MAINTAINING DRIVER AUTONOMY WITH COLLISION AVOIDANCE SYSTEMS: A SATISFICING APPROACH

Michael A. Goodrich and Erwin R. Boer

Cambridge Basic Research, Nissan Research and Development, Inc.

Cambridge, MA USA

Keywords: advanced safety vehicles, satisficing, collision warning systems

ABSTRACT

The decision to invoke a collision and accident avoidance system must account for not only the capabilities of safety-enhancing technologies, but also the autonomy and preferences of the human driver. In the absence of a general theory of human interaction with complex systems, it is difficult to define and find an optimal resolution to these competing design requirements. Instead we develop a satisficing decision system which employs both requirements by comparing the safety benefit of a collision avoidance action against the cost to the driver to identify when to assist the driver. We illustrate the design procedure with a lane departure example.

INTRODUCTION

Collision and accident avoidance systems (CAAS's) are an important component of advanced safety vehicles and may be realized with minimal or no changes to existing vehicles and highway infrastructure [1]. As a result of much current academic and industrial research, the complicated technological and human factors associated with CAAS design are being unraveled which enhances the desire to include CAAS's in vehicle design [2, 3]. However, a lesson learned from process automation is that, in the absence of human factors considerations, even technologically state-of-the-art systems can be more problematic than beneficial [4, 5, 6]. This lesson indicates the importance of including human factors in CAAS design so as to prevent "ironies of automation" [4]. Consequently, it is desirable to design for the complete system, which consists of both CAAS technology as well as a human operator. We use the term *human-CAAS* system to emphasize the human operator "in the loop" [7].

In the absence of a general theory of human interaction with complex systems, it is difficult to define and find optimal human-CAAS solutions [8]. We present a design method which replaces the objective of optimal design with the objective of *avoiding error* and illustrate its application to a lane departure example. This design paradigm employs the satisficing principle of Simon [9, 10], the domination principle from multi-attribute utility theory (MAUT), and the mathematics of Levi's error avoidance principle [11] in a theory called strongly satisficing decision theory (SSDT) [12, 13].

STRONGLY SATISFICING DECISION THEORY

Simon [9, 10] addressed the issue of bounded rationality by defining an *aspiration level* [14], such that once this level is met, the corresponding solution is deemed adequate, or *satisficing*. SSDT provides a formal definition of the aspiration level by utilizing Levi's mathematical approach to avoiding error.

Satisficing Decisions

Practical decision making. In its original philosophical context, Levi's approach to decision making addresses the question of "What to believe." For practical decision making such as in CAAS's, the question shifts to the issue of "How to act." SSDT was developed as an application of Levi's philo-

0-7803-4269-0/97/$10.00 © 1998 IEEE

sophical theory to those practical problems which require action.

SSDT employs the *accuracy* f_A and *rejectability* f_R decision attributes, which are generalizations of Levi's philosophical constructs, and which are analogous to benefit and cost values in MAUT. To form a design procedure, it is necessary to give operational definitions of these notions.

- *Accuracy* means conformity to a given standard, where the standard for practical decision making corresponds to whatever goal or objective is relevant to the problem.
- *Rejectability* asses the undesirable consequences of an action which are typically manifest in the form of costs or other penalties that would accrue simply as a result of taking the action, regardless of its accuracy.

For human-CAAS systems, the safety of the system is the basis for accuracy, and f_A corresponds the degree of success the CAAS has in achieving this goal. The cost to driver autonomy is the basis for rejectability, and f_R corresponds to the degree to which the CAAS interferes with driver autonomy. By defining these two system attributes, we have acknowledged the potential differences between the desire to improve safety and the desire to consider the cost of action to the driver. In practical terms, the CAAS system should go unnoticed by the attentive driver, but issue warnings to alert the inattentive driver and intervene to prevent collisions and accidents.

The satisficing decision rule. Formally, let U denote the set of possible decisions, and let Θ denote the states of nature. The states of nature represent those conditions which affect the consequence of a decision, but which cannot be controlled. For each decision $u \in U$ and for each state of nature $\theta \in \Theta$, a consequence results. The accuracy $f_A : U \times \Theta \mapsto \Re$ and rejectability $f_R : U \times \Theta \mapsto \Re$ functions are defined for each consequence (i.e., action/state-of-nature pair). Thus, as in MAUT, consequences are partitioned into two attributes.

According to Levi [11], *to avoid error, a decision maker eliminates those decisions which are more rejectability than accurate.* In terms of the lane departure system, an action is satisficing if it contributes to driver safety more than it interferes with the driver's autonomy. The set of all decisions which cannot be justifiably eliminated for a given θ is called the *satisficing set*, and is defined as

$$S_b(\theta) = \{u : f_A(u;\theta) \geq b f_R(u;\theta)\}.$$

The parameter $b \in [0, \infty)$ allows a decision-maker to alter the relative weight between accuracy and rejectability as well as ensures that the two attributes are comparable. Levi's rule provides a set-based mathematical formalism for the satisficing principle where the aspiration level, and hence the notion of adequacy, is defined in terms of the accuracy and the rejectability functions.

Strongly Satisficing Decisions

Although the set S_b contains all possible actions that are legitimate candidates for adoption, they generally will not be equal in overall quality. Thus, we are motivated to further refine the set of satisficing actions by employing the domination principle from MAUT. Define $B(u;\theta)$ as the set of actions that are *strictly better* than u; i.e., the set of all possible actions that are less rejectable but not less accurate than u, or are more accurate but not more rejectable that u. If $B(u;\theta) = \emptyset$, then no actions can be preferred to u in both accuracy and rejectability, and u is a (weakly) *non-dominated* action with respect to θ. The set

$$\mathcal{E}(\theta) = \{u \in U : B(u;\theta) = \emptyset\}$$

contains all non-dominated actions. The intersection of this set with the satisficing set yields the *strongly satisficing* set

$$\mathbf{S}_b(\theta) = \mathcal{E}(\theta) \cap S_b(\theta).$$

We can define the support of a decision as those states of nature θ for which u is strongly satisficing

$$\text{support}_b(\theta) = \{u : u \in \mathbf{S}_b(\theta)\}.$$

This set will be used to identify those conditions which justify the application of a CAAS action.

CAAS EXAMPLE: LANE DEPARTURE

It is desirable to design human-CAAS systems which are designed to not only make the vehicle more safe, but also allow the driver to retain vehicle control, thereby creating systems which bridge the gap between unassisted driving and fully autonomous vehicles. We demonstrate how SSDT can be used in the design of the decision logic for a lane departure CAAS.

CAAS Decision Logic

Assuming that the driver will not change current vehicle heading, the consequences of a CAAS action are parameterized by estimates of the Time to Lane Crossing (TLC), which we denote by $\theta = \tau$. For lane departure, we restrict attention to the set of CAAS actions $U = \{u_W, u_I\}$, where u_W and u_I indicate a lane departure warning, and intervention, respectively. The decision problem is to determine which $u \in U$ to invoke given an estimate $\hat{\tau}$ of the TLC. For any $\hat{\tau} \in \text{support}_b(u_W)$ a warning is issued, and for any $\hat{\tau} \in \text{support}_b(u_I)$ an intervention occurs.

Lane Departure Values

To use SSDT in the lane departure decision logic, it is necessary to independently formulate f_A and f_R using sound design principles and measurements [15]. Given these attributes, b parameterizes the critical threshold levels τ_W and τ_I which are defined as those TLC values below which warning and intervention actions are taken, respectively. In effect, b may provide a means to trade between driver autonomy and safety while accommodating differing driver preferences and environmental dependencies. These critical thresholds are chosen to trade between increasing safety and maintaining driver autonomy. If the TLC threshold is too low, accidents can occur which would have been prevented for higher thresholds; if the TLC threshold is too high, inappropriate and undesirable interventions can be issued.

Since the purpose of the system is to prevent road departures (maximize vehicle safety), the accuracy $f_A(u;\tau)$ should reflect this objective. Since unwanted warnings and interventions incur a cost to the driver (compromises driver autonomy which can affect driver patience, comfort, and attention), the rejectability $f_R(u;\tau)$ should reflect the objective to minimize unwarranted warnings and interventions. By formulating f_A and f_R in terms of expected consequences [16], these competing criteria can both be conveniently described by two factors: the *valuation* $J(u;\tau)$, meaning the payoff or cost, of a CAAS action; and the *likelihood* $\ell(u;\tau)$ that the CAAS action will produce a particular consequence. The attributes $f_A(u;\tau)$ and $f_R(u;\tau)$ are thus obtained from $f(u;\tau) = J(u;\tau)\ell(u;\tau)$.

Accuracy. The accuracy attribute is based on the desire to "prevent road departures for the largest possible set of departure conditions" [2, p.68]. The function $f_A(u;\tau) = \alpha(u)\tau e^{-\alpha(u)\tau}$, which is diagrammed in Figure 1, represents this desire. The following factors determine the accuracy (see [16]): (a) CAAS actions are most beneficial if they are issued early enough for the driver/vehicle to respond to them which implies that the payoff for a CAAS action increases as TLC increases whence $J_A(u;\tau) = \tau$, and (b) early CAAS actions (those issued at large TLC values) are less likely to effect driver/vehicle corrective behavior than later actions (those issued at small TLC values) whence $\ell_A(u;\tau) = \alpha(u)e^{-\alpha(u)\tau}$. We now discuss how $\alpha(u)$ can be chosen for for the "average" driver. A detailed discussion of these issues is given in [16].

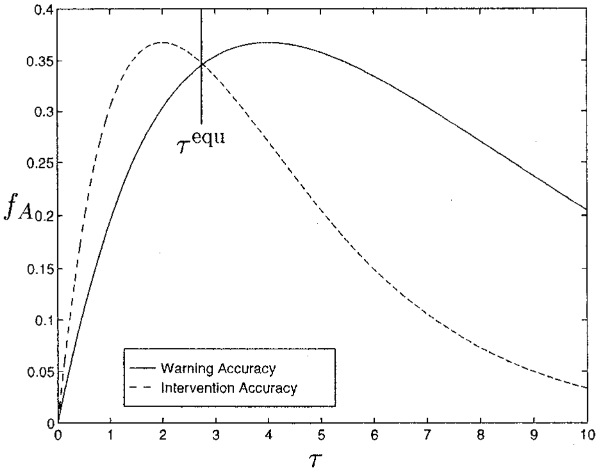

Figure 1: Accuracy attributes $f_A(u_W;\tau)$ and $f_A(u_I;\tau)$ as a function of the decision u and TLC.

- *Warning:* Most drivers respond to an unexpected driving situation within an interval of 1.5 and 4.0

seconds, with an average value of 2.5 seconds [17]. The value of $\alpha(u_W) = 1/4$ (implying $\tau_W^{\max} = 4.0$) is chosen because the warning will best accomplish its purpose (for most drivers) if signaled at four seconds before lane crossing. Thus, independent of rejectability, the accuracy $f_A(u_W;\tau)$ (represented by the solid line in Figure 1) achieves its maximum at $\tau_W^{\max} = 4.0$.

- *Intervention:* For intervention, the same factors determine the shape of $f_A(u_I;\tau)$ (represented by the dashed line in Figure 1) but $\tau_I^{\max}$ and, hence $\alpha(u_I)$, depend on two human factor considerations (a) an intervention should not occur until some time after the warning occurs, giving the driver a chance to react to the warning, and (b), an intervention should occur early enough to allow the CAAS intervention controller to smoothly and safely intervene. We select the value $\alpha(u_I) = 1/2$ (implying $\tau_I^{\max} = 2$ seconds) because the intervention will best accomplish its purpose if applied at two seconds before lane crossing and two seconds after warning $\tau_W^{\max}$.

Rejectability. One of the advantages of MAUT in general and SSDT in particular is the ability to independently assess different decision attributes. Therefore, the preceding development has addressed the issue of increasing safety, but ignored the issue of unwanted interventions and warnings. Clearly, warning a driver or intervening in vehicle control are undesirable if done too early because early CAAS actions imply more "false warnings and unwanted interventions" [2, p.68] thereby interfering with driver autonomy by demanding attention. The function $f_R(u;\tau) = \beta(u)\tau^2$, which is diagrammed in Figure 2, represents this decision criterion. The following factors determine the rejectability (see [16]): (a) early CAAS actions (those issued at large TLC values) interfere with driver performance more than later actions [15] (those issued at small TLC values) whence $J_R(u;\tau) = \beta(u)\tau^2$, and (b) the likelihood of interfering with driver autonomy requires information about false alarms [16], but when such information is unavailable, the likely response can be set to a constant for all TLC values whence $\ell_R(u;\tau) = 1$.

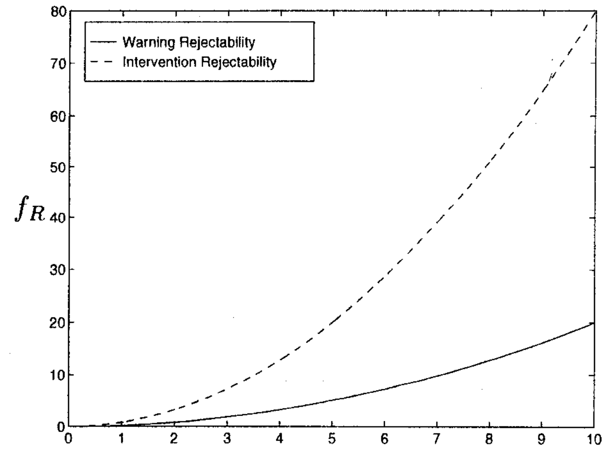

Figure 2: Rejectability attributes $f_R(u_W;\tau)$ and $f_R(u_I;\tau)$ as a function of the decision u and TLC.

- *Warning:* The nominal value $\beta(u_W) = 0.2/e$ yields a warning threshold consistent (for $b = 1$) with that derived from the subjective driver preferences reported in [2]. Note that this parameter is approximately driver independent since it is determined for the "average" driver.
- *Intervention:* The nominal value $\beta(u_I) = 0.8/e$ yields an intervention threshold consistent with that reported in [2]. Note that $\beta(u_I) = 4\beta(u_W)$ which indicates that interventions are four times more costly to driver autonomy than warnings.

Strongly Satisficing Actions

To make a decision, $\hat{\tau}$ is estimated and those $u \in \mathbf{S}_b(\hat{\tau})$ are implemented. Given f_A and f_R, we can describe the conditions for which $u \in \mathbf{S}_b(\tau)$ by (a) determining $S_b(\tau)$, (b) determining $\mathcal{E}(\tau)$, and (c) determining support$_b(u)$. In determining $S_b(\tau)$, the critical TLC values τ_W' and τ_I' occur when accuracy and rejectability are equal,

$$\begin{aligned} \tau_W' &= \arg_\tau \{f_A(u_W;\tau) = bf_R(u_W;\tau)\} \quad (1) \\ \tau_I' &= \arg_\tau \{f_A(u_I;\tau) = bf_R(u_I;\tau)\}, \quad (2) \end{aligned}$$

whence

$$\begin{aligned} u_W \in S_b(\tau) &\qquad \text{for } \tau \le \tau_W' \\ u_I \in S_b(\tau) &\qquad \text{for } \tau \le \tau_I'. \end{aligned}$$

From Figure 2, it is apparent that $f_R(u_I;\tau) > f_R(u_W;\tau)$ for all $\tau > 0$. Thus, the set of dominating actions can be determined by comparing

$f_A(u_W;\tau)$ to $f_A(u_I;\tau)$. The critical value $\tau^{\mathrm{equ}} = 2.77$ seconds is that value for which $f_A(u_W;\tau) = f_A(u_I;\tau)$ (see Figure 1),

$$\tau^{\mathrm{equ}} = \arg_\tau \{f_A(u_I;\tau) = f_A(u_W;\tau)\} \quad (3)$$

For $\tau < \tau^{\mathrm{equ}}$ both warning and intervention actions are permissible, but for $\tau \geq \tau^{\mathrm{equ}}$, only u_W is permissible since it dominates u_I. Thus,

$$\mathcal{E}(\tau) = \begin{cases} \{u_W, u_I\} & \text{for } \tau \leq \tau^{\mathrm{equ}} \\ \{u_W\} & \text{for } \tau > \tau^{\mathrm{equ}} \end{cases} .$$

In summary, the critical thresholds are defined using (1)-(3) as $\tau_W = \tau_W'$ and $\tau_I = \min(\tau^{\mathrm{equ}}, \tau_I')$, and the regions of support are

$$\begin{aligned} \mathrm{support}(u_W) &= \{\tau : \tau \leq \tau_W\} \\ \mathrm{support}(u_I) &= \{\tau : \tau \leq \tau_I\}. \end{aligned}$$

For $b = 1$, $\tau_W = \tau_W' = 2.03$ seconds and $\tau_I = \tau_I' = 1.01$ seconds, which agree with the values presented in [2].

Driver dependence. Note that for many drivers, $\tau_W' = 2.03$ seconds is within the range of normal operation which implies that $b = 1$ may be inappropriate for such drivers. Since different drivers have different operating ranges, it is desirable to adapt system operation to match an individual driver's preferences. Since $\alpha(u)$ and $\beta(u)$ are determined for the "average" driver, the b parameter is used to adapt the system operation to individual driver preferences and changing situations [18].

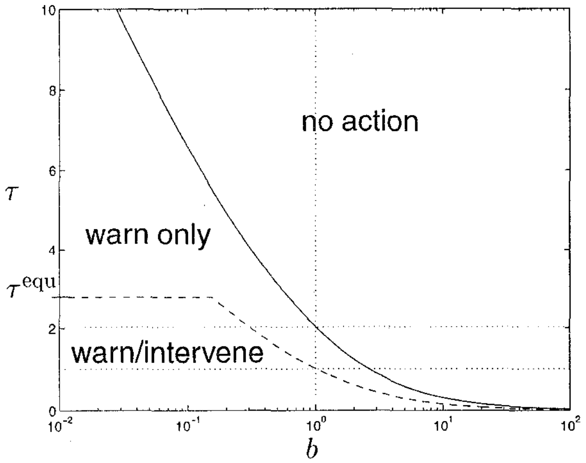

Figure 3: Critical times for warning and intervention as a function of b. The b parameter is a driver-dependent measure of the subjective disposition toward accepting assistance from the warning/intervention system.

In Figure 3, τ_W and τ_I, represented by the solid and dashed lines, respectively, are plotted as functions of b. The two horizontal dotted lines indicate the two critical values of τ for a nominal value of $b = 1$, which is represented by the vertical dotted line. Observe that the τ_I curve becomes constant for values less than about $b = .15$. This occurs because warning dominates intervention for $\tau \geq \tau^{\mathrm{equ}}$. If there are drivers for whom a nominal $b = 1$ value is inappropriate, then b can be adjusted to conform to their individual preferences. At one extreme, drivers who do not want the CAAS system to be a factor can be accommodated by letting $b \to \infty$. For these drivers, the strongly satisficing set is empty unless $\tau = 0.0$. At the other extreme, drivers who, for the sake of safety, do not mind frequent interventions or warnings can be accommodated by letting $b \to 0.0$. These drivers receive frequent warnings (since $\tau_W \to \infty$), but no intervention unless $\tau \leq \tau^{\mathrm{equ}}$ (since $\tau_I = \tau^{\mathrm{equ}}$). A similar reasoning can be employed to adjust the system to changing driving conditions.

DISCUSSION

We have presented strongly satisficing decision theory as a method for designing collision and accident avoidance systems. The motivation for this approach is the need to simultaneously account for the design criteria of driver safety and driver autonomy. Rather than arbitrarily aggregating these design criteria in a single performance valuation and then extremizing this valuation, two independent numerical attributes which are compared to determine when CAAS actions are appropriate. This comparison, and the selection of dominating actions, determine which set of CAAS actions are appropriate for the observed environment. In practical system design, tradeoffs between system safety and driver autonomy are intuitively handled by setting thresholds. Strongly satisficing decision theory provides a formal method for establishing such thresholds using two independent performance criteria, and motivates future research efforts to adaptively adjust these thresholds to individual driver preferences and changing driving situations.

References

[1] August Burgett. Crash avoidance holds key to safer US highways. *ITS: Intelligent Transportation Systems*, pages 94–98, September 1996.

[2] D. J. LeBlanc, G. E. Johnson, P. J.Th. Venhovens, G. Gerber, R. DeSonia, R. D. Ervin, C.-F. Lin, A. G. Ulsoy, and T. E. Pilutti. CAPC: A Road-Departure Prevention System. *IEEE Control Systems*, 16(6):61–71, December 1996.

[3] D. J. LeBlanc, P. J. Th. Venhovens, C.-F. Lin, T. E. Pilutti, R. D. Ervin, A. G. Ulsoy, C. MacAdam, and G. E. Johnson. A warning and intervention system to prevent road-departure accidents. In L. Segel, editor, *The Dynamics of Vehicles on Roads and Tracks, Proceedings of the 14th IAVSD Symposium*, Ann Arbor, MI, August 1995.

[4] L. Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983.

[5] J. Reason. Cognitive aids in process environments: prostheses or tools? *Int. J. Man-Machine Studies*, 27:463–470, 1987.

[6] Y. Xiao, P. Milgram, and D. J. Doyle. Planning behavior and its functional role in interactions with complex systems. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 27(3):313–324, May 1997.

[7] N. Moray. Designing for transportation safety in the light of perception, attention, and mental models. *Ergonomics*, 33(10/11):1201–1213, 1990.

[8] IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans. Special issue on Human Interaction with Complex Systems, May 1997.

[9] H. A. Simon. A behavioral model of rational choice. *Quart. J. Economics*, 59:99–118, 1955.

[10] Herbert A. Simon. *The Sciences of the Artificial.* MIT Press, 3rd edition, 1996.

[11] I. Levi. *The Enterprise of Knowledge.* MIT Press, Cambridge, Massachusetts, 1980.

[12] W. C. Stirling, M. A. Goodrich, and R. L. Frost. Procedurally rational decision-making and control. *IEEE Control Systems*, 16(5):66–75, October 1996.

[13] M. A. Goodrich. *A Theory of Satisficing Control.* Ph.D. Dissertation, Brigham Young University, 1996.

[14] T. Matsuda and S. Takatsu. Characterization of satisficing decision criterion. *Information Sciences*, 17(2):131–151, 1979.

[15] A. D. Horowitz and T. A. Dingus. Warning signal design: A key human factors issue in an in-vehicle front-to-rear-end collision warning system. In *Proceedings of the Human Factors Society 36th Annual Meeting*, pages 1011–1013, Atlanta, Georgia USA, October 1992.

[16] M. A. Goodrich and E. R. Boer. Maintaining driver autonomy with collision avoidance systems: A satisficing approach. Technical Report TR-97-2, Cambridge Basic Research, Nissan Research and Development, Inc., Cambridge, MA USA, 1997.

[17] L. Evans. *Traffic Safety and the Driver.* Von Nostrand Reinhold, New York, 1991.

[18] T. Inagaki. Situation-adaptive responsibility allocation for human-centered automation. *Transactions of the Society of Instrument and Control Engineers*, 31(3), 292-298 1995.