Towards A Formalization of the Critical Friend and Socio-Moral Decision-Making in Autonomous Systems

Joel Wester* , **Andreas Brännström** , **Juan Carlos Nieves**
Department of Computing Science
Umeå University
SE-901 87, Umeå, Sweden
c.joel.wester@gmail.com, andreasb@cs.umu.se, jcnieves@cs.umu.se

Abstract

The aim of this study is to investigate the Critical Friend (CF) and socio-moral decision-making, by understanding the CF in a formal way. CF, utilized to be understood as characterized behavior, is in contrast with state-of-the-art AI-systems argued to be a responsible construct, suitable for characterizing such formalization. This is done by asking 1) How to characterize a formal definition of the CF?, and 2) How to assess socio-moral decision-making of the CF? Results are a formal framework (in terms of an ontology and a transition system) of the CF, and an initial assessment of socio-moral decision-making of the CF. Results can be utilized by researchers to better understand socio-moral decision-making in AI-systems.

1 Introduction

Responsibility and fairness are crucial concepts for researchers and developers to consider before developing autonomous systems [Dignum, 2018]. Nevertheless, there are already implemented autonomous systems, such as conversational agents interacting with humans in critical situations. This in spite of the fact that there is no autonomous system fully capable of moral reasoning [Cervantes *et al.*, 2016], which reflects the current state of machine ethics [Charisi *et al.*, 2017]. However, there are recent efforts, focusing on investigating social behavior of autonomous systems, such as social robots. Work as such highlights less obsequious Artificial Moral Agents (AMAs) as an opportunity for Human-Robot interaction researchers to "avoid reinforcing [...] and positively challenge stereotyping and inappropriate user behaviors" (p. 4) [Winkle *et al.*, 2022]. This means that focus should be on what type of critical social behavior is desirable and what is not, as well as what critical social behavior is moral and what is not. Approaches as such aids current notions of going against e.g. stereotypical factors, instead advocating for feminist, anti-racist and prosocial approaches as part of responsible development of autonomous systems [Winkle *et al.*, 2021]. However, the question remains on how to explicitly do so.

To learn and understand more about morality in autonomous systems, socio-moral decision-making of autonomous systems needs to be assessed. Recent approaches highlight different challenges following moral aspects. For example, a recent study collected human moral preferences from individuals scattered over 100 countries, investigating a range of moral dilemmas in the context of automated vehicles, i.e. must-choose scenario, e.g. crash into 1 child vs. 3 adults [Awad *et al.*, 2018]. The authors presented 13 different dilemmas, from which they deduct three strong universal preferences; preference for sparing human lives [vs. animal lives], more lives [vs. a few lives] and young lives [vs. old lives]. However, there are highly relevant (socio-)moral dilemmas with less drastic outcomes than life-or-death. Socio-moral decision-making in social contexts regards how and when to say what in different situations, and how it will affect or influence a recipient. To illustrate, googling 'how do you tell someone they have bad breath' results in a large amount of varying results, which reflects the complexity of choosing between plague or cholera; should you avoid telling someone they have bad breath, or should you address it?

To further investigate socio-moral decision-making as such, the Critical Friend (CF) is introduced. The Critical Friend is in line with current notions of responsible autonomous behavior, and is here argued to have implicit potential to further such developments. Historically, the Critical Friend is a concept that has been discussed in educational research, and is understood in many different ways. For example, a well cited paper defines CF as "a trusted person who asks provocative questions [...]" (p.50) [Costa *et al.*, 1993] and a more recent paper as; "a challenging critic and a trusting friend" (p.1)[Storey and Wang, 2017]. The Critical Friend is usually employed in educational settings, such as in a relationship between teachers, where one of the teachers takes the role of being a Critical Friend and the other to be a recipient of the Critical Friend approach. The general purpose of the Critical Friend is under debate, but is essentially about increasing competence for both the Critical Friend and recipient of the Critical Friend.

Bearing the Critical Friend in mind, conversational agents (CAs) are progressively being developed in different contexts of health and well-being, such as mental health agents [Lu *et al.*, 2022] and social companions [Skjuve *et al.*, 2021]. CAs

*Contact Author

behavior and decision-making abilities are in line with different human abilities, such as being friendly or funny. One example of state-of-the-art CA with such abilities is Replika. Replika is described as a caring, always available AI companion to always be on your side. This is an appealing description of a CA with human-like abilities. But what does it mean to be a companion who cares, and to always be on your side? This is a difficult question to answer. However, introducing the CF as CA allows various investigations to take place, such as better understanding socio-moral decision-making. Characterizing such definitions and tools are especially important considering the pressing need for autonomous agents to have abilities such as being able to decide when criticizing (in)appropriate user behavior is morally correct.

In this study, the aim was to address these topics, and learn how individuals understand the Critical Friend conceptualized as a Conversational Agent. This is done by approaching the following research questions; **RQ1)** How to characterize a formal definition of the Critical Friend? **RQ2)** How to assess socio-moral decision-making of the Critical Friend

As implied, state-of-the-art autonomous systems do not fully consider socio-moral decision-making in human-computer interactions. Therefore, the main contribution of this study is a characterization of a formal framework of the Critical Friend, and an assessment of socio-moral decision-making of the Critical Friend, in terms of an ontology and a transition system. These formal models are available to researchers, designers and engineers, to further utilize, develop and implement in various human-AI interactions. Results are argued to be in line with the notion of responsible development of autonomous systems.

Sections are organized as follows. Section 2 presents recent efforts on the Critical Friend and socio-moral assessments. In Section 3, the structure of method and process are described. In Section 4, the resulting formal framework including an ontology and transition system are presented. In Section 5, the formal framework is discussed and compared with recent efforts on moral notions. Section 6 concludes the study with a discussion regarding limitations, possible applications and future directions.

2 Related Work

The Critical Friend is a concept or notion stemming from educational research. The construct of the Critical Friend is characterized as a set of behaviors that aids an individual to take a certain role in a given context with an overall purpose. Often, the Critical Friend is utilized in professional relationships, such as between teachers.

2.1 Models of the Critical Friend

A variety of models of the Critical Friend in a variety of contexts have been conceptualized, with the purpose of defining necessary items to be used for utilizing desirable behavior between e.g. peers. For example, a government initiative resulted in a longitudinal study, focusing on the nature of the Critical Friend in a fixed environment [Baskerville and Goldblatt, 2009]. This study identified items related to the Critical Friend, to be understood as necessary steps (or rules/heuristics) for utilizing the Critical Friend in a specific context. Such high-order taxonomies provide important layers to serve as a basis for further developing conceptual models of the Critical Friend. A different example focuses on professional learning and development (PLD). The Critical Friend is suggested to be a good fit with PLD, and is suggested to further support and enhance professional learning between peers [MacPhail *et al.*, 2021]. Here, items of interest are captured in a more independent way, suggested to be used as behavioral cues. Developing such models is necessary for capturing essential information about what the Critical Friend is, how it behaves and what may be the effects considering various influencing contexts and factors. However, to the author's knowledge, there is no formal definition of the Critical Friend.

2.2 Moral Assessment

Moral is understood in different ways, due to its subjective nature, and is therefore difficult to capture or describe in a formal way. However, there are such recent efforts, which aim to fill this gap and understand what most individuals perceive as moral. For instance in [Awad *et al.*, 2018], participants were presented with different dilemmas and had to decide on a moral action for a self-driving car following different outcomes. Results from this study may be utilized and guide future implementations of autonomous behavior. However, such monotonous moral decision-making is not realistic for socio-moral decision-making, thus it is not applicable for social-moral dilemmas. However, socio-moral notions are highly relevant for autonomous decision-making regardless of being a self-driving car or a social agent.

The following section presents an approach towards characterizing a formal framework of the Critical Friend, to assess socio-moral decision-making of autonomous systems.

3 Procedure

In this section, the methodology is described including data collection and analysis (Figure 1). The initial step of this study was to collect available studies online that included a definition of the Critical Friend. Using Google Scholar was suitable for this study, due to no additional studies of relevance using databases ACM Digital Library and arXiv. Articles had to include keywords Critical and Friend in the title, resulting in 299 studies. After reviewing titles and sorting out irrelevant studies, 52 studies were included in the next step. Reviewing abstracts of 52 studies excluded 20 studies, leaving 32 studies of relevance. A full reading of the 32 studies generated a final number of 18 studies to be included in analysis (see Table 1). In an iterative way, following a Grounded Theory process [Chun Tie *et al.*, 2019], collected studies were treated as raw data. First step in analysis consisted of initial coding, focusing on generating a large set of codes, allowing to establish a rigid basis for further analysis. Initial coding generated 206 codes. Secondly, intermediate coding helped aggregate initial codes, merging or deleting similar codes. Intermediate coding resulted in 75 codes sorted into 11 meaningful units. Thirdly, advanced coding focused on aggregating and revising codes and units. This step aggre-

Study	Sample Codes
[MacPhail *et al.*, 2021]	Honest; Suggestive
[Sjögren and Köhler, 2021]	Develop coping ability
[Petroelje Stolle *et al.*, 2019]	Support; Trust
[Martin and Russell, 2018]	Challenge; Discomfort
[Storey and Wang, 2017]	Provocative; Critical
[Evans, 2015]	Encouraging; Criticize
[Carlson, 2015]	Engagement; Commit
[Özek *et al.*, 2012]	Constructive; Capable
[Baskerville and Goldblatt, 2009]	Express; Attentive
[Carlson, 2009]	Appropriate; Relevant
[Deuchar, 2008]	Not always needed
[Gibbs and Angelides, 2008]	Liberate; Just
[Swaffield, 2008]	Detached; Advocacy
[Dahlgren *et al.*, 2006]	Decentre; Enhancing
[Schuck and Russell, 2005]	Frank; Insightful
[Swaffield, 2004]	Explain; Assistive
[Macbeath and Jardine, 1998]	Broadening; Sensitive
[Holden, 1997]	Understand; Encourage

Table 1: Studies included for analysis

gated 75 codes into 54 codes, and 11 meaningful units into 9 categories (see Section 4).

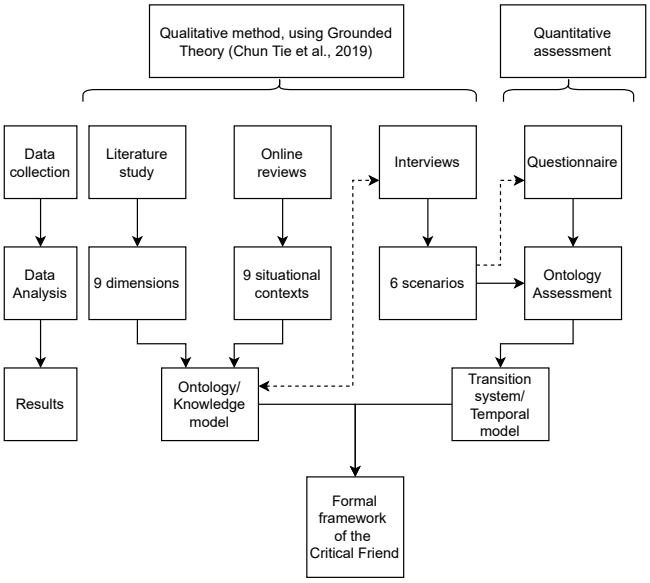

Figure 1: Flowchart of overall procedure

The second part of data collection focused on collecting anonymous user reviews of included CAs: Replika, Wysa and Woebot, via the App store website, in line with available ethical guidelines [Smedley and Coulson, 2021]. The reason for choosing these CAs is due to being among the most popular CAs contextualized in health and well-being [Wasil *et al.*, 2021]. A total of 35 users submitted a total of 34 comments dispersed over included CAs, each with different reasons, such as why interacting with the CA and how to interact with the CA. Aggregating 34 comments resulted in 24 meaningful units. Of 24 meaningful units, 9 contexts of interest were decided to be included for the interview process (described in the following paragraph), due to palpable differences in contexts.

The third part of data collection focused on interviews [DiCicco-Bloom and Crabtree, 2006]. Following a purposive sampling method, 4 experts were interviewed; 1) psychologist with expertise in Cognitive behavioral therapy, 2) teacher with expertise in/of educational settings, 3) computational scientist with expertise in Human-Computer interaction and Human-Robot interaction, and 4) philosophy professor with expertise in ethics and morals [Etikan *et al.*, 2016]. Interviews were semi-structured, and was complemented by using a 'think aloud' method, meaning that participants were encouraged to describe and motivate comments and choices [Eccles and Arsal, 2017]. Firstly, participants (n=4) were introduced to an example scenario, illustrating a short scenario between a user and a CA. Participants were then asked to build short scenarios using the presented 9 situational contexts and 9 CF-dimensions, in a stepwise manner deciding on one situational context, and what dimensions should be included for such situational context, creating a unique set of scenarios. Interview sessions were recorded and transcribed. Transcriptions together with the conceptual design of participants' preference of scenario were used to form simple utterances in line with suggested dimensions. Further, interviews were analyzed using Grounded Theory [Chun Tie *et al.*, 2019]. Initial coding generated 321 codes.

Last part of data collection focused on a questionnaire, where a total of 26 participants took part in answering the questionnaire. Participants' ages ranged between 20-33, and a majority of the participants were located in Stockholm. There was equal balance considering gender (male/female). Participants were exposed to six different scenarios (see Section 4) and asked to answer a set of questions. The questionnaire draws inspiration from the Multi-Dimensional Measure of Trust (MDMT), consisting of a 0-7 scale, where 0 = Not at all and 7 = Very [Ullman and Malle, 2019]. Subscales include ethical, respectable, principled, has integrity, and is suggested to be suitable constructs capturing moral aspects of human-AI interaction. The questionnaire was designed in line with and based on qualitative results, capturing the 9 dimensions (e.g. Challenge dimension = "Chatbot being challenging") Thus, 9 dimensions x 6 scenarios resulted in 54 questions focusing on the Critical Friend dimensions, and 4 items of the MDMT x 6 scenarios resulted in 24 questions, altogether 78 questions. Additionally, participants are asked to rank the 9 dimensions in line with temporal suitability, in order to have a comfortable interaction (e.g. 5-3-2-1-4-6-7-9-8).

4 Results

Results are two-parted, following the research questions. Together, the results comprise a formal framework (in terms of an ontology and a transition system) of the Critical Friend, and an assessment of socio-moral decision-making.

4.1 Knowledge model of the Critical Friend

Utilizing Grounded Theory resulted in 9 dimensions forming the concept of the Critical Friend. In what follows, the dimensions constituting the Critical Friend are presented and

explained, including the Critical Friend ontology and taxonomy (Figure 2 and 3).

Altruism is about being unconditional in various ways, such as being supportive to ways of life, sharing joy and sorrow, and being the best friend that you possibly can be. For example, in a situation where a user seeks support from the Critical Friend, the Critical Friend will completely avoid questioning e.g. motives, and will instead listen and support the individual with the purpose of making the individual feel better.

Affection is about confirming a user's feelings, such as raising an individual's achievements, affirming positive aspects of the individual, emphasizing strong features of the individual, in order to support the individual by confirming feelings. For example, if a user seeks contact and shares personal information, the Critical Friend will amplify/increase what is positive about this information.

Flexibility is about being the right Critical Friend for a specific individual independent of confounding factors. In relation to an individual, salient concepts such as dynamic and obsequious are used with the goal of being a good friend. For example, completely independent of what the user does or says to the Critical Friend, the Critical Friend will act completely obsequious.

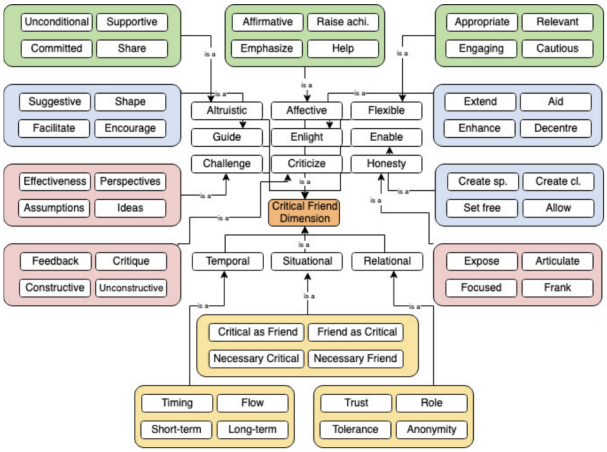

Figure 2: Knowledge model in terms of an ontology.

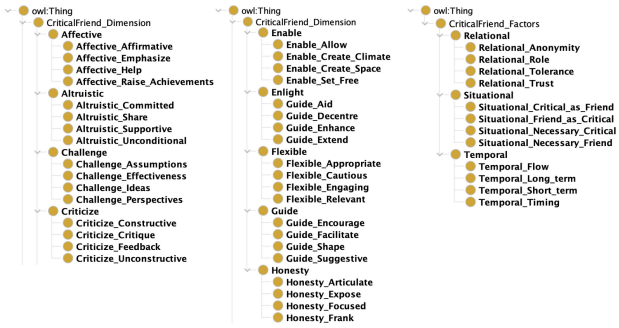

Figure 3: Critical Friend Taxonomy

Guide is about being a source of ideas and advice, in order to aid an user to make the right choice, without explicitly telling the user what to do. For example, if a user seeks help for irrational thoughts, the Critical Friend might suggest one or two desirable actions to the individual of which the individual may decide upon the most preferred.

Enlight is what the Critical Friend uses when aiming to teach a user about themselves in relation to others and to the world, as well as taking the role of being a "non-expert with expertise" or a "non-therapist with therapeutic abilities". For example, if a user seeks answers on e.g. practical or theoretical matters, the Critical Friend might provide reflections or perspectives with the purpose of further developing the interaction.

Enable is about providing tools, such as the ability to self-reflect, in order for an individual to develop a basis from which they can develop themselves. For example, a Critical Friend might identify a situation where the user seeks to share information, where the Critical Friend acts with the purpose of allowing the individual to do so.

Challenge is about challenging an individual about notions such as reasoning and assumptions. Challenging reasoning can be to question how the individual has come up with a certain thought, and question its implications. For example, the Critical Friend identifies a controversial thought, and challenges the user with conflicting facts based on science.

Criticize is about being able to give feedback, both constructive and nonconstructive, as well as providing critical insights that may cause uneasiness. For example, an individual might share information about a situation where the individual was acting malicious, where the Critical Friend might question why the user behaved in such a manner.

Honesty is about acting in an honest way, such as speaking your mind, focusing on what you want to say rather than what the user will feel, as well as expressing (possibly) negative things that are actually true. For example, the Critical Friend might be honest about disliking a certain comment made by the user.

Additionally, data retrieved from interviews helped to form factors influencing moral aspects of the formal definition of the Critical Friend. The Critical Friend needs to consider essential factors that highly influence the interaction. If the Critical Friend avoids considering the following factors, there is a possibility that e.g. being friendly and having bad timing to a new friend might instead be perceived as being critical. Thus, the 3 moral factors influencing the Critical Friend dimensions are presented in the following.

Temporal considers time as a factor, such as when to say/not to say something, at what point to say/not to say something, and how long/short-term influences temporal aspects of the interaction. For example, if a user has just started interacting with a Critical Friend, and the Critical Friend says something critical, it is likely that it will have a negative influence on the interaction/end it.

Relational has to do with factors such as trust which needs to be established to be able to have a relation, or faith which relates to what levels of faith an individual has towards a Critical Friend. Relational is a key factor to consider for the Critical Friend, and it is vital for the Critical Friend to understand

factual relations. For example, if trust has not been established between a Critical Friend and a user, it is likely that this will constrain and influence the interaction in a negative way.

Situational has to do with situations as a factor, where balancing between being Critical and Friendly is necessary. The Critical Friend needs to be able to consider the situation. For example, being friendly or critical in a certain (wrong) situation may be misperceived, and the user might understand the behavior as ingenuine, which may influence a relation in a negative way.

4.2 Heuristic Critical Friend scenarios

Additionally, interviewees were presented with the 9 Critical Friend Dimensions and 9 situational contexts. The interviewees were asked to form short scenarios, initially deciding on a specific context, in combination with using 9 dimensions to form short scenarios. Scenarios composition varied depending on the expert; Scenario 1-2 was formed by the human-computer interviewee (Figure 4), scenario 3-4 was formed by the teacher interviewee, and scenarios 6-7 was formed by the psychologist interviewee. The philosophy professor did not form any scenarios. Interviews resulted in six heuristic CF scenarios. Scenarios have the same relational context 'social companion', but differ in situational contexts.

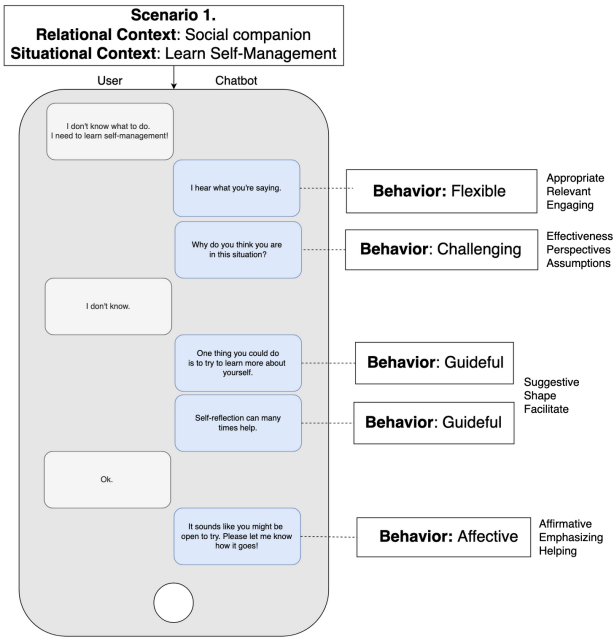

Figure 4: Scenario 1.

4.3 Temporal model of the Critical friend

Temporal model is based on extensive qualitative data, and participants' ratings of temporal suitability for the 9 dimensions for a comfortable interaction, which was included as a direct question in the quantitative assessment. Figure 5 illustrate a normalized and aggregated result as a transition system, where Critical (Challenge, Criticize, Honest), Neutral (Guideful, Enlight, Enabling) and Friend (Altruistic, Affective, Flexible) are compiled into three main dimensions to increase readability (Figure 5). For example, if a majority of transitions prioritized Altruistic, Affective or Flexible, this was considered as a transition prioritizing a Friend state.

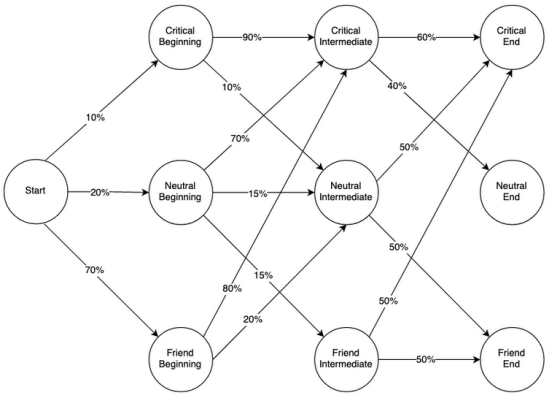

Figure 5: Temporal model in terms of a transition system

Looking at the transition system, a set of heuristics or rules representing (un)allowed transitions, as well as more desirable transitions are defined, presented in following paragraph.

1. Prioritize friend, if not yet been friend **2.** Prioritize critical, if have been friend **3.** Avoid neutral states, unless only been critical until the end **4.** If not yet started, avoid critical and neutral in a beginning state **5.** If critical in a beginning state, avoid friend and neutral in an intermediate state **6.** If neutral in a beginning state, avoid friend and neutral in an intermediate state **7.** If friend in a beginning state, avoid friend and neutral in an intermediate state **8.** If critical in intermediate state, avoid friend in an end state **9.** If neutral in an intermediate state, avoid neutral in an end state **10.** If friend in an intermediate state, avoid neutral in an end state.

Transition system is to be viewed as heuristics for strategic interactions between the Critical Friend and an user. The heuristics also provide a set of rules as presented, which is to be understood as abilities to maintain a comfortable interaction between the Critical Friend and the user.

4.4 Assessment of Socio-Moral Decision-Making

Following results reflects participants' ratings of the 9 dimensions in 6 heuristic Critical Friend Scenarios as a 'flow' (Figure 6). Results show Scenario 2 as the most deviant 'flow', e.g. 'Criticizing'-dimension which reflected high mean values (M=6.04, SD=1.18), suggesting this dimension as present. Scenario 2 also shows low mean values in a majority of dimensions, suggesting levels of Criticizing influenced participants' overall perception. Another result to highlight is Scenario 3, where ratings for 'Challenge'-dimension showed relatively low values (M=1.81, SD=1.70), suggesting participants perceived 'Challenge'-dimension as less present in this scenario.

Further, subscales capturing moral aspects of human-AI interaction was used to capture individuals' perception of

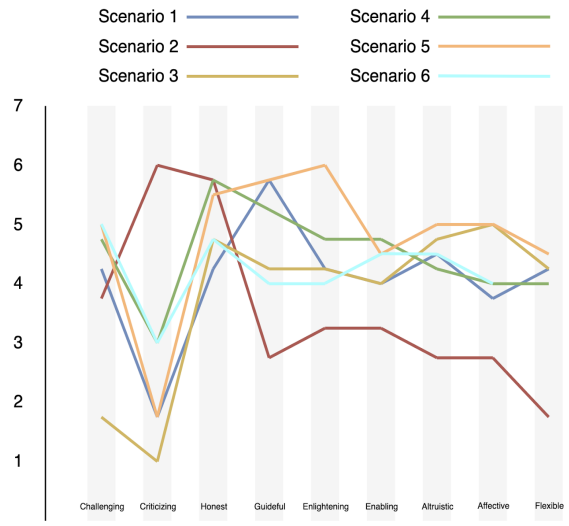

Figure 6: Illustration of flow for 9 dimensions in 6 CF scenarios.

moral decision-making in the different scenarios. Descriptive values were relatively balanced. In general, values were close to each other. However, results suggest that participants perceived scenario 1 (M=4.91, SD=1.06) and scenario 5 (M=5.23, SD=1.07) higher. This suggests that participants perceived scenario 1 and 5 as having acceptable levels of moral trust in the displayed interaction. Further, participants rated scenario 2 (M=3.51, SD=1.36) lowest, suggesting that scenario 2 was less preferred considering moral trust in displayed behavior. Thus, looking at Scenario 2, a majority of the 9 dimensions had low ratings, meaning that low levels on a majority of dimensions is perceived as less morally decision-making.

5 Discussion

The aim of this study was to 1) learn about the Critical Friend, and to characterize a formal definition of the Critical Friend, and 2) to learn about and assess socio-moral decision-making of the Critical Friend. A set of qualitative steps was carried out in order to answer RQ1); a literature review, online data collection and interviews, resulting in a knowledge model in the form of an ontology. Secondly, to answer RQ2), a quantitative measurement was conducted using an online questionnaire, resulting in a temporal model in the form of a transition system, assessing socio-moral decision-making of the Critical Friend in a precise way. Results from RQ1 and RQ2 form an initial formal framework of the Critical Friend and Critical Friend socio-moral decision-making.

5.1 Comparison

This study took inspiration from the Moral Machine experiment that focuses on decision-making in moral dilemmas [Awad *et al.*, 2018]. There are especially two comparisons to be made.

First, the self-driving car should be viewed as a self-driving 'actor', meaning that moral decision-making is more complex than a must-choose scenario. This has been demonstrated in this study, where a set of scenarios was used as tools for assessing socio-moral decision-making. In contrast to the Moral Machine experiment, these scenarios are constituted by a short dialogue, with a set of utterances based on qualitative data. These scenarios can be further utilized to capture the complex nature of socio-moral decision-making in a range of contexts, and should be viewed as an extension of using moral dilemmas to assess moral decision-making [LaCroix, 2022].

Secondly, the outcome of the Moral Machine experiment always results in life-or-death. Indeed, results from this study are very usable when trying to generalize morality in machines. However, is it really satisfactory that a majority, i.e. for some, are satisfied with the outcomes of the self-driving car's decision-making? To be able to view such results as applicable, we need to capture what is satisfactory for all. The proposed formal framework in this study provides a more sensitive assessment of socio-moral decision-making, and allows for a more generalizable conceptualization applicable for all, considering key terms, i.e. fairness.

5.2 Limitations

There are notable limitations in this study. Firstly, scenarios are heuristic, meaning that scenarios rely on the design done by a set of experts. Researchers, designers and engineers should strive to include a large variety of individuals to make models compatible for as many individuals as possible. It is likely that a younger/older sample would generate different results. Indeed, it is also likely that other factors influence results, which are not covered in this study, e.g. mood [Forgas *et al.*, 1984]. Additionally, different factors such as cultural or gender bias has not been covered in this study. Thus, results should be treated carefully, and further work needs to be done before being able to generalize results presented in this study.

6 Conclusion and Future Work

Results from this study forms an initial formal framework of the Critical Friend that may be used to characterize Critical Friend behavior for autonomous systems, or for autonomous systems to assess their own levels of Critical Friend behavior, or to assess Critical Friend behavior in already implemented systems. A possible way is to utilize results by considering recent research on automated planning for dialogue systems [Botea *et al.*, 2019]. Implementing such work requires libraries of information, i.e. knowledge bases such as ontologies, provided by this study. Additionally, such efforts would aid dialogue designers to introduce control in autonomous behavior [Muise *et al.*, 2019]. Lastly, the area of computational ethics still poses many open questions regarding what autonomous systems are and what they should be able to do. Recent research highlights the need for developing precise formal models in order to capture moral decision-making in autonomous systems [Awad *et al.*, 2022]. However, there is no consensus on how these abilities should look like. Thus, researchers from various disciplines are encouraged to utilize generated results presented in this study, to further shape autonomous behavior in responsible ways.

Acknowledgments

We thank the anonymous reviewers for their valuable and useful comments.

References

[Awad *et al.*, 2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[Awad *et al.*, 2022] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, MJ Crockett, Jim AC Everett, Theodoros Evgeniou, Alison Gopnik, Julian C Jamison, et al. Computational ethics. *Trends in Cognitive Sciences*, 2022.

[Baskerville and Goldblatt, 2009] Delia Baskerville and Helen Goldblatt. Learning to be a critical friend: From professional indifference through challenge to unguarded conversations. *Cambridge Journal of Education*, 39(2):205–221, 2009.

[Botea *et al.*, 2019] Adi Botea, Christian Muise, Shubham Agarwal, Oznur Alkan, Ondrej Bajgar, Elizabeth Daly, Akihiro Kishimoto, Luis Lastras, Radu Marinescu, Josef Ondrej, et al. Generating dialogue agents via automated planning. *arXiv preprint arXiv:1902.00771*, 2019.

[Carlson, 2009] Brian Carlson. School self evaluation and the critical friend perspective. *Educational Research and Reviews*, 4(3):078–085, 2009.

[Carlson, 2015] Elisabeth Carlson. Critical friends: a way to develop preceptor competence?, 2015.

[Cervantes *et al.*, 2016] José-Antonio Cervantes, Luis-Felipe Rodríguez, Sonia López, Félix Ramos, and Francisco Robles. Autonomous agents and ethical decision-making. *Cognitive Computation*, 8(2):278–296, 2016.

[Charisi *et al.*, 2017] Vicky Charisi, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Sombetzki, Alan FT Winfield, and Roman Yampolskiy. Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*, 2017.

[Chun Tie *et al.*, 2019] Ylona Chun Tie, Melanie Birks, and Karen Francis. Grounded theory research: A design framework for novice researchers. *SAGE open medicine*, 7:2050312118822927, 2019.

[Costa *et al.*, 1993] Arthur L Costa, Bena Kallick, et al. Through the lens of a critical friend. *Educational leadership*, 51:49–49, 1993.

[Dahlgren *et al.*, 2006] Lars Owe Dahlgren, Björn E Eriksson, Hans Gyllenhammar, Maarit Korkeila, Annika Sääf-Rothoff, Annika Wernerson, and Astrid Seeberger. To be and to have a critical friend in medical teaching. *Medical education*, 40(1):72–78, 2006.

[Deuchar, 2008] Ross Deuchar. Facilitator, director or critical friend?: Contradiction and congruence in doctoral supervision styles. *Teaching in Higher Education*, 13(4):489–500, 2008.

[DiCicco-Bloom and Crabtree, 2006] Barbara DiCicco-Bloom and Benjamin F Crabtree. The qualitative research interview. *Medical education*, 40(4):314–321, 2006.

[Dignum, 2018] Virginia Dignum. Ethics in artificial intelligence: introduction to the special issue, 2018.

[Eccles and Arsal, 2017] David W Eccles and Güler Arsal. The think aloud method: what is it and how do i use it? *Qualitative Research in Sport, Exercise and Health*, 9(4):514–531, 2017.

[Etikan *et al.*, 2016] Ilker Etikan, Sulaiman Abubakar Musa, Rukayya Sunusi Alkassim, et al. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1):1–4, 2016.

[Evans, 2015] Scotney D Evans. The community psychologist as critical friend: Promoting critical community praxis. *Journal of Community & Applied Social Psychology*, 25(4):355–368, 2015.

[Forgas *et al.*, 1984] Joseph P Forgas, Gordon H Bower, and Susan E Krantz. The influence of mood on perceptions of social interactions. *Journal of Experimental Social Psychology*, 20(6):497–513, 1984.

[Gibbs and Angelides, 2008] Paul Gibbs and Panayiotis Angelides. Understanding friendship between critical friends. *Improving schools*, 11(3):213–225, 2008.

[Holden, 1997] Gary Holden. 'challenge and support': the role of the critical friend in continuing professional development. *The Curriculum Journal*, 8(3):441–463, 1997.

[LaCroix, 2022] Travis LaCroix. Moral dilemmas for moral machines. *AI and Ethics*, pages 1–10, 2022.

[Lu *et al.*, 2022] Guang Lu, Martin Kubli, Richard Moist, Xiaoxiao Zhang, Nan Li, Ingo Gächter, Thomas Wozniak, and Matthes Fleck. Tough times, extraordinary care: A critical assessment of chatbot-based digital mental healthcare solutions for older persons to fight against pandemics like covid-19. In *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 735–743. Springer, 2022.

[Macbeath and Jardine, 1998] John Macbeath and Stewart Jardine. I didn't know he was ill–the role and value of the critical friend. *Improving Schools*, 1(1):41–47, 1998.

[MacPhail *et al.*, 2021] Ann MacPhail, Deborah Tannehill, and Rebecca Ataman. The role of the critical friend in supporting and enhancing professional learning and development. *Professional Development in Education*, pages 1–14, 2021.

[Martin and Russell, 2018] Andrea K Martin and Tom Russell. Supervising the teacher education practicum: A self-study with a critical friend. *Studying Teacher Education*, 14(3):331–342, 2018.

[Muise *et al.*, 2019] Christian Muise, Tathagata Chakraborti, Shubham Agarwal, Ondrej Bajgar, Arunima Chaudhary, Luis A Lastras-Montano, Josef Ondrej, Miroslav Vodolan, and Charlie Wiecha. Planning for goal-oriented dialogue systems. *arXiv preprint arXiv:1910.08137*, 2019.

[Özek *et al.*, 2012] Yvonne Hultman Özek, Gudrun Edgren, and Katarina Jandér. Implementing the critical friend method for peer feedback among teaching librarians in an academic setting. *Evidence Based Library and Information Practice*, 7(4):68–81, 2012.

[Petroelje Stolle *et al.*, 2019] Elizabeth Petroelje Stolle, Charlotte Frambaugh-Kritzer, Anne Freese, and Anders Persson. Investigating critical friendship: Peeling back the layers. *Studying Teacher Education*, 15(1):19–30, 2019.

[Schuck and Russell, 2005] Sandy Schuck and Tom Russell. Self-study, critical friendship, and the complexities of teacher education. *Studying Teacher Education*, 1(2):107–121, 2005.

[Sjögren and Köhler, 2021] Elaine Sjögren and Anita Kärner Köhler. The critical friend–a way to develop as a tutor in problem-based learning groups. *Högre utbildning*, 11(3), 2021.

[Skjuve *et al.*, 2021] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601, 2021.

[Smedley and Coulson, 2021] Richard M Smedley and Neil S Coulson. A practical guide to analysing online support forums. *Qualitative Research in Psychology*, 18(1):76–103, 2021.

[Storey and Wang, 2017] Valerie A Storey and Victor CX Wang. Critical friends protocol: Andragogy and learning in a graduate classroom. *Adult Learning*, 28(3):107–114, 2017.

[Swaffield, 2004] Sue Swaffield. Critical friends: Supporting leadership, improving learning. *Improving schools*, 7(3):267–278, 2004.

[Swaffield, 2008] Sue Swaffield. Critical friendship, dialogue and learning, in the context of leadership for learning. *School leadership and Management*, 28(4):323–336, 2008.

[Ullman and Malle, 2019] Daniel Ullman and Bertram F Malle. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 618–619. IEEE, 2019.

[Wasil *et al.*, 2021] Akash R Wasil, Emma H Palermo, Lorenzo Lorenzo-Luaces, and Robert J DeRubeis. Is there an app for that? a review of popular apps for depression, anxiety, and well-being. *Cognitive and Behavioral Practice*, 2021.

[Winkle *et al.*, 2021] Katie Winkle, Ryan Blake Jackson, Alexandra Bejarano, and Tom Williams. On the flexibility of robot social identity performance: benefits, ethical risks and open research questions for hri. In *HRI Workshop on Robo-Identity*, 2021.

[Winkle *et al.*, 2022] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Brščić, Iolanda Leite, and Tom Williams. Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pages 120–129, 2022.