

$$\Delta E,\ L_H,\ \delta_H.$$

Therefore, the illustrative value

$$p_R = 0.6944$$

should not be interpreted as a universal empirical threshold.

This is consistent with empirical mixed-traffic research showing that safety effects vary with vehicle composition and traffic context [20–24]. Oikonomou and Yannis [22], for example, report nonlinear hotspot patterns across automation penetration levels, while Hu and Lee [23] demonstrate that the interaction structure of different vehicle pairs can change as AVs enter the traffic stream.

9.4 Why static communication can be dynamically fragile

The main dynamic result is

$$M_S > 0$$

does not imply

$$\alpha(J) < 0.$$

The first condition establishes static feasibility; the second establishes local dynamic stability.

The numerical results show both behaviors in the same structural environment. A near-separating state is stable, while another stationary state has a positive eigenvalue.

This has a direct safety-science interpretation: **a technically feasible communication policy may still be behavioral ly fragile**.

9.5 Relationship to empirical V2V findings

The quantity

$$x_1 - x_0$$

has an empirical interpretation as the behavioral separation produced by communication.

When

$$x_1 - x_0$$

is large, communication materially changes human behavior and the Signaling benefit $B_\theta$is large.

When

$$x_1 \approx x_0,$$

the message is behavioral ly ineffective even when transmission reliability is high.

This mechanism is consistent with recent evidence that V2V communication can reduce driver confusion [27], that imperfect communication can alter decisions [28], and that information modality influences takeover performance [29].

9.6 Socio-technical safety interpretation

The model represents the AV–human interaction as a feedback system rather than as an isolated technical problem:

AV capability → information → human interpretation → behavior → risk → Signaling incentive. (93

Such a perspective is consistent with hazard-analysis approaches that treat AI-enabled safety-critical systems as joint human–technical systems [30].

10. Practical implications

The results support three safety-policy implications.

First, AV communication standards should evaluate **informativeness** rather than transmission reliability alone. A signal transmitted by every vehicle may be reliable but uninformative.

Second, communication policy should be evaluated jointly with AV deployment. Equations (47)–(49) show that the feasible penetration interval depends on both signal cost and behavioral incentives.

Third, safety certification should consider human response. Reliability, training and information quality have measurable effects on trust and driver behavior [25,26,29]. A safety communication standard that ignores those responses may overestimate its practical effect.

11. Limitations and future research

The model has several limitations.

First, the collision-risk functions are reduced-form theoretical relationships. Future research should estimate $\bar{r}_{D\theta}$ and δ_θ from naturalistic driving data, simulator experiments or connected-vehicle trajectories.

Second, the communication channel is binary. Actual automated vehicles may transmit continuous information concerning intent, confidence, braking capability, predicted trajectory and hazard severity.

Third, q and p are exogenous. A richer model could make vehicle adoption endogenous.

Fourth, the human response model uses conditional population shares rather than explicit psychological classes. Future work could introduce cautious, opportunistic and risk-tolerant driver types.

Fifth, finite-population and network effects are not modelled. Evolutionary Signaling research shows that mutation, population structure and learning mechanisms can materially affect long-run outcomes [6–8,30,37].

Sixth, the numerical parameterisation is illustrative. It should not be interpreted as crash-rate calibration.

Finally, the present model establishes local dynamic stability through the Jacobian. Global basin structure, bifurcations and stochastic finite-population effects require additional analysis.

12. Conclusions

This paper developed a coupled Signaling–evolutionary safety model for interactions between autonomous and human-driven vehicles under asymmetric information.

The static analysis showed that informative Signaling requires

$$\boxed{L_A\delta_L p < c < L_A\delta_H p}, \qquad (94)$$

which represents a type-specific cost window rather than a single prior-weighted threshold.

The receiver condition is

$$\boxed{p > \frac{\Delta E}{L_H\delta_H}}, \qquad (95)$$

and the complete penetration region is defined by

$$\boxed{\underline{p}(c) < p < \bar{p}(c)}. \qquad (96)$$

Under the illustrative parameterization,

$$p_R = 0.6944$$

and

$$0.20 < c < 0.72$$

at full penetration.

At

$$p = 0.90,\ c = 0.40,$$

all static separating inequalities are satisfied and the static separation margin is

$$M_S = 0.148.$$

The dynamic extension yields a four-dimensional coupled system. The near-separating state has

$$\alpha(J) = -0.025000,$$

whereas a second stationary state has

$$\alpha(J) = 0.038572.$$

Thus, informative Signaling can be statically feasible while the associated dynamic system contains locally unstable behavior.

The broader safety conclusion is therefore that AV communication should not be assessed solely in terms of whether messages are transmitted or whether a static equilibrium exists. Communication effectiveness depends on the interaction between signal cost, vehicle type, AV penetration and human adaptation.

The model provides an analytical framework for evaluating these mechanisms jointly. Its next empirical step should be calibration against trajectory-level mixed-traffic data and controlled V2V experiments.

Declarations

Ethics approval and consent to participate: Not applicable. This study used mathematical modeling and previously published aggregate data and did not recruit human participants or analyze identifiable individual-level data.

Consent for publication: Not applicable.

Competing interests: The author declares that he has no competing interests.

Funding: No external funding was received for this study.

Authors' contributions: Mohammad Reza Omidi: Conceptualization, methodology, mathematical modeling, formal analysis, and writing—original draft. Methodology, interpretation, and writing—review and editing. Supervision, validation, and writing—review and editing. The author approved the final manuscript and agrees to be accountable for all aspects of the work.

Acknowledgements: The author acknowledges the researchers whose published studies on mixed autonomous–human traffic, Signaling games, evolutionary game theory, and V2V communication provided the theoretical and empirical context for the present model.

Data availability: The mathematical model, equations, parameter definitions, and baseline parameter values are fully reported in this article. No individual-level participant dataset was generated. Numerical simulation code can be made available by the corresponding author upon reasonable request.

Use of artificial intelligence: During preparation of the manuscript, generative AI-assisted tools were used for language editing, structural refinement and mathematical consistency checking. The author reviewed the resulting manuscript and takes full responsibility for its scientific content, mathematical derivations and final version.

References

[1] R. Elvik, A review of game-theoretic models of road user behavior, Accid. Anal. Prev. 62 (2014) 388–396. https://doi.org/10.1016/j.aap.2013.06.016.

[2] T. Bjørnskau, The Zebra Crossing Game – Using game theory to explain a discrepancy between road user behavior and traffic rules, Saf. Sci. 92 (2017) 298–301. https://doi.org/10.1016/j.ssci.2015.10.007.

[3] M. Spence, Job market signaling, Q. J. Econ. 87 (1973) 355–374. https://doi.org/10.2307/1882010.

[4] Q. Zhou, Q. Wang, Assessing the safety regulatory process of compliance-based paradigm in China using a Signaling game model, Saf. Sci. 126 (2020) 104678. https://doi.org/10.1016/j.ssci.2020.104678.

[5] X. Yang, Y. Tian, K. Feng, J. Yang, S.-h. Zhang, S. Wang, Signal game analysis on the effectiveness of coal mine safety supervision based on the affective events theory, Complexity (2020) 5710419. https://doi.org/10.1155/2020/5710419.

[6] J. Hofbauer, S.M. Huttegger, Feasibility of communication in binary signaling games, J. Theor. Biol. 254 (2008) 843–849. https://doi.org/10.1016/j.jtbi.2008.07.010.

[7] S.M. Huttegger, B. Skyrms, P. Tarrès, E.O. Wagner, Some dynamics of signaling games, Proc. Natl. Acad. Sci. U.S.A. 111 (Suppl. 3) (2014) 10873–10880. https://doi.org/10.1073/pnas.1400838111.

[8] J. Hofbauer, S.M. Huttegger, Selection-mutation dynamics of signaling games, Games 6 (2015) 2–31. https://doi.org/10.3390/g6010002.

[9] R.D. McKelvey, T.R. Palfrey, Quantal response equilibria for normal form games, Games Econ. Behav. 10 (1995) 6–38. https://doi.org/10.1006/game.1995.1023.

[10] P.D. Taylor, L.B. Jonker, Evolutionary stable strategies and game dynamics, Math. Biosci. 40 (1978) 145–156. https://doi.org/10.1016/0025-5564(78)90077-9.

[11] J.W. Weibull, Evolutionary Game Theory, MIT Press, Cambridge, MA, 1995.

[12] H. Shao, M. Zhang, T. Feng, Y. Dong, A discretionary lane-changing decision-making mechanism incorporating drivers' heterogeneity: a Signaling game-based approach, J. Adv. Transp. 2020 (2020) 8892693. https://doi.org/10.1155/2020/8892693.

[13] I. Bitar, A. Solernou Crusat, R. Romano, D. Watling, To signal or not to signal? A non-cooperative game-theoretic approach to discretionary communication between road users, SN Comput. Sci. 7 (2026) 3. https://doi.org/10.1007/s42979-025-04533-w.

[14] Y. Zhang, S. Wu, D. Mubbala, N. Du, Human-machine cooperation strategies in game-theoretic scenarios within mixed traffic: a simulator study on driving styles, Accid. Anal. Prev. 223 (2025) 108254. https://doi.org/10.1016/j.aap.2025.108254.

[15] W. Guo, J. Li, X. Song, W. Zhang, A game-theoretic driver steering model with individual risk perception field generation, Accid. Anal. Prev. 211 (2025) 107869. https://doi.org/10.1016/j.aap.2024.107869.

[16] P. Panahandeh, A.R. Alghooneh, M. Pirani, B. Fidan, A. Khajepour, Game-theoretic motion planning with perception uncertainty and right-of-way constraints, Sensors 24 (2024) 8177. https://doi.org/10.3390/s24248177.

[17] W. Xie, S. Lu, L. Dai, R. Jia, Interactive strategy of shared control vehicle and automated vehicles based on adaptive non-cooperative game, Proc. Inst. Mech. Eng. Part D J. Automob. Eng. 239 (2025) 5650–5663. https://doi.org/10.1177/09544070241285144.

[18] M. Abdel-Aty, S. Ding, A matched case-control analysis of autonomous vs human-driven vehicle accidents, Nat. Commun. 15 (2024) 4931. https://doi.org/10.1038/s41467-024-48526-4.

[19] K. Hou, F. Zheng, X. Liu, Enhancing mixed traffic safety assessment: a novel safety metric combined with a comprehensive behavioral modeling framework, Accid. Anal. Prev. 208 (2024) 107766. https://doi.org/10.1016/j.aap.2024.107766.

[20] S. Rajia, H. Naka, Y. Muromachi, M. Hossain, Crash risk assessment in a mixed traffic environment of autonomous and human-driven vehicles, J. East. Asia Soc. Transp. Stud. 16 (2025) PP4065. https://doi.org/10.11175/easts.16.PP4065.

[21] R. Li, Y. Liu, J. Sun, Y. Tian, Cooperative or competitive? Resolving social dilemmas in autonomous vehicles through evolutionary game theory, Accid. Anal. Prev. 228 (2026) 108402. https://doi.org/10.1016/j.aap.2026.108402.

[22] M.G. Oikonomou, G. Yannis, Spatial dynamics of crash hotspots under autonomous vehicle adoption scenarios, Accid. Anal. Prev. 233 (2026) 108562. https://doi.org/10.1016/j.aap.2026.108562.

[23] J. Hu, J.J. Lee, Car following dynamics in mixed traffic flow of autonomous and human-driven vehicles: complex networks approach, Physica A 665 (2025) 130519. https://doi.org/10.1016/j.physa.2025.130519.

[24] M. Zhang, N. Uno, X. Yang, Traffic safety evaluation for mixed traffic flow caused by degradation of connected automated vehicles, Physica A 659 (2025) 130353. https://doi.org/10.1016/j.physa.2025.130353.

[25] M. Lee, B.J. Pitts, The effects of automated vehicle reliability, self-estimated confidence, and repeated interactions on drivers' trust and takeover decisions, Saf. Sci. 195 (2026) 107062. https://doi.org/10.1016/j.ssci.2025.107062.

[26] S. Kim, S. Calvert, M.P. Hagenzieker, Exploring ADAS driver training in driving academies: perspectives from driving instructors, Saf. Sci. 194 (2026) 107042. https://doi.org/10.1016/j.ssci.2025.107042.

[27] H. Chen, Y. Zhang, Mitigating driver confusion and aggression in mixed traffic: effects of V2V communication and driving style, Accid. Anal. Prev. 236 (2026) 108676. https://doi.org/10.1016/j.aap.2026.108676.

[28] S. Ye, T. Chen, O. Oviedo-Trespalacios, Y. Ali, T. Oh, I. Kim, How do the drivers react to different C-V2X-based communication conditions in dilemma zones? A driving simulator study, Accid. Anal. Prev. 221 (2025) 108208. https://doi.org/10.1016/j.aap.2025.108208.

[29] Z. Yu, L. Zhai, K. Jiang, S. Yu, B. Zhang, Q. Deng, Z. Huang, Impact of information-assisted modalities and levels on driver takeover performance in conditional automated driving, Accid. Anal. Prev. 220 (2025) 108158. https://doi.org/10.1016/j.aap.2025.108158.

[30] M.L. Cummings, A taxonomy for AI hazard analysis, J. Cogn. Eng. Decis. Mak. 18 (2024) 327–332. https://doi.org/10.1177/15553434231224096.

[31] B.T. Gould, P.N. Brown, On partial adoption of vehicle-to-vehicle communication: when should cars warn each other of hazards? in: Proc. 2022 American Control Conference, IEEE, Atlanta, GA, 2022, pp. 627–632. https://doi.org/10.23919/ACC53348.2022.9867165.

[32] B.T. Gould, P.N. Brown, Rationality and behavior feedback in a model of vehicle-to-vehicle communication, in: Proc. 62nd IEEE Conference on Decision and Control, IEEE, 2023. https://doi.org/10.1109/CDC49753.2023.10384214.

[33] S. Li, J. Li, Y. Wang, S. Xue, Study on the evolution of incentive mechanisms for multibody collaborative governance of occupational safety and health in China, Saf. Sci. 187 (2025) 106867. https://doi.org/10.1016/j.ssci.2025.106867.

[34] D. Liu, X. Xiao, H. Li, W. Wang, Historical evolution and benefit–cost explanation of periodical fluctuation in coal mine safety supervision: an evolutionary game analysis framework, Eur. J. Oper. Res. 243 (2015) 974–984. https://doi.org/10.1016/j.ejor.2014.12.046.

[35] C. Chen, Y. Zhao, X. Zhang, Emergency resource sharing in chemical industry parks: an evolutionary game perspective, Saf. Sci. 201 (2026) 107271. https://doi.org/10.1016/j.ssci.2026.107271.

[36] Y. Xu, L. Yang, J. Zhu, Evolutionary game and simulation analysis of production safety regulation in chemical enterprises, SAGE Open 15 (2025) 1–20. https://doi.org/10.1177/21582440251393450.

[37] E.O. Wagner, Deterministic chaos and the evolution of meaning, Br. J. Philos. Sci. 63 (2012) 547–575. https://doi.org/10.1093/bjps/axr039.