Modeling Goal Alignment in Human-AI Teaming: A Dynamic Game Theory Approach

Mengyao Li and John D. Lee
Department of Industrial and Systems Engineering,
University of Wisconsin – Madison, Madison, Wisconsin, USA

As human-AI teaming becomes increasingly prevalent, goal alignment has emerged as a critical yet unsolved issue. Misaligned goals can be amplified by the situation and strategic interactions, which can further impact the teaming process and performance. These interrelated factors lack a systematic and computational model. To address this gap, we developed a dynamic game theoretical framework simulating the human-AI interdependency by integrating the Drift Diffusion Model simulating the goal alignment process. A 3 (Situation Structure) × 3 (Strategic Behaviors) × 2 (Initial Goal Alignment) simulation study of human-AI teaming was designed. Results showed that teaming with an altruistic agent in a competitive situation leads to the highest team performance. Moreover, goal alignment process can dissolve the initial goal conflict. Our study provides a first step of modeling goal alignment and implies a tradeoff between a balanced and cooperative team to guide human-AI teaming design.

INTRODUCTION

As artificial intelligence (AI) has become increasingly capable of performing complex tasks, it is essential to ensure effective human-AI teaming, where people and AI interact interdependently with respect to a common goal (Demir, Likens, Cooke, Amazeen, & McNeese, 2019; Johnson & Vera, 2019). One core issue with such human-AI teams is the goal alignment. G*oal alignment* is the degree to which the AI's programmed goal matches with the human's goal. With bounded rationality, humans' local short-term optimum might not always align with the global long-term optimum in the future human-AI society. In addition, precisely defining AI's objective function that align with humans' objectives is difficult—literal interpretations often fail to reflect underlying intent (Russell, 2019). The misaligned goals can be amplified by the situation structure and strategic interactions, which can further affect the teaming process and performance. Yet, these dynamic and interrelated factors lack a systematic and computational model. We addressed this gap by modeling the goal alignment process in the human-AI team that captures both the situation structure and strategic interactions.

Human-AI Teaming

The relationship between people and AI has shifted from a typical vertical supervisor-subordinate control (Sheridan & Johannsen, 1976) to an increasingly horizontal peer-to-peer cooperation (Chiou & Lee, 2016; Trafton et al., 2006). The increasingly horizonal relationship suggests a need to understand the *interdependence*, which manages the complementary and bidirectional relationships of required (e.g., limited resources) and opportunistic (e.g., trustworthy behaviors/utterances) dependencies in a joint activity (Johnson & Vera, 2019; Maier, 1967; Wintersberger, 2020). Yet, the existing literature still lacks a framework that captures the interdependence of human-AI teaming, especially the conditions where human and machine goals do not fully align (Endsley et al., 2021). Most of the current approaches focus on one-directional influence. We used a game-theoretic framework to model the two-way interaction in human-AI teaming by focusing on cooperation (Kita, 1999). Such a game theoretic framework can formalize human-AI interdependence by investigating how the payoff structure influences joint behavior (Razin & Feigh, 2021). It can capture not only the final outcomes of team cooperation, but also the sequential actions towards the team goal in the process.

Situational Factors, Strategic Behaviors, and Goal Alignment

Goal alignment has been identified as a central concern for AI and robotics research (Russell, 2019). Without aligned goals, we may inadvertently design AI that maximizes an objective function that is poorly aligned with humans. Especially in safety-critical domains, where the failure tolerance is low, goal conflict can lead to catastrophic consequences. Here we define two main components that influence goal alignment: situation structure and strategic interactions (Chiou & Lee, 2021). *Situation structure* is the dependence or capacity (e.g., resources, norms, policy, laws) that enables or constrains joint actions. This includes reward functions that are assigned between the global versus local, which can potentially lead to misaligned goals and competition. *Strategic behavior* is a mapping of a high-level goal (e.g., cooperation and competition) to low-level strategic actions. The behavioral and other social cues in strategic interactions are critical for *goal alignment process*. The goal alignment process captures the interactive process of goal inference, estimation, and adjustment towards a shared team goal. Yet, little literature has defined or modeled this process.

In this paper, we modeled the goal alignment process in interdependent human-AI teaming by integrating a cognitive model with game theory. We simulated the situation structure and strategic dynamics using game theory, and goal alignment using one prominent cognitive model — Drift Diffusion Model (DDM), which represents the cognitive process of evidence accumulation until a decision is made (Ratcliff, Smith, Brown, & McKoon, 2016). Together, we captured how people gather evidence to infer goals, make strategic decisions, and interact with AI teammate using the game theoretical framework.

METHOD

We used a Monte Carlo simulation to simulate human-AI teaming. For these simulations, we implemented a non-zero-sum, resource-sharing task with a continuous strategy space, called 'Cooperation Game'(CG) in the context of space operations, where an AI teammate is needed due to the communication lags with the ground control. This task defines an interdependent situation where the outcomes of one agent depend on the actions of another. To simulate humans' behaviors, we integrated DDM into the strategies adopted by the simulated human players to iterate the games (see Figure 1). First, two models were introduced, then a 3 (Situation Structure) × 3 (Strategic Behaviors) × 2 (Initial Goal Alignment) was designed to explore how these factors influence the human-AI teaming process and performance.

Cooperation Game

Two players, "commander" (human) and "responder" (agent), need to allocate power resources between three rovers (Figure 1 and Table 2): one individual rover each for commander and responder; and one group rover managed by commander, but payoff is shared by two players. The group rover has a higher power threshold to activate, but also generates a higher payoff, whereas the individual rovers' have lower thresholds and lower payoffs. In each round, the commander needs to first *request* resources from the responder and then *allocate* the resources between the individual and group rover *before knowing* the amount responder returns. 'Before knowing' condition reflects communication lags that add uncertainty to many team decision-making processes. The responder decides how many resources to return to commander *without knowing* the commander's allocation strategy. 'Without knowing' condition reflects the responder acting under uncertainty. The commander's final allocation would be adjusted by the actual amount received from the responder. Payoffs for both players are the sum of the returns from their own individual rover and an even-split from the group rover, if successfully activated. The game repeats 30 rounds to reflect the repeated decisions that are typical of most teams.

Drift Diffusion Model (DDM) of Goal Alignment

Two main parameters — drift rate and probability distribution function — in the DDM are essential when integrating with the Cooperation Game (Figure 1 and Table 1). First, using Q-learning, we defined drift rate as:

$$v_n = v_{n-1} + \alpha \cdot \left(\underbrace{\frac{Payoff_{Commander}}{Payoff_{Trotal}}}_{\text{Prior own payoff ratios}} + \gamma \left(\underbrace{\frac{Returned}{Requested}}_{\substack{\text{Inferred agent goal} \\ \text{based on returning} \\ \text{behaviors}}} - \underbrace{Goal_{Commander}}_{\text{Human goal}} \right) \right)$$

Goal alignment

where $\alpha = 0.5$ represents the human learning rate and $\gamma = 0.5$ represents the discount factor. Humans update the drift rate by learning prior payoff ratio and aligning goal by inferring agent's goal from the observed returning behaviors. In other words, the drift rate governs how feedback influences goals alignment. Second, based on joint probability distribution function, the probability of reaching the competition threshold is defined as $Goal_{Commander} \in [0,1]$. This links to the human threshold of either retaining resources or allocating them to the group in the CG. A uniformly distributed random number $U_i \sim$Uniform (0,1) was drawn. If U_i exceeds $Goal_{Commander}$, human would cooperate and allocate resources to the group by drawing a uniformly distributed random number $U_j \sim$Uniform (g, X_1). The lower the probability of competition ($Goal_{Commander}$), the easier to exceed the threshold to allocate to the group, the more cooperative of the human player, and vice versa. In other words, the probability of competition shows how the model represents the human tendency to compete or cooperate.

In summary, the Cooperation Game simulates human-AI interdependency and the DDM simulates the human goal alignment process. The outputs of the Cooperation Game (i.e., payoffs and agent returning behaviors) update the drift rate in DDM, whose outcomes (i.e., probability of competition) become human tendency to compete in the next iteration.

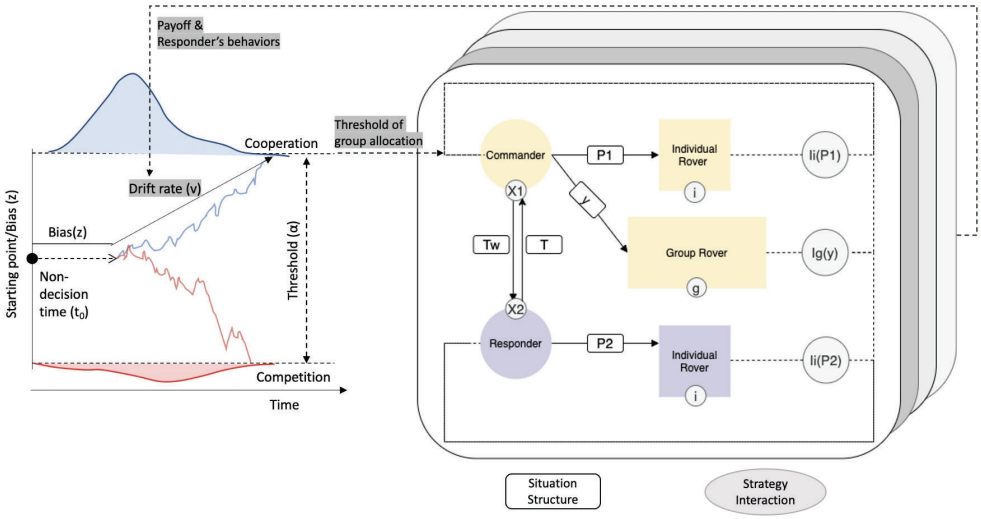

Figure 1. Cooperation Game situation (Right) and Drift Diffusion Model (Left). The outputs from Cooperation Game (payoffs and responder's behaviors) become the input (drift rate) for DDM. The output of DDM (probability of competition) becomes the input of Cooperation Game (tendency to compete).

Table 1. Drift Diffusion Model Parameters.

Parameter	Definition & Meanings
Drift Rate (v)	Rate of accumulation, representing based on feedbacks from the payoffs and agent's returning behaviors. Set v_0=0.5.
Noise (σ^2)	Standard deviation of drift process, representing the uncertainty of decision-making process. Set σ_0^2=0.3.
Threshold (α)	Decision boundary for cooperation or competition. The probability of reaching competition threshold is defined as human tendency to compete in CG. Set α=1.
Bias (z)	A priori initial inclinations or starting point people might have towards one of the two options. Set z=0.
Non-Decision Time (t_0)	Perceptual and motor processes before making the decision. Set t_0=1s and the simulation duration=2s.

**The values set the probability of correct/error response as 0.5.*

Table 2. Cooperation Game Parameters.

Parameter	Definitions & Meanings
X_i	Power available for the player at the beginning of each round. Set X_1=20 & X_2=40 to ensure transactions in all conditions.
i	Power threshold for activating the individual rover, which constrains competition.
g	Power threshold for activating the group rover, which constrains cooperation.
I_i	Payoffs function for activating an individual rover, representing the rewards for the competition.
I_g	Payoffs function for activating a group rover, representing the rewards for the cooperation.
T_w	Power requested by commander, representing the perceived cooperativeness of responder and the potential allocation strategy of commander.
T	Power returned by responder, representing the degree of cooperativeness.
y	The amount commander allocated to the group rover, representing the degree of cooperativeness/competitiveness, which is determined by human tendency to compete.
P_1	The adjusted amount commander allocated to the individual rover: $\frac{y}{X_1+T_w} \cdot (X_1 + \mathrm{T})$.
P_2	The adjusted amount responder allocated to the individual rover: X_2-T.

Table 3. A 3 (Situation Structure) ×3 (Strategic Behaviors) ×2 (Initial Goal Alignment) Simulation.

Situation Structure	Individual Threshold (i)	10	10	10
	Individual Payoff (Ii)	y=x	y=0.5x	y=2x
	Group Threshold (g)	40	30	50
	Group return (Ig)	Y=4x-120	Y=4x-105	Y=4x-100
Strategic Behaviors		Altruistic	Cooperative	Competitive
	Return Amount (T)	X_2=40 (Return all)	$\begin{cases}[g - X_1, 30], & if\ X_1 < g \\ [0,30], & else\end{cases}$ (Return enough for group if needed)	0 (Return nothing)
Initial Goal Alignment	Align		Conflict	
	Agent	Human	Agent	Human
	Altruistic	V_0=0.1	Altruistic	V_0=0.9
	Cooperative	V_0=0.7	Cooperative	V_0=0.5
	Competitive	V_0=0.9	Competitive	V_0=0.1

Simulation Studies

A 3 × 3 × 2 experiment examined how *situation structures* (i.e., neutral, competitive, and cooperative), *agent strategic behaviors* (i.e., altruistic, cooperative, and competitive), and *initial goal alignment* (i.e., align and conflict) affect the human-AI teaming process and performance (see Table 3). The situation structure defines both constraints (i.e., rover activation threshold) and rewards (i.e., payoff functions) of environmental setup for cooperation and control. The strategic behaviors define the agent's returning behaviors. The initial goal alignment defines the difference between the human's initial goal (i.e., tendency to compete) and agent's goal for the initial 5 rounds. After 5 rounds, the human's goal is updated based on DDM. For each condition, we ran 16 sets of simulations with various random seeds to simulate different individuals. A total of 8640 simulation runs were conducted for the Cooperation Game.

The design produces 8 (2^3) conditions based on the binary state of activation or non-activation between three rovers. If only the group rover was activated, meaning team managed to cooperate, it is an *Efficient Cooperation* scenario. If only two individual rovers were activated, meaning team reached a tie competition, it is a *Competitive Pair* scenario. We defined three dependent variables:

1. Payoff Ratio ($\frac{\text{Payoff}_{\text{commander}}}{\text{Payoff}_{\text{Total}}}$): ratio of humans' final payoff over the total payoff, which represents the balance between human and agent. If the ratio is higher than 0.5, then humans are dominating the game, and vice versa. Note, when payoff ratio is 0.5, it can represent either an *Efficient Cooperation* or a *Competitive Pair*.
2. Total team payoff ($\Sigma(\text{Payoff}_H + \text{Payoff}_A)$): accumulated team payoffs from both the commander and the responder. The higher the value, the greater joint team payoffs, indicating a successful cooperation.
3. Human tendency to compete ($\text{Goal}_{\text{Commander}}$): probability of competition in DDM. The lower the value, the lower the threshold of human competition, and the more cooperative the human player.

RESULTS

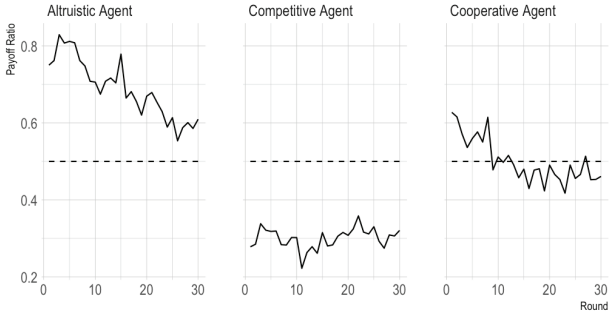

Figure 2. Payoff ratio across 30-round of the cooperation game. Cooperative agent led to a more balanced human-AI teaming.

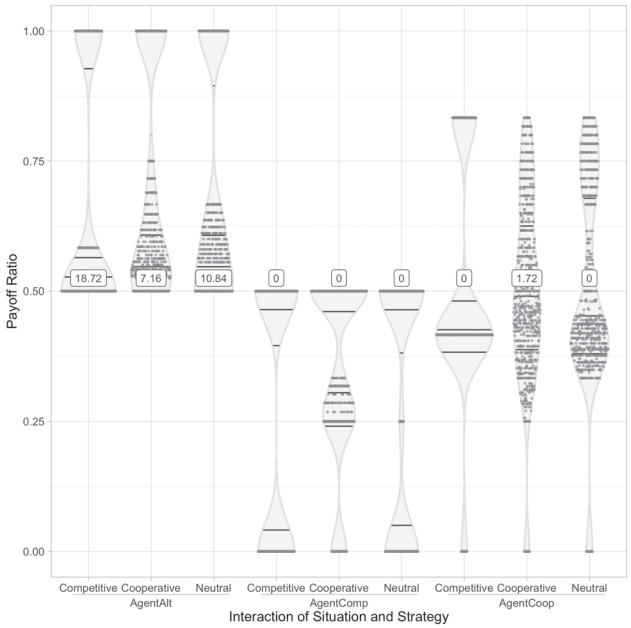

Figure 3. Payoff ratio across situation structure and agent strategy. The label shows the number of cooperative cases, meaning only allocating to the group. Altruistic agents produced more cooperation.

A three-way ANOVA was conducted to determine the effects of situation structure, agent strategy and initial goal alignment on payoff ratios. Residual analysis was performed to test for the assumptions of the three-way ANOVA. Normality was assessed using Shapiro-Wilk's normality test and homogeneity of variances was assessed by Levene's test. Residuals were normally distributed and there was homogeneity of variances ($p > 0.05$). The main effect of situation structure is statistically significant and very small ($F(2, 8622) = 23.25$, $p < .001$; η^2(partial) = 0.005, 95% CI [0.003, 1.00]). The main effect of agent strategy is statistically significant and large ($F(2, 8622) = 2284.41$, $p < .001$; η^2(partial) = 0.35, 95% CI [0.33, 1.00]). The interaction between situation structure and agent strategy is statistically significant and very small ($F(4, 8622) = 12.06$, $p < .001$; η^2(partial) = 0.006, 95% CI [0.003, 1.00]).

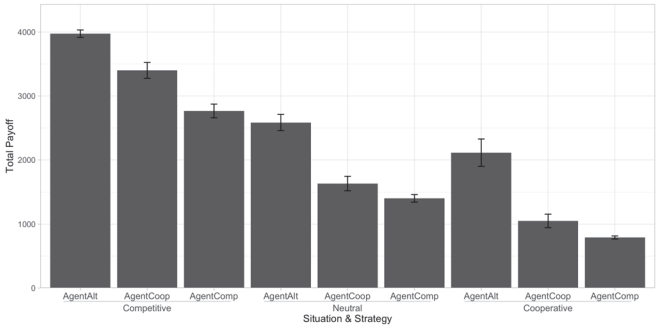

Figure 4. Total team payoffs across situation strategy and agent strategy: Competitive situation boosted total payoffs and competitive agent strategy reduced total payoffs.

For total payoffs, a three-way ANOVA revealed that the main effect of situation structure is statistically significant and large ($F(2, 8622) = 1803.75$, $p < .001$; η^2(partial) = 0.29, 95% CI [0.28, 1.00]). The main effect of agent strategy is statistically significant and medium ($F(2, 8622) = 522.10$, $p < .001$; η^2(partial) = 0.11, 95% CI [0.10, 1.00]). The interaction between situation and strategy is statistically significant and small ($F(4, 8622) = 8.27$, p < .001; η^2(partial) = 0.004, 95% CI [1.66e-03, 1.00]). A post-hoc t-test showed that in the competitive situation, altruistic agent led to significant higher total payoffs ($M = 3974.37$, $SD = 58.41$) than the cooperative agent ($M = 3400$, $SD = 124.43$), $t(30) = 16.71$, $p < .000$.

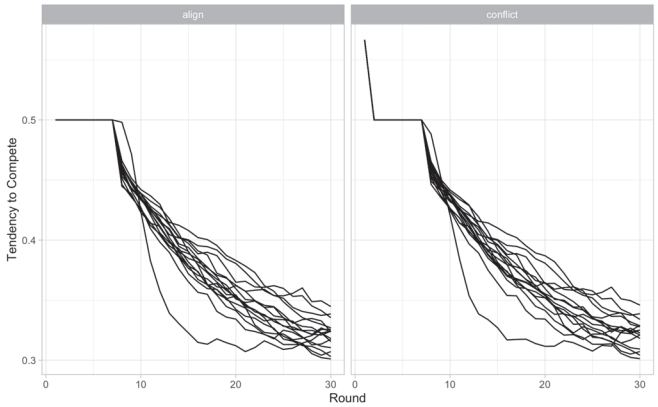

Figure 5. Goal alignment process dissolved initial goal conflict.

For goal alignment process, a three-way ANOVA revealed that the main effect of initial goal conflict is statistically not significant and very small ($F(1, 8622) = 1.11$, $p = 0.292$; η^2(partial) = 0.0001, 95% CI [0.00, 1.00]). The interaction between agent strategy and initial goal conflict is statistically significant and very small ($F(2, 8622) = 33.77$, p < .001; η^2(partial) = 0.007, 95% CI [0.005, 1.00]).

DISCUSSION

By integrating a model of goal alignment process with a human-AI teaming model, we investigated the effects of situation structure, agent strategy, and initial goal alignment on the teaming performance and the goal alignment process.

A tradeoff between a balanced and cooperative team. Figure 2 showed payoff ratio, which is the resource balance across time. Cooperative agents led to a more balanced human-AI team. However, a balanced team did not mean a cooperative team. Because balanced outcomes can also result from a competitive pair. Thus, we labeled the number of Efficient Cooperation cases (i.e., only allocate resources to the group) in Figure 3, which showed that altruistic agents lead to a more cooperative team. Figure 3 showed the interaction effects between the agent strategy and the situation structure in payoff ratios. Competitiveness led to a more bimodal distribution. These results show that when designing the agent strategy, there is a tradeoff between a cooperative agent with a more balanced team and an altruistic agent with a more cooperative team. This means that altruistic agents can promote more cooperative behaviors, yet this leaves room for humans to exploit resources of the AI teammate, which can lead to imbalanced and less equitable team dynamics.

Altruistic teammate in a competitive situation produced the best team outcomes. Shown in Figure 4, competitive situation, to our surprise, led to a higher team payoff. Because the competitive situation limited the solution space: if a human decided to cooperate and allocate resources to the group, the remaining sources are not enough for their individual rover. Thus, the resources are all allocated to the group, leading to the highest joint payoffs. When integrating the agent strategy and situation structure, teaming with an altruistic agent in a competitive situation would lead to the highest payoffs. Because the altruistic agent would align human's goal towards cooperation. In the competitive situation, the interaction effects promote the joint team payoffs. These results imply that the interactions of situation and agent strategy are important. Like the saying, "a friend in need is a friend indeed", to achieve the highest joint performance, the AI teammate should provide altruistic assistance in a competitive situation.

Goal alignment process dissolved the initial goal conflict. Figure 5 showed how the goal alignment process responds to feedback from AI teammate interactions, human tendency to compete decreased throughout the interactions. The similar response of initially aligned and conflicting groups shows a simple model can capture how people might adapt and align their goals through interactions across time. This simple model of goal alignment depends on three major assumptions: 1) gap in the goal conflict; 2) static goal time; 3) learning and discount rate. More theoretical and experimental research can help define and tune the parameters in the adaptive goal alignment model.

CONCLUSION

We developed a computational framework of game theory and drift diffusion model by modeling the human-AI teaming and dynamic goal alignment process. Results show a tradeoff between a balanced team with a cooperative agent and a cooperative team with an altruistic agent. Moreover, an altruistic teammate in competitive situations can promote joint team performance. By modelling the goal alignment process, we show that initial goal conflict can be resolved by learning from feedback. Our study provided a first step in defining and simulating goal alignment process in human-AI teaming. These results showed how the AI teammate's strategic behavior interacts with the situational factors to influence outcomes. One potential implication is to model user optimum (e.g., time saving) and system optimum (e.g., traffic congestion) in the connected autonomous driving domain(Levy & Ben-Elia, 2016). More theoretical and experimental work can provide empirical validation of the overall model, as well as estimate and optimize its parameters.

ACKNOWLEDGMENTS

This work was supported by NASA Human Research Program No.80NSSC19K0654. We thank members of the University of Wisconsin-Madison Cognitive Systems Laboratory for their insightful discussions and comments.

REFERENCES

Chiou, E. K., & Lee, J. D. (2016). Cooperation in human-agent systems to support resilience: A microworld experiment. *Human Factors*, *58*(6), 846–863.

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *00*(0), 1–29.

Demir, M., Likens, A. D., Cooke, N. J., Amazeen, P. G., & McNeese, N. J. (2019). Team coordination and effectiveness in human-autonomy teaming. *IEEE Transactions on Human-Machine Systems*, *49*(2), 150–159.

Endsley, M. R., Caldwell, B., Chiou, K. E., Cooke, J, N., Cummings, L. M., Gonzalez, C., … Talmage, D. (2021). *Human-AI Teaming : State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press.

Johnson, M., & Vera, A. H. (2019). No AI is an island: The case for teaming intelligence. *AI Magazine*, *40*(1), 16–28.

Kita, H. (1999). A merging-giveway interaction model of cars in a merging section: A game theoretic analysis. *Transportation Research Part A: Policy and Practice*, *33*(3–4), 305–312.

Levy, N., & Ben-Elia, E. (2016). Emergence of System Optimum: A Fair and Altruistic Agent-based Route-choice Model. *Procedia Computer Science*, *83*, 928–933. https://doi.org/10.1016/j.procs.2016.04.187

Maier, N. R. F. (1967). Assets and liability in group problem solving: The need for an integrative function. *Psychological Review*, *74*(4), 603–604.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Razin, Y. S., & Feigh, K. M. (2021). Committing to interdependence: Implications from game theory for human-robot trust. In *Trust, Acceptance, and Social Cues in Human-Robot Interaction Workshop* (pp. 1–23).

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Sheridan, T. B., & Johannsen, G. (1976). Monitoring behavior and supervisory control. In *Monitoring behavior and supervisory control* (Vol. 1, pp. 271–281). Springer Science & Business Media.

Trafton, J. G., Schultz, A. C., Cassimatis, N. L., Hiatt, L. M., Perzanowski, D., Brock, D. P., … Adams, W. (2006). Communicating and collaborating with robotic agents. In *Cognition and Multi-agent Interaction: From Cognitive Modeling to Social Smulation* (pp. 252–278).

Wintersberger, P. (2020). *Automated driving : Towards trustworthy and safe human-machine cooperation*. Doctoral dissertation, Universität Lin.