FAU
COLLEGE OF ENGINEERING & COMPUTER SCIENCE
Florida Atlantic University

A Probabilistic Model for Personality Trait Focused Explainability

by

Mohammed Alharbi
malharbi2016@fau.edu

Shihong Huang
shihong@fau.edu

David Garlan
garlan@cs.cmu.edu

Outlines

1. Introduction
2. Motivation
3. Approach
4. A Probabilistic Model for Personality Trait Focused Explainability Framework
5. Example Scenario
6. Results and Analysis
7. Conclusion

Introduction

- **A co-adaptation system** is symbiotic human-in-the-loop system where human-system cooperation is required in achieving shared goals, and system and human actions mutually impact each other's behavior in accomplishing coordinated tasks [1].
- **Explainability** refers to the degree to which a software system's actions or solutions can be understood by humans [2,3].
- **The Need For Explainability** [2,3]:
  1. **Explain to Justify**: we use explanations to *justify* results of decisions to the human, particularly when decisions are made suddenly.
  2. **Explain to Control:** explanations can help not only to justify, but also to control and prevent systems from going wrong.
  3. **Explain to Improve:** *improving* the systems utility continuously through human involvement.
  4. **Explain to Discover:** *discovering* and gathering new facts that help human to learn and to gain knowledge.

Problem Statement

- Effective explanations to humans can improve effectiveness of system-human collaborative systems [4].
- Key issues to resolve:
  - What should the **content** of an explanation be?
  - How **frequently** should explanations be given?
  - **How do the answers to these questions vary from person to person?**
  - Can we **mechanize** the decision process that a system uses in determining the answers to these questions?
- Our approach:
  - Use relevant personality traits to capture differences in people.
  - Formalize these so that a system can automatically determine appropriate amounts of explanation.

Personality Traits

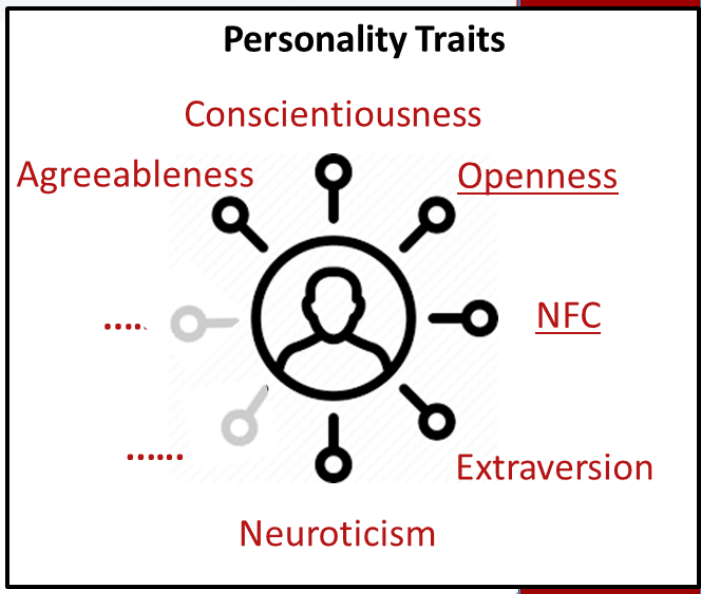
- We chose the traits Need for Cognition (NFC) and Openness to experience, since there is a direct relationship between *NFC* and *explainability* and between *openness* and *capability* in OWC [5].

A. Need for Cognition [6]:

- Need for Cognition (NFC) is defined as the "individual's tendency to engage in and enjoy effortful cognitive tasks."
- People with higher NFC levels typically prefer more detail.
- A score above 80 is generally considered to be High NFC (or high personality trait), and below 50 is Low NFC [6].

B. Openness to experience:

- Open people tend to be intellectually curious, creative and imaginative [7].
- Open people have a high openness to embrace new things, fresh ideas, and novel experiences.

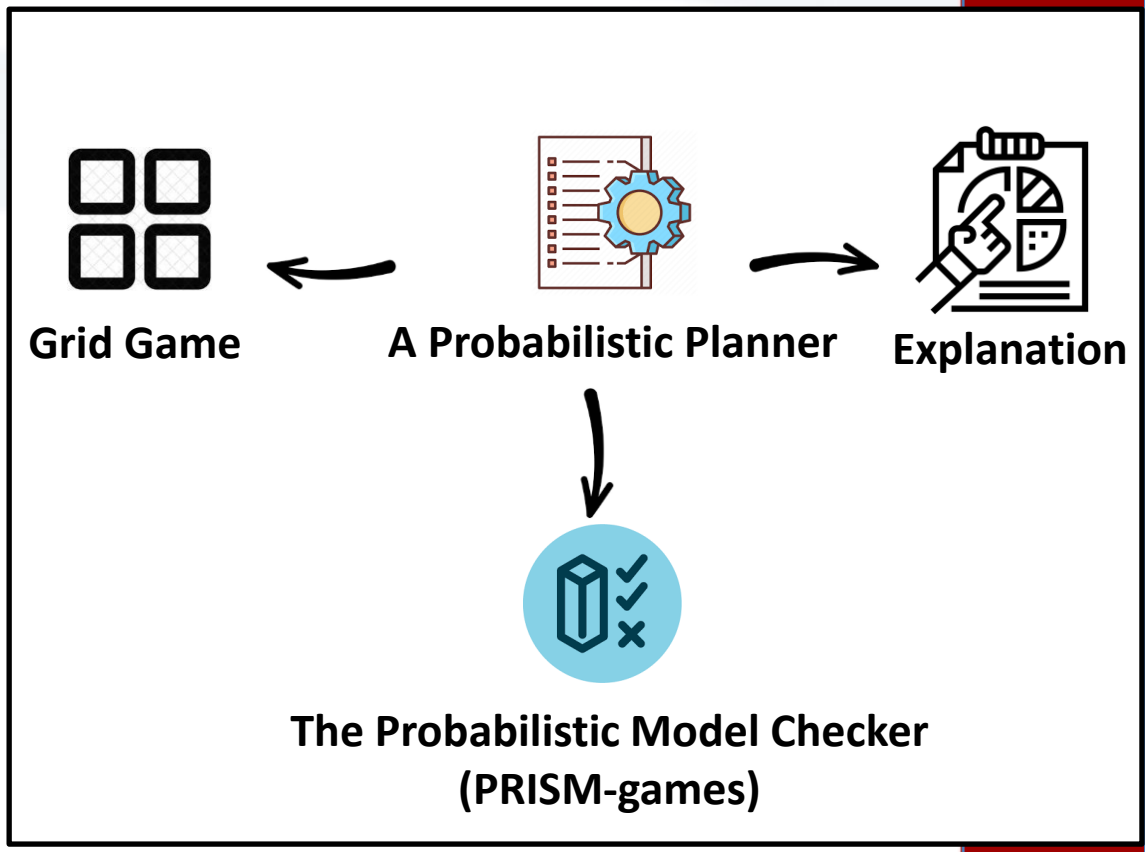


Research Objectives

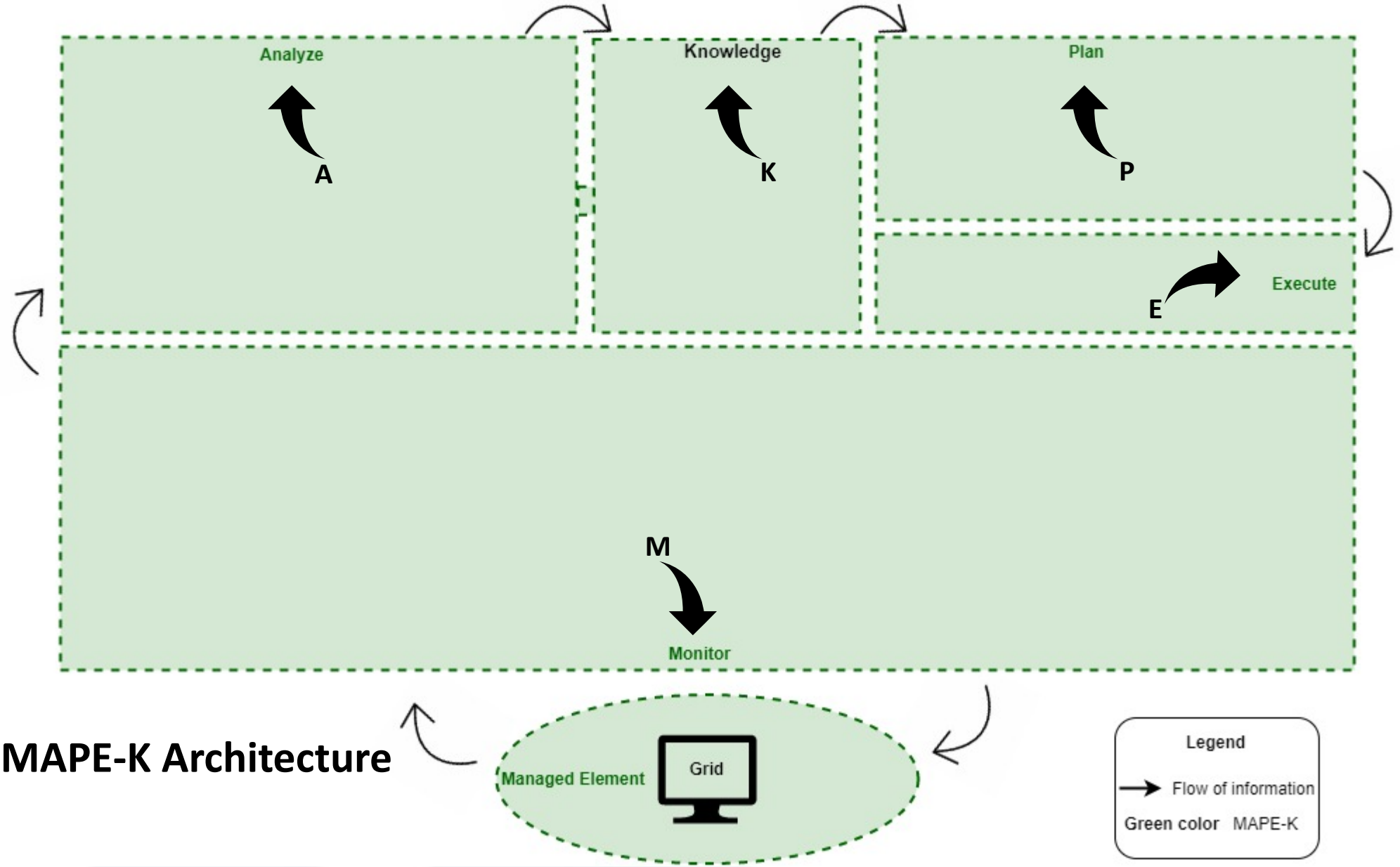
- We aim to define a formal framework for reasoning about **how systems should determine the ideal amounts of the explanations that should be considered**.
- **We want to answer the following research question:**
  - **How to use knowledge about an individual's personality traits to improve the overall system utility?**
- **The main contribution of this research is:**
  - Define a formal framework that **incorporates human personality traits** and guides adaptive human-in-the-loop systems to decide how **much explanation** should be given in order to **improve system utility**.

Approach

- We utilize **a probabilistic planner** [8] to determine the optimal amount of explanation according to those personality traits.
- We use **explanation** as a tactic (or action) that systems can use to improve the effectiveness of human-system co-adaptation based on human personality traits.
- The **probabilistic model checker (PRISM-games)** is utilized for formally model our approach [9].
- We defined the **Grid** as a game that embodies a representative scenario for human-system co-adaptation.

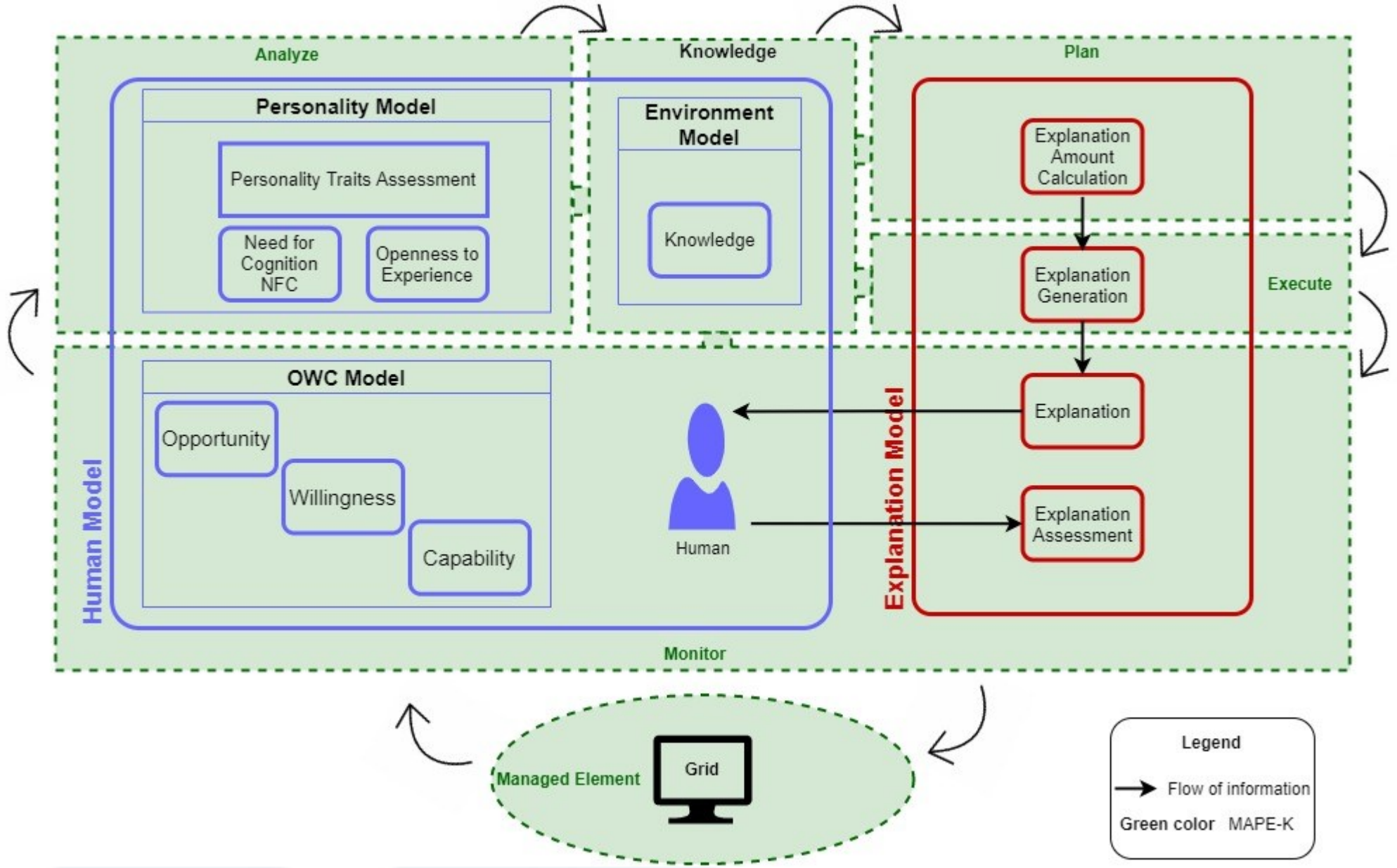


A Probabilistic Model for Personality Trait Focused Explainability Framework



The MAPE-K Architecture

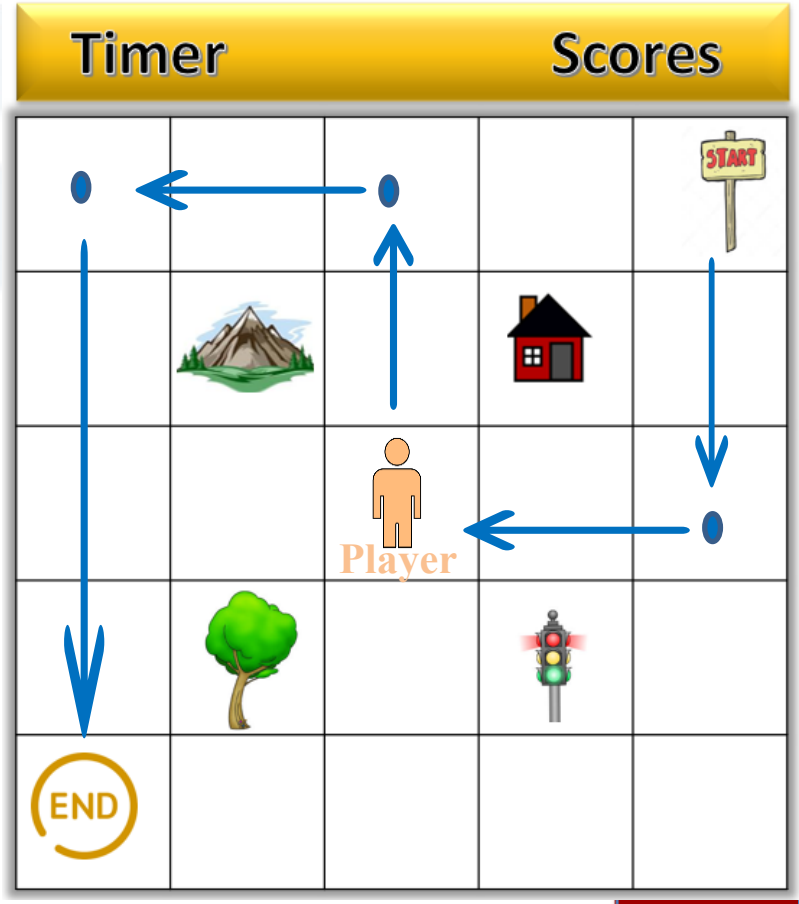
A Probabilistic Model for Personality Trait Focused Explainability Framework



Representative Scenario

Grid Game

- We defined the **Grid** game (a virtual game) as a game that embodies a representative scenario for human-system co-adaptation.
- The system S instructs a player P verbally to move on a 5×5 grid from the top right corner (start) to the bottom left corner (end).
- **Game objectives:**
  - Human follow the system instruction through a certain path within certain maximum amount of time (60 seconds).
  - Minimize the time *t* to complete the task.
  - Traverse an optimal number of blocks *B* to complete the end-to-end task, avoiding obstacles.
- **Game rules:**
  - The player can move either horizontally or vertically.
  - Game scores (100 points): points are deducted for traversing extra blocks or moving into or through obstacle squares.



Grid Game Tactics

- **The Grid game can generally use five tactics for interacting with the player:**

Model	Categories	Tactics	Role	Example
System	Less Explanation	lessExplain (lExp)	Commands the human to carry out an action	"Go 2 blocks left" "Move south 4 blocks"
	More Explanation	moreExplain (mExp)	The system further explains information when the human is confused and loses track	"You will go between a house and traffic light" "You go straight, and you see a car on your left side"
Human	Clarification Request	Check (Chk)	The human requests the system to confirm information that they not entirely sure about	"North?" "Should I continue above the tree?"
	Feedback	playerFeedback (pF)	Human feedback is collected about his satisfaction for each given explanation	Helpful, Not helpful, Neutral
	Acknowledgement	confirm (conf)	The human confirms information and follows the instructions.	"Yeah", "Thanks" "Okay"

Utility Attributes

- **The four utility attributes of the game are:**

Game Scores

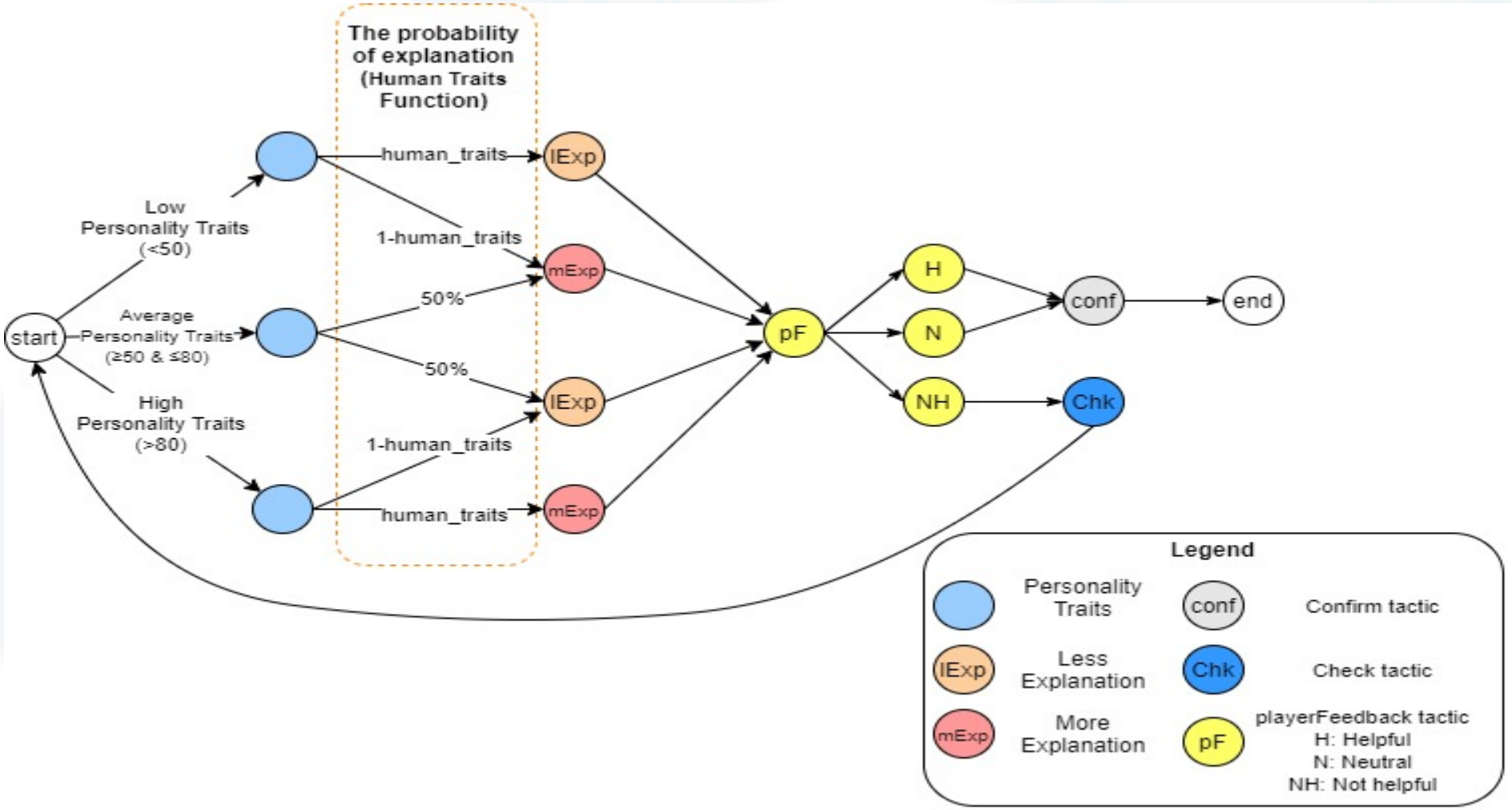
1. **RequiredTime (t):** the total elapsed time for completing the game.
2. **Blocks (B):** the number of the blocks traversed to complete the task.

Explainability Attributes

3. **LengthOfExplanations (xL):** the amount of delay (or time) required to explain.
4. **ExplainEfficiency (xE):** a measurement that determines how happy the player is with the given explanations (Helpful, Not helpful, Neutral).

The Strategy

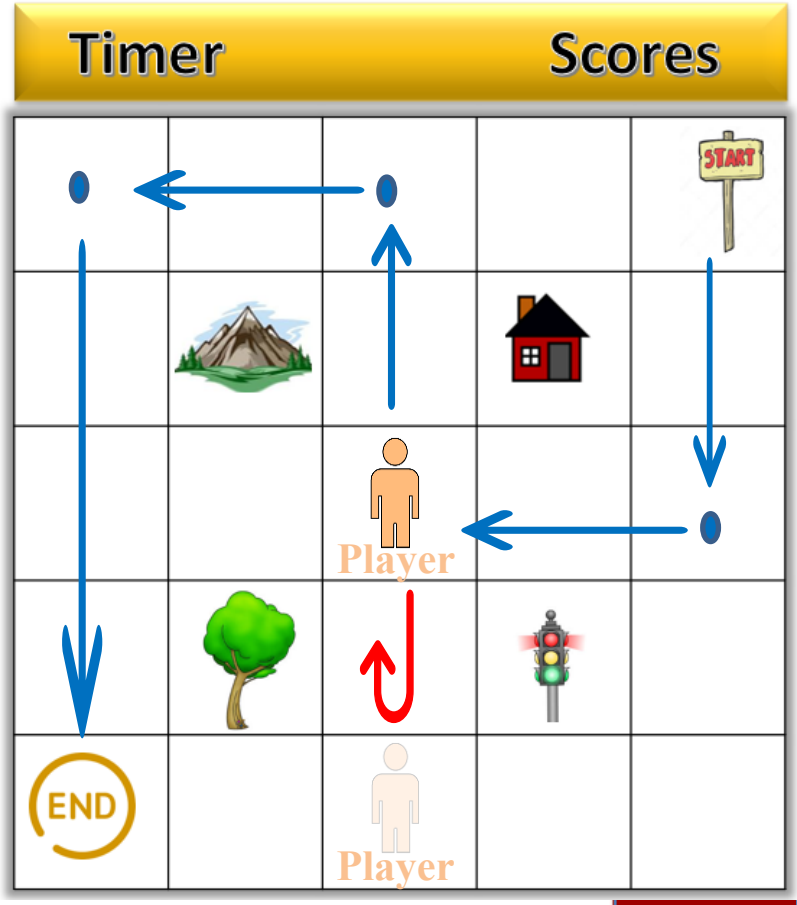
$$Human\ Traits = \frac{Openness + NFC}{Openness^{max} + NFC^{max}}$$



Example Scenario

- **An example dialogue between the system (S) and a human (H) :**

	Tactics	Time	pF
S: Can you go 2 blocks down?	(lExp)	3s	Helpful
H: Yeah	(conf)	3s	
S: Then go 2 blocks left.	(lExp)	3s	Not helpful
H: Could you repeat that?	(Chk)	3s	
S: Go west. You will go between a house and traffic light.	(mExp)	6s	Helpful
H: Okay	(conf)	3s	
S: Go after that 2 blocks up.	(lExp)	3s	Neutral
H: The human is on the wrong track			
S: No, not south. You go north	(mExp)	6s	Helpful
H: Okay	(conf)	3s	
S: Go 2 blocks left	(lExp)	3s	Neutral
…….			
S: Go south 4 blocks.	(lExp)	3s	Helpful
H: Okay, thanks a lot.	(conf)	3s	



Scores: 75

t= 42s B=15

xL= 27 s xE ≈43

Results and Analysis

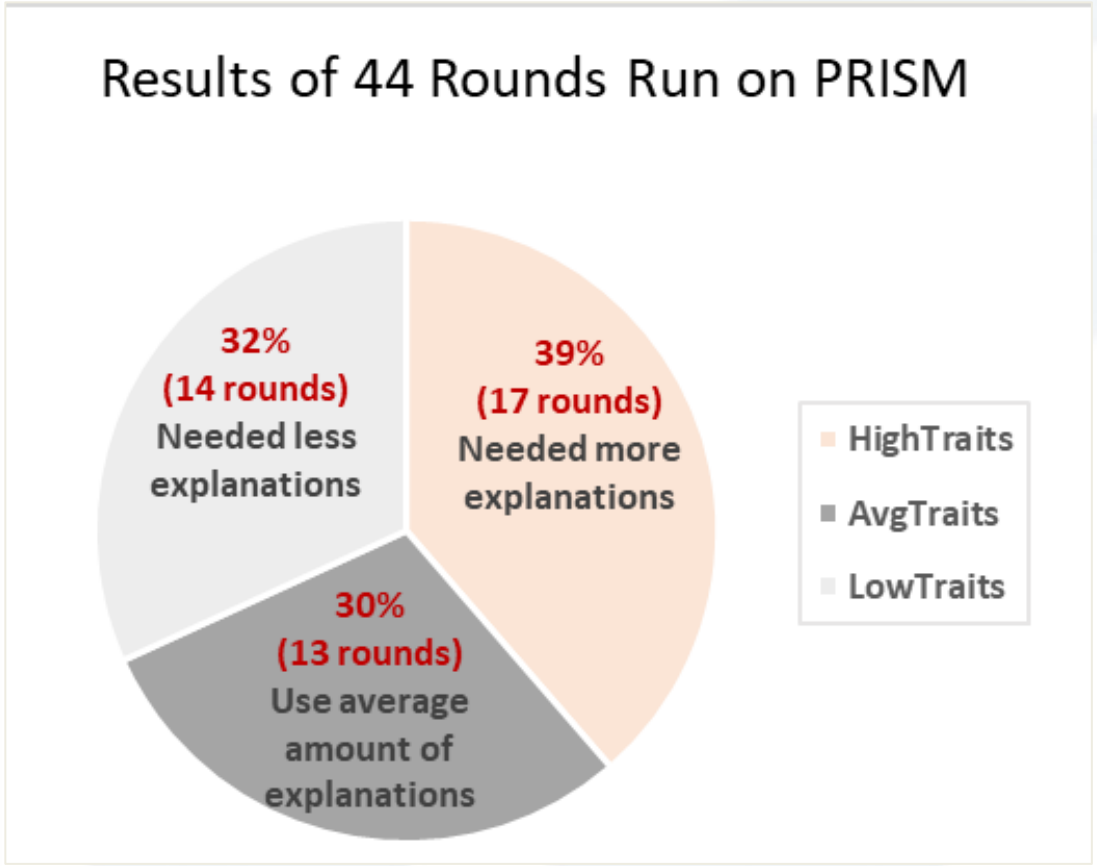
Results and Analysis

FLORIDA ATLANTIC UNIVERSITY FAU

#	Human Traits		Combined Traits	Utilities		
	Openness	NFC		LengthOfExplanations (xL)	ExplainEfficiency (xE)	Scores
1	75	90	82.5	27	28.5	93.4
2	100	100	100	21	50	96.7
3	50	90	70	15	80	100
4	95	30	62.5	21	50	96.7
5	95	85	90	15	80	100
6	45	88	66.5	15	80	100
31	47	47	47	33	12.5	90
32	83	83	83	21	50	96.7
33	22	19	20.5	15	80	100
34	96	77	86.5	15	80	100
35	69	55	62	27	28.5	93.4
36	39	11	25	21	50	96.7
37	33	19	26	15	80	100
38	17	15	16	27	28.5	93.4
39	9	30	19.5	21	50	96.7
40	49	29	39	15	80	100
41	51	71	61	15	80	100
42	93	100	96.5	21	50	96.7
43	100	90	95	15	80	100
44	81	80	80.5	15	80	100
min	9	11	16	15	-12.5	90
max	100	100	100	36	80	100
avg	66.02	61.98	64	20.39	57.07	97.23

	Avg Utilities		
	xL	xE	Scores
HighTraits > 80	21.18	53.50	91.39
AvgTraits > 50<80	19.15	62.27	97.71
LowTraits < 50	20.57	56.57	96.92

Results and Analysis



- From the results we can conclude that:
  - A human with high personality traits needs more detailed information (i.e., explanations),
  - While a human with low personality traits needs less detailed explanation.
- These conclusions are all consistent with psychology studies [6][10].

Conclusion

- **Summary:**
  - In this research we presented a formal framework that **incorporates human personality traits** as one of the important elements in guiding system decision-making about the **proper amount of explanation** that should be given to the human to **improve overall system utility**.
  - We use **probabilistic model analysis (SMG)** to determine how to utilize explanations in an effective way.
  - **Grid** was developed to illustrate our approach, to represent scenarios for human-system co-adaptation.
- **Future work:**
  - **Conducting an empirical study** to validate these models on actual real-world systems with humans in the loop.
  - **How explanations modality should be presented: graphically, textually, verbally?**

References

[1] E. Lloyd, S. Huang, and E. Tognoli, "Improving Human-in-the-Loop Adaptive Systems Using Brain-Computer Interaction," Proceedings - *2017 IEEE/ACM 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, SEAMS 2017. pp. 163–174, 2017.

[2] G. Vilone and L. Longo, "Explainable Artificial Intelligence: a Systematic Review." arXiv preprint arXiv:2006.00093, 2020.

[3] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6. pp. 52138–52160, 2018.

[4] M. Alharbi and S. Huang, "A Survey of Incorporating Affective Computing for Human-System Co-adaptation," in *Proceedings of the 2020 The 2nd World Symposium on Software Engineering*, 2020, pp. 72–79.

[5] D. Eskins and W. H. Sanders, "The multiple-asymmetric-utility system model: A framework for modeling cyber-human systems," *Proc. 2011 8th Int. Conf*. Quant. Eval. Syst. QEST 2011, pp. 233–242, 2011.

[6] C. J. Sadowski and H. E. Cogburn, "Need for cognition in the big-five factor structure," *Journal of Psychology: Interdisciplinary and Applied*, vol. 131, no. 3. pp. 307–312, 1997.

[7] R. R. McCrae, "Openness to Experience as a Basic Dimension of Personality," *Imagin. Cogn. Pers.*, 1993.

References

[8] M. Kwiatkowska, G. Norman, and D. Parker, “PRISM 4.0: Verification of probabilistic real-time systems,” in Lecture Notes in Computer Science (including subseries Lecture Notes in *Artificial Intelligence and Lecture Notes in Bioinformatics*), 2011.

[9] M. Kwiatkowska, G. Norman, D. Parker, and G. Santos, “PRISM-games 3.0: Stochastic Game Verification with Concurrency, Equilibria and Time,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020.

[10] R. E. Petty, J. T. Cacioppo, R. E. Petty, J. A. Feinstein, and W. B. G. Jarvis, “Dispositional Differences in Cognitive Motivation : The Life and Times of Individuals Varying in Need for Cognition Dispositional Differences in Cognitive Motivation : The Life and Times of Individuals Varying in Need for Cognition,” *Psychol. Bull.*, vol. 119, no. August, pp. 197–253, 2015.

Thank You !

Mohammed N. Alharbi
malharbi2016@fau.edu