ATTENTION TO CONTEXT IS WHAT xAI NEEDS: A FRAMEWORK FOR GENERATING CONTEXT-DEPENDENT REQUIREMENTS FOR HUMAN-CENTRIC EXPLAINABLE AI (xAI)

UNNATI PATEL

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the degree of

Doctorate in Philosophy
in Digital Transformation and Innovation

Faculty of Engineering, University of Ottawa

uOttawa

Abstract

As the rapid integration and widespread deployment of Artificial Intelligence (AI) systems across vital societal domains have yielded promising results, their opacity – the "black box" problem – poses significant barriers to trust in, accountability of, broad acceptance, and adoption of this transformative technology. While Explainable AI (xAI) seeks to address the black box problem, current initiatives often prioritize computational and algorithmic transparency over the actual explanatory needs of human users. Since explanations are ultimately for the benefit of humans, it is crucial to tailor them to their needs.

This interdisciplinary thesis addresses two major gaps that persist in current xAI development: an overemphasis on computational approaches due to the underutilization of Social Science and Humanities (SSH)-based knowledge on the concept of explanation; and insufficient attention to the role of context in shaping what constitutes an adequate explanation for diverse, non-AI expert stakeholders.

To bridge these gaps, this research project proposes the **Framework for Generating Context-Dependent Requirements (FGCR)**. The FGCR is a multidisciplinary practical guide and toolkit designed to guide xAI developers in identifying contextually suitable xAI solutions for diverse stakeholders by placing the *explainee* at the center of the xAI development cycle. It provides a structured process and accompanying suite of tools to support this goal across varied contexts. The framework integrates insights from social sciences, humanities, engineering requirements management, and xAI research. Its components include novel taxonomies (an SSH-based taxonomy of explanation (classifying the “what” and “how” of human explanation) and a taxonomy of the xAI landscape) alongside established engineering requirements management practices to identify what constitutes an adequate and contextually appropriate human-centric explanation.

The research was conducted in three phases. Phase 1 comprised a multi-disciplinary literature review synthesizing SSH-based concepts of explanation, the current landscape of xAI approaches (with

a focus on artificial neural networks (ANNs)), and engineering requirements management processes. Phase 2 involved the design and construction of the FGCR. Phase 3 undertook proof-of-concept testing of the FGCR using a qualitative case study of autonomous vehicle (AV) crashes, specifically, a real-world series of incidents in which Tesla's Autopilot system disengaged shortly before colliding with stationary first-responder vehicles. The study tested whether the FGCR could effectively categorize varied stakeholder inquiries and generate adequate xAI model requirements.

The FGCR was evaluated through a proof-of-concept qualitative study involving semi-structured interviews with 26 stakeholders representing a diverse cross-section of society including: pedestrians, drivers, legal experts, first responders, and AI developers. The study elicited stakeholders' explanatory needs and applied the FGCR's components to generate specific xAI model requirements. Findings from the proof-of-concept testing show that the FGCR can identify and classify varied explanatory needs and translate them into xAI model requirements and potential xAI solutions that meet those requirements.

Importantly, findings indicate that while the diversity of AI applications and the complexity of context both play a pivotal role in explanations and drive the need for tailored xAI solutions, it is the need of *explainees* that take primacy in determining an explanation that is both relevant and sufficient. Essentially, while technical accuracy is necessary, the adequacy of an explanation is primarily determined by the *explainee(s)'* context and intended use of the information.

This research contributes to the field by providing a structured and practical framework for xAI developers to move beyond a "one-size-fits-all" technical approach for development of transparent, and understandable human-centric AI systems. By bridging theoretical SSH and engineering requirements management knowledge, it offers a roadmap for creating explainable AI systems that are genuinely intelligible to humans they affect. The result is a practical framework and toolkit that prioritizes context, stakeholder diversity, and the needs of those affected by AI systems at the center of the xAI development cycle.

Table of Contents

List of Figures

List of Tables

Acknowledgements

A Village, A Ghost Cat, and the Long Scenic Route

Writing a PhD thesis is often described as a solitary marathon. In reality, it feels more like trying to assemble a high-precision aircraft engine while riding a unicycle through a library. At age 61, I have come to appreciate that the adages are true: "No man is an island," "It takes a village," and, most importantly, "Man makes plans and the Gods have a truly wicked sense of humour."

The Non-Human (and Non-Existent) Contributors

I began this section stumped. When I asked my supervisor, André Vellino, for guidance, he suggested that people generally thank those around them—including their cats. So, here goes, I would like to thank my cat for providing immeasurable emotional support throughout this endeavour. There is, however, one minor technicality: I don't have a cat. But in the spirit of academic thoroughness—and to satisfy my supervisor's advice—I felt it important to acknowledge the hypothetical feline that would have sat on my keyboard had it existed. I suspect its absence was actually a net positive for my word count, though, I apologize in advance to the readers and reviewers now facing down the considerable volume of pages that follow, as any PhD monograph opus is wont to produce.

In a similar vein, I must acknowledge the AI tools—not the civilization-reshaping, paradigm-obliterating forces their press releases would have you believe, but genuinely useful software that helped with actual tasks. In accordance with the University of Ottawa's AI policy, I used Large Language Models for both language refinement and certain aspects of data analysis – documented honestly, including their infallibility and pitfalls, in the chapters and appendices that follow. Their assistance with language refinement during the final edits of this thesis including checking sentence structure, suggesting synonyms, and generally preventing me from sounding like an outdated operating system was genuinely useful. And yes, the irony of using AI to avoid sounding like a machine is not lost on me. In the interest of fairness to the non-AI generative world: the Oxford English Dictionary; the ripped-and-

tattered 1970s-era copies of Webster's New Collegiate Dictionary and Collins English Dictionary on my bookshelf; and WordHippo. If seeking a better synonym for "additionally" is a crime, I am a repeat offender. WordHippo remains the unsung hero of this dissertation.

But enough about the tools—digital, and hypothetical feline alike. The **people** are what actually matter, and the best, as they say, is always saved for last.

The Human Village

My "village" is vast, spanning four continents and six decades. It begins with my parents, Kanti Patel (1932–2019) and Indu Patel (age 90 and still going strong!). Other family members include my sister Prerana and my nephew Anup. I wish to send out a heartfelt thank you to Anup Patel, who, despite his age, has stood by me and cheered me on with kindness through every "against the grain" life pivot I have made.

My parents instilled in me the immigrant's sacred creed: the value of education. There is a photograph from my first graduation (Electrical Engineering, U of T, 1988) that tells the whole story. In the photo, my father is the one physically holding my degree. A friend pointed out the look on his face—it was the look of a fiercely stern old-school immigrant man saying, "I know exactly what this piece of paper means for my daughter's future, and I am not letting go of it until we get home." He was right. That degree, and the MBA that followed, opened doors that would have otherwise been bolted shut. But while my father gave me the "practical" foundation, my mother gave me her love of books. Deep down, I was still that eight-year-old girl who arrived in England in 1972 after Idi Amin's expulsion of Indians from Uganda. At the time, I had fallen through the cracks of the system and was labelled "illiterate" by my first principal. Three months of intensive tutoring later, I didn't just learn to read; I fell in love with books, the worlds they opened up to me, and the irresistible pull of many, many libraries!

Because of them, I went on to spend twenty-five years as a senior manager in the aerospace and manufacturing industries. This career taught me many things, chief among them being that the "hair-on-fire" emergencies and "mission-critical" crises that corporations love to inflict upon their employees are,

in the grand scheme of things, significantly less important than connecting with other humans, the many small acts of actually savoring one's life, and of course, a rejuvenating nap.

The Mid-Life "Detour"

Upon accomplishing "Freedom 55"—or in my case, a slightly more ambitious "Freedom 50" (a term that my fellow Canadians of a certain vintage will recognize with either a nod of respect or a twinge of envy)—I decided it was time to stop managing other people's emergencies and start feeding my own curiosity. Having secured the luxury of time, I treated myself to the humanities, a pursuit of pure passion that turned into a decade-long, full-blown mid-life detour.

This PhD is my fifth degree, a fact I share not as a trophy count, but as evidence of a lifelong compulsion of curiosity that I have never quite been able to cure. The first two were deliberate: engineering and an MBA, the credentials that built a career and opened doors that would otherwise have been bolted shut. The last three have been something else entirely: sheer joy. The joy of reading, of following an idea until it leads somewhere unexpected, of discovering that curiosity, left unsupervised, has no natural ceiling. A forty-year runway, as it turns out, is exactly long enough to find your way back to what you loved first. About ten years ago, a conversation with a dear friend, Joanna Szurmach, at the UTM campus reignited a spark. I decided to pivot from the boardroom to philosophy, and in the process, ended up completing an HBA and an MA at the University of Toronto before landing here at the University of Ottawa. The four years of this PhD have been equal parts exhilarating and humbling—a rather fitting combination for a project that examines how humans (and machines) explain themselves.

What drew me to this particular topic was a tale of two AI moments, worlds apart in context but eerily linked in the questions they raised. The first was a report about Tesla automated vehicles deactivating Autopilot seconds before crashing into emergency vehicles—a case study in what happens when an AI “pilot” decides that “not my problem anymore” is an acceptable handoff strategy to its sometimes very surprised human co-pilot. The second was AlphaGo: The Movie, depicting Google DeepMind's AI program AlphaGo defeating Go world champions with moves so unexpected that they

were repeatedly described as “creative”, a word the world had previously reserved exclusively for humans. Two systems. Two domains. One nagging question: how does anyone—human or machine—explain what they just did, and why? That question became this thesis. The four years spent chasing it have been among the most rewarding of my career—a career that has already spanned more fields than most people's bookshelves. I hope the research that follows proves worthy of the journey.

I owe my deepest gratitude to my supervisor (and Dharma Teacher) Dr. André Vellino. His scholarly guidance was essential, but it was his "Zen-ness" and unconditional support that kept me tethered to reality. When the world felt like a chaotic data set, André remained the calmest variable in the equation. I am also grateful to my committee, Dr. Diana Inkpen and Dr. Vincent Bergeron, for their participation in the essential milestones of this process and for providing valuable feedback on my draft to keep this project moving forward.

The Ground Crew

A project of this scale requires "super-networkers," and I found them in Dr. Channarong Intahchomphoo and Paula Stellato. Channarong, thank you for being my steadfast companion through this process. Our walks along the Rideau Canal—where we strategized my way through the rigors of the DTI program—were fuel for my academic stamina. And Paula, thank you for your unwavering and patient support across four years of conversations, whether in person or over Zoom. Paula and Channarong were also instrumental in recruiting half of the 26 participants in my research study.

To those **26 anonymous participants**: I am especially grateful for your time and the invaluable insights you shared. To those of you who went above and beyond by helping recruit additional participants, you didn’t have to do that, and yet you did. A reminder that the most important variable in any research project is the goodwill of the humans involved—who, as it happens, are precisely what this research set out to put back at the center of AI development.

To my long-time companions from the U of T Electrical Engineering Class of 8T8—Peter Vanderheyden, Joe Fuda, Jeannette Chau, Eugene Girolimetto, Victor Ferreira, Greg Gayowsky (and

Bonnie Weston Gayowsky), and the rest of the GNO gang: Some friendships are not measured in frequency but in weight. We may see each other once in a while, but forty-plus years of knowing each other produces something rare: people who cheer you on through every unlikely pivot without once asking "but what is it good for?" In a world full of skeptics, that kind of unconditional enthusiasm is worth more than I can say. To Brandeen McDonald, a steadfast friend whose loyalty and support have followed me through every twist and turn of this life — thank you for never once wavering. And to Hairong Xu and Chantal Ripp, newer but no less cherished members of my Ottawa village: your enthusiasm and encouragement in these final stretches have been a genuine gift. Also, to my other support network in U of T's philosophy community–Jane Medeiros, Chris Longley (and Meryn Mercer), Phil Clark, Zach Irving, Lu-Vada Dunford, Brendan DeKennessey, Owen Ware, and many others, thank you for your encouragement especially through the tumultuous final year of that journey. To the balancers in my life, Meral Al-Ameer, Sandy Sajko, Jumanah Sbeih, and another Zen-master Terry Hodgkinson: Thank you for keeping my mind and body in equilibrium when the thesis threatened to tip the scales.

There are many more friends whose names are not etched here, especially those who supported me in the decades before this final academic chapter of my life's journey. To them: thank you for showing up, becoming confidants, and sticking by me through every endeavour and venue. You know who you are. Throughout my life, people have always, always, always shown up—I am a testament to the fact that no man is an island, and it does indeed take a village.

This PhD is the culmination of a rare and deeply rewarding journey that spans decades and continents. It bridges the worlds of engineering and technology development, management, philosophy, and now AI. I am profoundly grateful to the community that sustained me, and I hope that my research will make a positive impact on society.

P.S. — A correction to the record: I wrote above that I have no cat, which remains technically true. What I failed to mention is that I do have Ganesh—an African Grey parrot with the intelligence of

a two-year-old human, absolute certainty that he is the boss of this household, and a gift for derailing concentration at precisely the wrong moment. He did not help with this thesis. He interrupted it at will, interjected his unsolicited opinions at volume, and occasionally reviewed drafts by walking across the keyboard. And yet, when the thesis felt impossible, it was Ganesh who reminded me of what this entire research project was ultimately about: that at the center of any system worth building, there is a living being who needs to be heard, understood, and never, ever ignored. Ganesh required no explanation — only attention. And in loudly, persistently demanding it, he kept reminding me what the field is still learning: that attention is not all you need. Living beings (a.k.a. humans) need both attention and explanations. And they deserve explanations that actually make sense to them.

1 Introduction

Explainable Artificial Intelligence (xAI) is a field dedicated to providing humans with explanations that enable them to gain insight into and understand the underlying decision-making processes and outputs of complex and opaque AI systems. Since people are the ultimate beneficiaries of explanations, it is essential to pay attention to the contextual nature of their explanatory needs. What counts as an adequate explanation is wholly dependent on those seeking it *(explainees)*, and the situation. In short, context is what xAI needs, and it is therefore essential to create human-centric xAI solutionss that adequately meet the explanatory needs of the people they serve. This thesis provides a framework for generating context-dependent requirements (FGCR) that serves as a guide for identifying a suitable xAI solution for a given context. The FGCR prioritizes the contextual needs of stakeholders – the *explainees* – throughout the development cycle of human-centric xAI solutions. It provides a systematic means of eliciting and specifying these needs and mapping them into a set of requirements that drive the engineering development cycle of human-centric xAI models.

Explanations, by being an integral part of human communication, are a fundamental part of our lives. Despite their ubiquity and familiarity, at their core, they are incredibly complex. The ubiquity of explanations is evident in the fact that humans regularly and universally engage in the common act of seeking and providing explanations. We are constantly asking and answering questions to elicit explanations about many diverse phenomena: "*Why did our friends get divorced? Why did that car accident occur? Why did I get cancer? Why was I rejected for a bank loan? Why was there an attack on the Twin Towers on 9/11? How do batteries generate electricity? How do plants grow? Why is the economy failing?"* These questions yield different answers that reflect the diverse perspectives and circumstances of the individuals seeking and providing explanations. Our everyday lived experiences attest to the fact that explanations are not only ubiquitous and commonplace but also multifaceted and varied. Their complexity arises from the fact that they are highly dependent on the situation and the

explainee(s) – context matters! While their constant presence might make explanations seem straightforward, the act of explaining is far from simple. Everyday explanations are multi-predicated relational acts where someone *(explainer)* explains (communicates or provides) something *(explanans)* to someone else and/or oneself *(explainee)* about a phenomenon to be explained *(explanandum)*. Furthermore, the complexity of these relational acts is driven by context. For when it comes to grasping the context of explanations, what makes them so complex is that we have to consider multiple factors at once. An adequate explanation involves juggling the various elements of not only the situation and *explainee,* but also the sub-aspects of the time and place of the *explanandum*, *explainee* and *explainer.*

AI systems are essentially machines that perform acts requiring human intelligence and cognitive abilities such as problem solving, learning, perceiving their environment, acting to achieve specific goals, etc. As a result of increasingly rapid and burgeoning technological advancements, AI systems have been applied broadly throughout society including health, transportation, finance, retail, commerce, food, farming, manufacturing, public administration, personal assistants, etc. (News: European Parliament, 2020). However, one challenge that AI systems continue to face is their opacity, which hinders the ability to provide suitable explanations for how they arrived at their outputs, thus leading to what is referred to as "the black box problem". This opacity drives the need for explainable AI (xAI), a suite of AI systems and machine learning techniques, which, when combined with effective explanation techniques, can provide explanations to human users about the decisions, predictions, results, and outputs produced by AI systems. By providing insights into the inner workings of AI systems' decision-making processes, xAI can enable human users to understand, trust, and effectively manage these systems. (Gunning & Aha, 2019; Gunning et al., 2021).

The broad application of AI systems throughout society combined with the contextuality of explanations will influence and inform the requirements that xAI must fulfill. To be effective, xAI will therefore have to be adaptable and responsive to the distinct and the varied needs of a wide range of *explainees* — the stakeholders who interact with, are affected by, or have a legitimate interest in AI

systems and their outcomes. Stakeholders are entities (individuals or organizations) who may affect, be affected by, or perceive themselves to be affected by a decision, activity, or outcome of a product or system (INCOSE, 2023; Project Management Institute, 2021), and in the context of AI systems, they could include end-users; AI experts; engineers; lawmakers; government agencies and regulatory bodies; multiple players in industries that embed AI technology, such as suppliers, manufacturers, and operators; customer organizational decision-makers; and bystanders in environments where AI systems operate within society at large. Crucially, most of these stakeholders are not AI experts and therefore require non-technical, non-computational explanations expressed in human-centric layperson terms. By providing a description of the needs of various stakeholders, requirements are essential factors in defining the features, functions, and characteristics of the end-product and serve as inputs to and form a basis for the specification of the end-product (Project Management Institute, 2021).

Establishing such requirements for human-centric xAI, however, is not straightforward. Two significant gaps persist in current xAI development. Gap 1, Paucity of Human-Centric xAI Solutions Due to Limited Cross-Disciplinary Integration and Static Requirements Fallacy: Most xAI initiatives focus on computational approaches leading to a paucity of human-centric xAI solutions. While emerging software engineering trends have begun to treat explainability as a baseline requirement (Köhl et al., 2019; Habiba et al., 2022), current methodologies carry an inherent risk of treating explainability as a static, one-time engineering checkbox rather than an adaptive process of requirements management. This leads to xAI models that are built by AI developers for AI developers without fully and systematically leveraging the extensive body of knowledge from the Social Sciences and Humanities (SSH), specifically, the disciplines of cognitive science, psychology, and philosophy, about how humans participate in act of giving and receiving explanations (Miller et al., 2017; Miller, 2019). Gap 2, Insufficient Attention to Contextuality of Explanations: The complexity of the context-dependent nature of everyday explanations, where what counts as adequate depends on the *explainee*, the situation, and

their interplay, poses a challenge to xAI designers. These gaps, and the FGCR's approach to addressing them, are elaborated in the Problem Statement that follows (section 1.1.1).

1.1 Research Problem and Objectives

1.1.1 Problem Statement

To provide an adequate explanation for the outputs of AI systems, xAI needs to meet the explanatory needs of a broad range of stakeholders beyond AI domain experts. And because they are not AI domain experts, most stakeholders require explanations that are understandable by the layperson. Furthermore, xAI developers face several challenges in meeting the explanatory needs of most stakeholders and these challenges will result in different types of and multiple human-centric xAI solutions. The following combination of challenges drives the need for such solutions: contextuality of real-world explanations; explanations being sought for the outputs of different AI domains, applications, actions and models; and a paucity of human-centric xAI models. Therefore, when developing context-driven human-centric layperson termed xAI solutions, xAI developers need to take into consideration three intertwining issues:

1) Problem 1 (The Contextuality Challenge): Real-world explanations are context-dependent.

2) Problem 2 (The Technological Heterogeneity Challenge): AI systems are not uniform and homogenous. Even though people refer to "AI" as if it is one universal entity, it is not – AI systems are technologies that vary widely across domains, applications, functions, and models.

3) Problem 3 (The Computational Over-Emphasis Challenge): The current state of xAI over-emphasizes computational approaches that are mostly built for AI developers rather than end-users or a broad range of non-technical stakeholders.

1.1.1.1 Problem 1: Contextuality in Real-World Explanations

Intuitively, we know that explanations are highly contextual. Explanations depend on the contextuality of both the situation and the recipient of an explanation (*explainee*).

Hesslow's (1988) example about seeking an explanation for two different occurrences of a fire illustrates the situationally-driven contextuality of explanations where different situations call for different explanations. The need for different explanations stems from the context which is comprised of the combination of factors of the same *explainee* seeking explanatory information for the same phenomenon (fire) which occurs in different situations. Even though oxygen is a necessary and crucial condition for any fire, depending upon the context, it may or may not be cited as the answer to the 'why' question. The explanation for a house fire may cite a lighted match even though oxygen is a necessary condition. In contrast, the explanation for a fire in a normally oxygen-free container that holds a highly inflammable substance may cite the entry of oxygen into the container.

An example about seeking an explanation for a lung cancer diagnosis illustrates the *explainee*-driven contextuality of explanations where different *explainees* call for different explanations. The need for different explanations stems from the context which is comprised of the combination of factors of different *explainees* seeking explanatory information for the same phenomenon (cancer diagnosis) occurring in the same situation. For someone who is not a medical expert, a patient's history of smoking tobacco could be a sufficient explanation. However, for a medical expert, an adequate explanation might require information about a patient's blood test results that measure the levels of exposure to tobacco carcinogens such as NNK. NNK can bind to DNA and impact the modulation of critical cells that can transform normal cells into tumor cells which can ultimately result in uncontrolled cellular proliferation and tumorigenesis (Centers for Disease Control and Prevention (US), et al., 2010).

Furthermore, when we take a broader view on the subject of contextuality, one crucial consideration stands out: context matters, and context is a complex matter. Although the above examples neatly illustrate the two perspectives of an explanation's situation-driven and *explainee*-driven contextual components, note that they are providing an overly cursory and narrow view. When it comes to the context of explanations, there are multi-layered factors at play. Let's begin with the fact that everyday explanations are multi-predicated relational acts where someone *(explainer)* explains

(communicates or provides) something *(explanans)* to someone else and/or oneself (*explainee)* about a phenomenon to be explained *(explanandum)*. Now let's overlay contextuality into this act where we have to consider multiple layers of complexity: We not only have the context of the situation of the *explanandum* to consider but also the context of the situation of the *explainee*. When it comes to defining a situation, we would be looking at a situation's two sub-aspects of time and place. When it comes to considering the time aspect of the *explanandum*, we may not only be considering the time that the *explanandum* occurred, but also different times within the causal chain leading up to the *explanandum*. And, we have to consider the sub-aspects of time and place of the *explainee* and *explainer*. What most people do not realize is that when it comes to explanations, we are all juggling a multiplicity of factors with the result that multiple explanations (*explananda*) for an *explanandum* will be sought based on the interplay of all of the above.

An example illustrating the complex nature of contextuality within explanations is the multiple explanations for Malala Yousafzai's survival after being shot by the Taliban in 2012 (Yousafzai and Christina, 2013). Here, we have the scenario with the same phenomenon that needs an explanation (her survival), the same situation of the phenomenon (the time and place of her shooting), and the same explainer and *explainee* (Malala herself) in different situations (time and place at which Malala provides explanations to herself and others). In her accounts of explanation for her survival, the same person (Malala) holds or explores multiple explanations because each explanation serves a different function. When she recalls her doctors' explanations for her survival, she is looking at it from a medical and/or scientific explanatory perspective. They discussed how lucky she was to survive due to the facts that the bullet had entered through the side of her eye, traveled down her neck and lodged in her shoulder and how it had missed her brain and spinal cord, and the central blood vessels. Over time, when she speaks from a psychological explanatory perspective, she describes how the trauma shaped her identity and resilience and how she coped with fear and recovery through mental strength and support of her family and community. At other times, she speaks from a spiritual explanatory perspective when she said her

survival was due to God's will and power of global prayers. And at further points in time, she speaks from a social function perspective by interpreting her survival as impetus to advocate for girls' education and human rights. This example shows how one person can integrate and hold multiple explanations at different times for the same event each serving a different purpose including medical/scientific understanding; mental healing; meaning-making; and advocacy.

Another real-world example within the field of AI that illustrates the complex nature of contextuality within explanations involves Google DeepMind's AI computer program AlphaGo (Google DeepMind, 2024). The competition between AlphaGo and Lee Sedol, a 9-dan professional Go player (depicted in *AlphaGo: The Movie* (2020)) in 2016 serves as an example of multiple explanations being sought and given for the same phenomenon—AlphaGo's 4-1 victory over Lee Sedol. Here, we have the contextual configuration of the same phenomenon that needs an explanation (AlphaGo's victory); multiple and varying acts and events occurring at different times in the competition that combine to the end result of the phenomenon that needs an explanation; with different *explainers* and *explainees* participating in acts of explanation at different times. Some explanations from a technological innovations perspective involved DeepMind's use of multiple AI ML techniques and approaches including neural network, reinforcement learning, Monte Carlo tree search, value network, and policy network (Escape Velocity Labs, 2023; Google DeepMind, 2024). Some explanations from a strategic differences perspective involved AlphaGo's focus on probability of winning as opposed to the human approach of winning by as many points as possible (Swallow, 2023). Other explanations cited the innovative logic used by AlphaGo such as the example of AlphaGo's move 37 in game 2, which was described as a creative move (Google DeepMind, 2024). Other explanations involved the sheer computational power advantage held by AlphaGo that used ~1202 CPUS + ~176 GPUs + >100 scientists + ~1 megawatt of electrical power feeding the hardware against Lee Sedol's 1 human brain + 1 coffee + ~20 watts used by the human brain (Ceva, 2016). And to this day, the quest for more explanations continues – an internet search for explanations for this one phenomenon will yield many

more than those discussed above. This example demonstrates that multiple explanations will be sought and provided for the same phenomenon depending upon the complexity involved with multiple combinations of contextual perspectives (*explainer(s)*, *explainee(s),* time, place, etc.)

The complexities that I have discussed above may deter xAI developers from undertaking the task of providing human-centric explanations to a wider range of stakeholders. Nevertheless, it is crucial to keep in mind that humans are naturally inclined to explain and understand, and explanations are vital for building trust and fostering social connections. And building trust is crucial to gaining acceptance and increasing adoption of new technologies such as AI. Therefore, developing human-centric xAI models for diverse sets of stakeholders is worth undertaking. The key challenge lies in crafting explanations that are tailored to the needs of the *explainee(s)* and focusing on the fact that it is ultimately the *explainee(s)* who will determine the adequacy of an explanation. This underscores the necessity of diverse stakeholder consultation throughout the development of xAI models. Essentially, because explanations are not universal and context matters greatly, xAI solutions will need to be tailored for the context of not only the situation and *explainee(s)* but also their sub-aspects of time and place.

1.1.1.2 Problem 2: AI Systems are not a Uniform & Homogenous Technology

Machine learning (ML) is a branch of AI at the intersection of computer science and statistics that enables computers to identify patterns, learn, perform specific actions, make specific decisions, and provide accurate results. There are many different types of ML techniques used to create AI models. ML techniques vary from AI application to AI application and even within the same AI application depending upon the function undertaken. For example, facial recognition software might use neural net ML algorithms, whereas a Go game playing system might use a combination of techniques such as neural nets, reinforcement learning, decision trees, etc. The choice of model or combination of models created will depend upon the task undertaken (Alpaydin, 2014; Mohammed et al., 2017). This makes AI systems—the *explanandum and/or explananda*—varied and heterogeneous and therefore, xAI solutions

will need to be tailored to particular AI domains and applications and within that, particular AI models performing particular functions to take particular actions and/or make particular decisions.

1.1.1.3 Problem 3: Paucity of Human-Centric xAI Models

Currently, there is a scarcity of xAI development for a broad range of stakeholders who do not possess technical expertise in the field of AI. This emphasis on highly technical computational solutions is a result of the situation where most AI developers are building xAI for themselves rather than for end-users (Miller et al., 2017). xAI development efforts have been focusing on algorithmic functionality with a secondary consideration for the explanatory needs of diverse sets of end-users and stakeholders who require explanations in layperson terms.

This emphasis on and overabundance of highly technical and computational xAI solutions (e.g., decision trees, model distribution, rule extraction, etc.) is reflected in the findings of surveys of xAI initiatives such as Adadi and Berrada (2018) and Burkart and Huber (2021) and in the composition of DARPA's Explainable Artificial Intelligence Program (2019). In DARPA's program (2019), eleven out of twelve research teams tackled computational methods with a focus on developing new ML techniques to produce explainable models and designing effective explanation interfaces, while just one explored the topic of applying psychologic theories of explanation. Furthermore, these highly computational xAI solutions could leave human AI domain experts with the two-fold problem of explaining not only the outputs of technically complex and opaque AI systems but also translating the outputs of xAI models which themselves may be technically complex and incomprehensible to most stakeholders. Therefore, in order to mitigate the paucity of human-centric xAI models for a broad range of stakeholders, complex computational xAI solutions will need to be tailored to the diverse needs of non-technical *explainee(s)* in human-centric layperson terms.

This computational over-emphasis exposes a deeper methodology gap within the AI community: a fundamental misunderstanding of requirements management for sociotechnical systems. While xAI developers are increasingly turning to traditional Requirements Engineering (RE) tools to identify what

types of algorithmic solutions to build (e.g., Köhl et al., 2019; Habiba et al., 2022), these frameworks, while valuable, remain under-operationalized for adaptive deployment. They provide important conceptual scaffolding for identifying and classifying explainability requirements, but do not supply the structured tools, stakeholder consultation mechanisms, or iterative processes needed to manage, trace, and dynamically redeploy those requirements as user contexts evolve. Because human cognitive needs are volatile and situation-dependent, xAI demands a shift from static requirements specification to adaptive requirements management. Without this, even technically sophisticated xAI systems remain misaligned with the non-technical stakeholders they are meant to serve.

Because of the overly narrow emphasis on computational xAI solutions, xAI developers frequently overlook the human dimensions of explainability, resulting in a corresponding lack of human-centric xAI. This oversight occurs because they do not sufficiently leverage the knowledge on the concept of explanation found in the social sciences and humanities (SSHs). It is important to note that the term "social sciences and humanities" encompasses a wide range of academic disciplines, including, among others, cognitive science, psychology, philosophy, communications, media studies, sociology, political science, linguistics, and history. However, for the purposes of this thesis, the term SSH is used in a deliberately bounded sense, aligning with Miller's (2019) influential and foundational work. Therefore, throughout this thesis, the acronym SSH refers strictly to this integrated triad of cognitive science, psychological, and philosophical dimensions of human explanation, filtering out other SSH disciplines, which, while valuable in their own right, fall outside the scope of this research project. These three disciplines have studied the concept of explanation and how people engage in acts of explanation for decades, and in the case of philosophy, for millennia. Due to the immense breadth and depth of SSH-based knowledge on the concept of explanation, it is unreasonable and unrealistic to expect AI developers to be experts in two vastly different knowledge areas. AI and xAI researchers are already facing the demands of mastering and being experts in their own field, therefore, it is impractical to expect them to also be experts in the field of SSH-based research on the concept of explanation.

SSH knowledge can assist in the establishment of requirements that xAI models need to fulfill in order to provide adequate explanations in layperson terms for a broad range of stakeholders beyond AI experts. Miller's (2019) influential survey of the existing SSH research on the concept of explanation began to address this gap. This thesis builds on and augments Miller's (2019) survey by synthesizing his ideas and those of many other SSH researchers into taxonomies for use by xAI designers. These SSH-based taxonomies of explanation are among the resources within the FGCR that can serve as aids for facilitating the efficient identification of crucial requirements for xAI model development, making SSH-based knowledge on the concept of explanation accessible for xAI developers.

Determining an appropriate xAI solution for outputs of an AI system is dependent upon a combination of the various elements outlined above. To provide more human-centric explanations for a broad range of stakeholders, xAI developers will need to consider three interrelated factors: 1) the particular stakeholder seeking an explanation within a particular context; 2) the particular AI application, and within that, the particular AI model performing a specific action; and 3) the articulation of an explanation in human-centric layperson terms for most stakeholders. The complexity of balancing these challenges means there will rarely (if at all) be a one-size-fits-all universal xAI solution – solutions will need to be tailored not only to diverse AI applications, but also to the same models and actions as they appear across different scenarios, accounting for the contextuality of explanations and the varying needs of different *explainees.*

1.1.1.4 The FGCR as a Proposed Solution

Together, the three issues elaborated above drive the two research gaps identified at the outset. The overemphasis on computational xAI solutions built by and for AI developers reflects an insufficient leveraging of SSH-based knowledge (Gap 1 and Problem 3). Meanwhile, the insufficient attention to account for the complexity involved in the context-dependent nature of everyday explanations (Gap 2) reflects an inadequate consideration of the interplay of multiple factors including: the *explainee* and *explainer,* both of which are layered with the situation of the time and place (Problem 1); as well the

inherently heterogenous nature of the *explananda*—AI systems, which are also layered with time and place (Problem 2). Because these factors are so deeply variable, there can be no universal "one-size-fits-all" xAI solution. Consequently, addressing the two research gaps therefore requires simultaneously balancing the tensions between all three intertwining issues. This interdisciplinary doctoral project proposes a framework for generating context-dependent requirements (FGCR) to do exactly that.

The FGCR is composed of a process and toolkit to assist xAI developers with navigating and balancing these tensions in order to determine a suitable xAI solution for any given case. By integrating insights from SSHs, engineering, and xAI, the FGCR will serve as a practical guide for xAI developers by a) enabling the establishment of a set of requirements that define the needs of various and diverse sets of stakeholders; b) providing user-friendly resources, such as taxonomies, to facilitate the identification of crucial requirements; and c) enabling developers to efficiently determine a suitable xAI solution. Ultimately, the FGCR will enhance current xAI initiatives by enabling the production of more human-centric and contextual xAI models that will increase AI systems' explainability for a wider stakeholder base.

1.1.2 Research Objective

The objective of this research project is to utilize multi-disciplinary insights within the intersection of xAI, engineering, and the social sciences and humanities to produce a framework for generating context-dependent requirements (FGCR). The FGCR will aid xAI developers in identifying a suitable human-centric xAI solution by establishing the contextual requirements essential to determining what is to be built. This will inform the development of more human-centric xAI models and enhance current xAI initiatives.

The FGCR is comprised of a process and a suite of resources for use in the xAI development cycle. The FGCR is composed of a combination of new and existing knowledge:

(1) The FGCR includes a new process (Figure 10 in Chapter 4) that outlines iterative activities to progressively elaborate the needs of the *explainee*s based upon context.

(2) The FGCR also includes new aids in the form of taxonomies that can be used to classify and convey complex and inter-related existing knowledge of SSH-based concepts of explanation and the current xAI landscape. Note that for the purposes of this thesis, these taxonomies can be thought of as resources that are aids for carrying out a particular function, namely, efficiently identifying a set of detailed requirements for xAI solutions. The FGCR's taxonomies will facilitate the filtering of extensive knowledge on the topic of SSH-based explanation and the current xAI landscape. They are designed to help xAI developers pinpoint and focus on relevant subsets of information within the broader context. The content of the SSH-based and xAI explanatory classifications will be elucidated in hierarchical and faceted forms. The taxonomies will serve two purposes – in Chapter 3's literature review, they will be used to articulate existing knowledge and in Chapter 4's proposed FGCR, they will serve as one of the resources for use in FGCR process. The faceted taxonomies will be constituted of and alternately represent the same information that is contained in the hierarchical taxonomies. These taxonomies will aid xAI developers to identify, clarify, and select the specific explanatory information sought by different stakeholders. They will therefore serve as crucial tools for efficiently identifying the key requirements that xAI models need to fulfill in order to meet the needs of multiple stakeholders.

(3) The FGCR also utilizes existing knowledge within engineering and project management disciplines including a selection of various stakeholder requirements definition and analysis tools.

The components of the FGCR will guide the process for determining what an appropriate explanation is for an AI application and its varied stakeholders. When used by xAI developers, the output of the FGCR will be a set of requirements that serve as inputs to the specification used throughout the xAI development cycle. Note that the proposed framework provides methodologies and tools to identify what xAI model to develop and not detailed design methodologies that prescribe how to develop an xAI model!

1.1.3 Motivations

The driving forces behind this research project are the human need for explanations in general and in particular, from agential entities such as AI systems, and the multiplicity of explanatory needs that humans possess. There are many reasons for wanting explanations from AI systems including the impact that AI systems have on society as well as wider acceptance and increased adoption of AI systems by the populace at large. Also, key factors such as trust, fairness, and accountability among others are essential in meeting the varied and multiple explanatory needs of human beings.

1.1.3.1 Societal Impact, Wider Acceptance, and Increased Adoption of AI Systems

According to Watson and Floridi (2021), "Perhaps, the most popular reason to explain algorithms is their large and growing social impact." A few controversial examples of the usage of AI systems that result in failures include: Predatory lending practices that disproportionately target poor communities (Eubanks, 2018); biased evaluations of loan applications (Munkhdalai et al., 2019); predictions of criminal recidivism (Dressel & Farid, 2018); inaccurate facial recognition classification of black/brown faces (Buolamwini & Gebru, 2018); etc. Currently, opaque AI systems seem to result in the automation, amplification, and rapid proliferation of systemic injustices already present in society. If these negative social impacts of AI deployment throughout society are not addressed and corrected, they could hinder wider acceptance and adoption.

Wider acceptance and increased adoption of AI systems are other reasons for wanting explanations. At present, despite the diverse and extensive range of deployment of numerous AI applications throughout society, they are not widely accepted or broadly adopted. Note that technologies can be broadly applied to many domains throughout society but can lack across-the-board mass adoption by billions of people. Examples of across-the-board mass adoption technologies can include cell phones or gasoline powered non-automated vehicles. In comparison, technologies such as facial recognition, digital personal assistants, healthcare diagnostic tools, recidivism prediction technologies, automated vehicles (AVs), etc. can be considered as technologies that are broadly applied

to many domains throughout society but lack across-the-board mass adoption. Perhaps, more transparency through provision of explanations will enable wider acceptance and increased adoption of AI systems.

1.1.3.2 Multiple Explanatory Needs of Humans

Meeting the explanatory needs of human beings has the potential to significantly enhance the acceptance and adoption of these new AI systems, especially those that act as agential entities that perform human tasks. If one takes the view that an agent is simply an entity that acts and is skillfully engaging with the world, we can consider some AI systems as agential entities (Schlosser, M., 2019, Section 2.3-2.4). When AI systems take autonomous actions without a human in the loop, such as performing the motion control functions of an automated vehicle (AV) system, we are dealing with an agent that is acting in our world amongst us. And, when things go awry, such as when AVs have accidents, we need to obtain explanatory information about the decisions made and actions taken by these agential entities. Having agents that suffer from the "black box problem" and cannot provide adequate explanations is an impediment to our understanding and acceptance of their decisions and actions, and in particular, our trust in them.

Some of the factors that contribute to meeting and/or satisfying the multiple explanatory needs of human beings include trust, fairness, accountability, transparency, ethicality, informativeness, causality, adjustability, and generalizability (Burkart & Huber, 2021).

According to McLeod (2023), ". . . 'trust is generally a three-part relation: *A* trusts *B* to do *X'*. . . or '*A* trusts *B* with valued item *C'* . . . or '*A* trusts *B* in domain *D'* . . .". Trust can be thought of as an attitude that we have towards others. Trust allows us to depend on others, and it requires reliance upon the other party to be competent to do what is required to achieve specific goals. For AI systems performing human tasks, especially those acting as agential entities amongst us, trust is an essential factor in our acceptance of their outputs and actions. Adequate explanations for the

outputs of AI systems could lead to our developing greater trust in them and this greater trust could lead to our greater acceptance of their outputs.

Increased fairness, accountability, and transparency (FAT) and conformity to ethical standards could also lead to increased acceptance of AI systems' outputs. An AI system needs to be accountable for its actions, be transparent about and able to justify its outputs, decisions, and actions, and conform to fair and ethical standards.

For domain experts, an AI system needs to meet the informativeness, causality and adjustability criteria in order to improve its performance. Domain experts need information and insight into the causal mechanisms such as attribute importance, within the underlying input-output relationship in order to understand a model and adjust a model's parameters to improve its performance.

In order to ensure that an AI system can be used for making decisions and taking actions appropriately for the contexts of their operating environments, it needs to meet the generalizability criterion. An AI system should be able to perform well in the real-world by generalizing beyond the training environment. Humans need to have confidence that its future behavior with yet-to-be seen data in real-world settings is appropriate, and we need to know in which contexts it generalizes well.

1.2 Research Questions

Under the overarching theme of "What is an adequate xAI-generated explanation?", this project has pursued the objective of informing the development of innovative xAI models that meet the explanatory needs of a broad range of stakeholders in human-centric terms. Because xAI aims to provide explanations to people, this project drew on knowledge from the social sciences and humanities (SSH) about how humans explain our own actions and thoughts to each other and ourselves in order to apply this knowledge to the field of xAI development.

While much work has been undertaken to solve the AI "black box problem", currently, meeting the contextual and human-centric layperson explanatory needs of diverse and non-AI expert sets of

stakeholders remains a challenge. To achieve my research objective of addressing the gap in xAI development for a broad range of stakeholders beyond AI experts, I created a framework that aids in identifying a contextually fitting xAI solution. In order to create and test the usefulness and practicality of a framework for identifying an adequate xAI solution for both the situation and the *explainee*, I explored several research questions.

Beginning with the research question, RQ1:

RQ1) "What multi-disciplinary insights and knowledge from the fields of social sciences, humanities, engineering, project management, and xAI can improve xAI development efforts by assisting with the identification of suitable context-dependent, human-centric xAI solutions?"

RQ1 involved a literature review that was driven by several sub-questions. Because xAI aims to provide explanations to people, this project first focused on how humans explain our own actions and thoughts to each other and ourselves.

In order to apply knowledge from the social sciences and humanities (SSH) to the field of xAI development, I investigated the sub-question, RQ1a:

RQ1a) "What constitutes an adequate explanation for humans?"

Because this project seeks to enhance current xAI initiatives, an understanding of the AI and xAI landscape was required and, therefore, I investigated the sub-question, RQ1b:

RQ1b) "What are the current state-of-the-art xAI approaches?"

And because the proposed framework for identifying a suitable xAI solution will focus on generating context-dependent requirements for human-centric xAI model development, I investigated current requirements management practices through sub-question, RQ1c:

RQ1c) "What insights and practices from the field of engineering and project requirements management can inform future xAI development?"

The knowledge obtained by answering RQ1 and its sub-questions was used to produce a framework for identifying a suitable context-dependent human-centric xAI solution. The creation,

testing, and optimization the operational features of this framework were steered by the main research question of this project, RQ2:

RQ2) "What is a framework that can be employed for identifying and classifying explanatory information sought by diverse stakeholders and generating context-dependent requirements for AI systems' explainability?"

Through the investigations of RQ1 and its sub-questions, I used insights from the SSHs, engineering, and xAI fields to create a framework that could aid xAI developers with the identification of the contextual explanatory needs of *explainees* and translation of those needs into a set of requirements that human-centric xAI solutions need to meet.

To test the functionality of and optimize this proposed framework, I used the qualitative research design of inquiry involving case studies research with interviews as the primary data collection method. I used data obtained from interviews with different and varied stakeholders who were queried about the explanatory information they would seek regarding a scenario outlined in the case study, namely, a series of automated vehicle (AV) car crashes. This testing enabled an assessment of whether the FGCR can be employed to generate specific xAI model requirements that meet the needs of different sets of stakeholders who would seek different types of explanations for the scenario outlined in the case study thereby determining a suitable xAI solution.

1.3 Research Scope

Because the purpose of xAI is to provide explanations to people, the scope of this project has been driven by their needs and placing them at the center of developing human-centric xAI solutions. To meet the explanatory needs of diverse sets of stakeholders in human-centric layperson terms, xAI developers must address the challenge of answering the "Are we building the right thing for relevant stakeholders?" question as outlined in "Section 2.3.5.2 Stakeholder Needs and Requirements Definition Process" of INCOSE's *Systems Engineering Handbook* (2023). In other words, identifying what an

adequate xAI solution is for a particular and/or sets of *explainee(s)*. And the engineering practices of stakeholder requirements definition and analysis processes are crucial to the endeavor of identifying the explanatory needs of diverse sets of stakeholders. Therefore, in order to encourage *explainee*-centric xAI development, the scope of this research project is focused on providing xAI developers with a framework to conduct requirements definition and requirements validation activities throughout the development cycle.

1.3.1 In-Scope

The scope of this research project is a proposed framework, the FGCR, which includes a process and a suite of tools and techniques that are applicable to and tailorable for xAI development. The framework will serve as a guideline for generating requirements that will be inputs to the development cycle of xAI solutions. Requirements are essential factors in defining and clarifying the features, functions, and characteristics of the xAI model to be developed. Requirements provide a description of the needs of various stakeholders and serve as inputs to and form a basis for the specification of the end-product (Project Management Institute, 2021). Contents of the proposed framework borrow from and extend existing systems engineering, software engineering, and project management knowledge utilized by many different fields in the development of products or services. The framework includes the practices, guidelines, tools, and techniques that are part of requirements management knowledge areas outlined in International Council on Systems Engineering's (INCOSE) *Systems Engineering Handbook* (2023), Project Management Institute's (PMI) *Project Management Body of Knowledge Guide* (*PMBOK)* (2021), and the *Guide to Software Engineering Body of Knowledge (SWEBOK)* (Bourque & Fairley, 2014).

The proposed FGCR will contribute to and augment the existing *SWEBOK*'s (Bourque & Fairley, 2014) software requirements knowledge area as outlined in "Section 1: Software Requirements". Through the addition of the proposed FGCR process that details step-by-step tasks, inputs, and outputs plus various tools and techniques as well as aids such as the proposed taxonomies,

the FGCR will facilitate xAI developers to generate context-driven requirements that feed into the software requirements specification. The intention is to assist with, fine-tune and spotlight context-driven requirements definition for a broad range of stakeholders thereby enhancing xAI development.

1.3.2 Out of Scope

This research project does not examine or tackle the domain of how to design or develop an xAI model. The framework does not include detailed design processes, practices, guidelines, or instructions targeted at individual engineering disciplines. For example, the tasks and guidelines involved in software engineering's design, construction, testing, maintenance, configuration management, QA, etc. knowledge areas are not part of the scope of this research project.

The focus of this research project is not about how to develop the xAI model, but rather, about determining what xAI model to develop, specifically, identifying a contextually appropriate one.

Furthermore, it is important to note that this research project's investigation and design of the FGCR's scope excludes ethical considerations and broader multidisciplinary domains beyond the disciplines of cognitive science, psychology, and philosophy. Therefore, note that the application of the FGCR is value-neutral regarding the underlying ethics, behaviours, and outputs of AI systems themselves. It is not intended to, nor does it evaluate, endorse, or critique those behaviours on ethical grounds, rather, it is strictly designed to systematically elicit the requirements of what constitutes a contextually adequate explanation that is transparent and human-centric for the stakeholders affected by the outputs of AI systems. Nevertheless, while the framework itself is evaluatively neutral, the outputs it produces are not isolated from ethical AI practice. By systematically capturing and uncovering the transparent, human-centric explanatory needs of affected stakeholders, the requirements generated by executing the FGCR's process can benefit and inform ethical dimensions of design choices made by AI and xAI developers. In the same vein, in defining the theoretical boundaries of this work, the multidisciplinary scope of the Social Sciences and Humanities (SSH) is explicitly constrained. This research does not evaluate broader sociological, political, or historical dimensions of AI deployment;

instead, exclusions extend to all non-engineering fields outside of the three specified disciplines of cognitive science, psychology, and philosophy.

1.4 Thesis Structure

This document is organized in five sections: Chapter 1 above outlined the introductory information including the problem statement; objectives; motivations; the research questions; and scope of this research project. Chapter 2 will outline the research design, qualitative research methodology and process taken by this project. Chapter 3 will present the results of the collection and analysis of relevant multi-disciplinary literature from the social sciences, humanities, project management, engineering and xAI that inform the development and serve as the foundation of the proposed theory of this research project. Chapter 4 will outline the proposed theory that this research project is investigating – the framework for generating context-dependent requirements (FGCR) for human-centric xAI model development. Chapter 5 will present the results of the proof-of-concept evaluation of the proposed theory, the FGCR. And Chapter 6 will summarize the thesis, including a discussion on contributions to xAI research initiatives as well as the impact of the findings on future research opportunities.

2 Research Design and Methodology

Based on Maxwell (2022)'s interactive model for qualitative research design, I undertook an interactive and exploratory qualitative approach for this thesis that suits the pragmatic nature of the qualitative research undertaken. The qualitative research design approach employed systematic literature synthesis for theory development followed by qualitative evaluation through proof-of-concept testing.

Maxwell's (2022) interactive model has five key components that are conceptualized as an interacting system and include: goals; conceptual framework; research questions; methods; and validity. The component of goals focuses on what the study seeks to accomplish and why it is worth doing. The conceptual framework component seeks to identify the theories, beliefs, prior research findings, literature, and/or personal experiences that serve as sources and guide the research. The research questions component identifies the specific questions that the research study will attempt to answer. The methods component identifies what actually will be done in conducting the study. The validity component will involve evaluation of the results of the analysis of the data collected and conclusions drawn from these. The interactive model has an interconnected and flexible structure that eschews other fixed and/or sequential design structures used by other types of qualitative research designs. It is an underlying scheme that governs

> the activities of collecting and analyzing data, developing and modifying theory in response to the data, elaborating or refocusing the research questions, identifying and addressing validity threats, and even rethinking the goals of the study in light of these may all be going on more or less concurrently, each potentially influencing all of the others. This process isn't adequately represented by a choice from a prior menu, or by a linear model, even one that allows multiple cycles (Maxwell, 2022).

2.1 Research Design

This research project took an interactive qualitative research approach and comprised three phases. The first phase involved the collection and analysis of relevant multi-disciplinary literature from the social sciences, humanities, engineering, project management, and xAI. Phase 1 tackled RQ1 and its sub-questions. This involved a structured review of the relevant literature to identify the theoretical gaps and inconsistences. The second phase iteratively synthesized the ideas from the literature analyzed and produced the proposed theory – a literature-derived conceptual framework for generating context-dependent requirements (FGCR) for human-centric xAI model development. Phase 2 tackled the main research question RQ2. And phase 3 involved proof-of-concept testing of the proposed theory, the output of phase 2's RQ2, the FGCR.

2.2 Research Methodology & Research Process

2.2.1 Phase 1: Collection and Analysis of Multi-Disciplinary Literature

Research Question (RQ1), "What multi-disciplinary insights and knowledge from the fields of social sciences, humanities, engineering, project management, and xAI can improve xAI development efforts by assisting with the identification of suitable context-dependent, human-centric xAI solutions?" and its sub-questions were explored through the collection and analysis of relevant multi-disciplinary literature from the social sciences, humanities, engineering, project management and xAI.

2.2.1.1 Search Strategy

To address RQ1 and its sub-questions, I employed a multi-disciplinary search strategy that involved exploring literature from the social sciences, humanities, engineering, project management, and xAI. In October 2022 I set out to identify pertinent and relevant literature with respect to RQ1's varied multi-disciplinary topics and knowledge areas. I initially designed search expressions around key

concepts pertaining to social sciences, humanities, explanation, and xAI and entered them into two source databases (University of Ottawa's OMNI and Google Scholar).

I subsequently used a combination of two search strategies to delve deeper into the topics pertaining to RQ1's sub-questions: pearl growing (Papaioannou et al., 2010) and snowballing (Wohlin, 2014). Pearl growing is beneficial for multi-disciplinary searches due to the dispersion of literature across a range of subject areas as well as the variability of and difficulty in defining concepts, terminology, and search terms. Pearl growing involves taking an exemplar or 'pearl' article to extract index terms and keywords in an iterative process of searching until the material found becomes less relevant. Snowballing is the technique of using both backward and forward searching strategies applied to an exemplar paper to identify additional papers and literature for investigation. Backward snowballing involves looking for references or cited articles by the exemplar paper. And forward snowballing involves looking for citing articles and literature that were published after and influenced by the exemplar paper.

2.2.1.2 Search Query

As stated above in 2.2.1.1, during the preliminary search phase, I iteratively conducted simultaneous searches that combined the multiple concepts pertaining to the social sciences, humanities, explanation, and xAI. The initial search queries were entered into the following two source databases: University of Ottawa (uOttawa) library's unified search engine (OMNI) and Google Scholar. The purpose of the preliminary search was to identify key journals and databases on the topic and to locate exemplar articles that were then used as the nucleus of pearl growing and snowballing strategies.

The initial search query for RQ1a targeting literature at the intersection of the topics of explanation in the social sciences and xAI was composed of the following phrases:

("Cognitive Psychology") AND (Explanation) AND (xAI OR "Explainable AI")

The initial search query for RQ1b targeting literature on the topic of xAI was composed of the following phrases:

(xAI OR "explainab* AI" OR "explainab* Artificial Intelligence") AND (survey)

I limited the search to metadata fields such as the title, keywords and abstract. The inclusion criteria for this search were limited to publication dates spanning 2017-2022; English publications only; and included all types of peer reviewed and grey literature papers and publications including conference, journals, workshops, books, book sections and thesis.

The above-mentioned search queries yielded the pearl documents of Adadi and Berrada's (2018) "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", and Miller's (2019) "Explanation in artificial intelligence: Insights from the social sciences".

In addition to snowballing, I used these two pearl documents listed above to extract index terms and keywords and conducted subsequent search queries in the following source databases:

1. Scopus – a multidisciplinary database covering research in areas such as science, technology, medicine, social sciences, arts, and humanities.
2. Web of Science – a database that includes conference papers and articles from key research journals in science, engineering, medicine, social sciences, and humanities.
3. Science Direct – a database that includes full-text-peer-reviewed journal articles and book sections from the core scientific literature.
4. IEEE Explore – a database that includes technical literature in electrical engineering, computer science and electronics. Full text published since 1988, and partial content published since 1913 from IEEE journals, standards, transactions, and conference proceedings.

2.2.1.3 Database Search Query Results

The initial searches for both search query phrases above yielded a combined total of 14,704 results. The screening process to identify the above-mentioned exemplar starter articles involved a review of the title, abstract and keywords and involved the research goals and questions, namely, did the publication provide introductory information and survey information on the topics of explanation and xAI.

This preliminary search approach yielded the above-mentioned two starter articles that were then used as the nuclei of both the pearl growing and snowballing strategies using the databases listed above. For RQ1a, Miller's (2019) in-depth survey of literature on the topic of explanation within various sub-

fields of the social sciences served as the exemplar "pearl" article. And for RQ1b, Adadi and Berrada's (2018) survey on xAI served as the exemplar "pearl" article.

To bridge the three-year gap between the initial literature conducted in 2022 and the current submission of this thesis in 2026, I re-executed the search protocols. This enabled me to ensure the continued relevance of the findings of the 2022 literature review. The results of the follow-up search demonstrated consistency with the original search because there were no significant discrepancies or fundamental shifts between the results of the two searches. Although a small number of contemporary articles were identified in the recent search, they served to corroborate rather than challenge, contradict or supersede the findings of the original search. A discussion of selected contemporary literature and its relevance to the present research project is provided in section 5.4. This consistency and corroboration justify the ongoing reliance of the results of the 2022 search. Consequently, the literature identified in the 2022 search remains a cornerstone of this research project, informing the core of the foundational concepts of the theory that I have proposed.

2.2.1.4 Collection and Analysis of Literature for RQ1a

I conducted an exploration of **RQ1a, "What constitutes an adequate explanation from humans?" and/or "How do we (humans) explain our own actions and thoughts to each other and to ourselves?"** by identifying relevant literature from the social sciences and humanities on the topic of explanation. To do so, as stated above, I used Miller (2019) as an exemplar article for both pearl growing and snowballing purposes. And I conducted subsequent search queries using the source databases identified in section 2.2.1.2.

The output of RQ1a, is a synthesis of SSH knowledge of the concept of explanation (section 3.1). The prevailing characterizations of the concept of explanation is explicated in the form of textual descriptions as well figures of SSH-based hierarchical and faceted taxonomies of human explanation classes. Note that these taxonomies are part of the FGCR's set of tools and aids (Chapter 4).

2.2.1.5 Collection and Analysis of Literature for RQ1b

I conducted an exploration of **RQ1b: "What are the current state-of-the-art xAI approaches?"** by identifying relevant literature from the fields of engineering, AI, and xAI.

Again, to do so, as stated above, I used Adadi and Berrada (2018) as an exemplar article for both pearl growing and snowballing purposes.

For section 3.2's Artificial Intelligence and machine learning concepts and definitions literature review, I used a search query that included the following phrases and synonyms for:

(AI OR artificial intelligence) AND "introduction to" AND ("machine learning") AND ("for beginners")

To identify the relevant literature for the exploration of RQ1b, I conducted subsequent search queries using the source databases identified in section 2.2.1.2.

The outputs of RQ1b include a synthesis of AI and machine learning concepts and definitions (section 3.2) and a synthesis of the current xAI landscape (section 3.3). This is explicated in the form of textual descriptions of the various xAI approaches as well as a figure of a faceted taxonomy of the landscape of xAI approaches. Note that this taxonomy is also a part of the FGCR's set of tools and aids (Chapter 4).

2.2.1.6 Collection and Analysis of Literature for RQ1c

I conducted an exploration of **RQ1c: "What insights and practices from the field of engineering and project requirements management can inform future xAI development?"** by identifying relevant literature from the fields of engineering and project management. For RQ1c, I used codified engineering standards pertaining to product development requirements management practices: International Council on Systems Engineering's (INCOSE) *Systems Engineering Handbook* (2023), Project Management Institute's (PMI) *Project Management Body of Knowledge Guide* (*PMBOK)* (2021), and the *Guide to Software Engineering Body of Knowledge (SWEBOK)* (Bourque & Fairley, 2014).

The output of RQ1c is a synthesis of product development requirements management knowledge from the fields of engineering and project management in the form of a set of processes, tools and techniques.

The findings from RQ1 and its sub-questions drove the exploration of RQ2.

2.2.2 Phase 2: Development of Theory – The Framework

The findings from phase 1 served as crucial inputs to the activities involved in answering the main research question of **RQ2, "What is a framework that can be employed for identifying and classifying explanatory information sought by diverse stakeholders and generating context-dependent requirements for AI systems' explainability?"** The activities included the usage of the outputs of phase 1 in the development of a proposed framework for generating context-dependent requirements (FGCR) for human-centric xAI model development.

The FGCR comprises a process and a suite of resources that combines new and existing knowledge to guide xAI development. The cornerstone of the FGCR relies on crucial new knowledge, which includes a process (see Figure 10 in Chapter 4) and essential taxonomies that provide valuable aid in guiding xAI development.

The FGCR's process is a synthesis of the knowledge that resulted from RQ1c's insights and practices from the field of engineering and project requirements management. I extracted descriptive requirements management knowledge from International Council on Systems Engineering's (INCOSE) *Systems Engineering Handbook* (2023), Project Management Institute's (PMI) *Project Management Body of Knowledge Guide* (*PMBOK)* (2021), and the *Guide to Software Engineering Body of Knowledge (SWEBOK)* (Bourque & Fairley, 2014) and synthesized it into a process map. The process map outlines iterative activities to progressively elaborate the needs of the *explainee*s based upon context and to capture them into a set of requirements that feed into the xAI development cycle.

The taxonomies serve as resources that can be used to facilitate the process of filtering and classifying of extensive knowledge on the complex topics of SSH-based concepts of explanation and the

current xAI landscape (outputs of RQ1a and RQ1b). The taxonomies will help xAI designers to identify and pinpoint relevant information about the explanatory needs of stakeholders thereby enabling the tasks of efficiently identifying a set of detailed requirements for xAI solutions.

The FGCR also includes existing knowledge from the engineering and project management disciplines on the topic of stakeholder requirements definition and analysis tools (outputs of RQ1c) and incorporates them as a list of potential tools that xAI developers can use throughout the process of identifying key requirements.

All of these parts of the FGCR are intended to enable xAI engineers to generate requirements for developing robust and effective xAI models. The output of RQ2 is the FGCR (Chapter 4), the proposed theory.

2.2.3 Phase 3: Testing the Theory

Phase 3 involved limited testing of the output of phase 2, the proposed theory, the FGCR to demonstrate proof-of-concept. According to INCOSE (2023), proof-of-concept is the ". . . realization of an idea or technology to demonstrate its feasibility". The objective of proof-of-concept testing for a theory such as a framework is to assess its feasibility and functionality before proceeding to its full development.

During testing, I evaluated the functionality of the FGCR's process workflow and components and the usefulness of its tools and aids, especially the new aids in the form of taxonomies. Specifically, the primary aim was to assess whether it enables the generation of detailed xAI model requirements that reflect the needs of various and diverse sets of stakeholders thereby facilitating the determination of an appropriate xAI solution.

To test the FGCR's functionality, I followed two sequential stages: In stage 1, I employed the qualitative research design of inquiry involving case studies to gather real world data about the explanatory needs of stakeholders with respect to the actions taken and/or decisions made by an AI system. I used participatory and empirical research methods common to the social sciences such as

interviews as the primary data collection method for case studies research. In stage 2, I used the data obtained from the interviews from stage 1 to assess whether the FGCR's components can be applied to real data to generate specific xAI model requirements for different stakeholders who would seek different types of explanations for the case study scenario.

This enabled me to identify the flaws in, finetune, optimize, and enhance the FGCR's design features. Note that the scope of testing did not include comprehensive acceptance testing in a full field trial with the participation of xAI developers to assess if the framework is truly useful and practical in real-world development contexts. Such comprehensive validation would be the purview of future research opportunities to determine if it can be used by xAI developers in the field to produce usable human-centric xAI models.

2.2.3.1 Stage 1 of Testing the FGCR: Data Collection Using Case Studies Research

During stage 1 of testing the FGCR, to elicit the explanatory needs of stakeholders pertaining to the actions taken and/or decisions made by an AI system, I used interviews as the primary data collection method for case studies research.

Case studies enable researchers to collect and document detailed information about the experiences of participants and answer pertinent and open-ended 'how' and 'why' type questions about a case in a real-life setting or involving a real-life case. They can focus on concrete entities such as individuals or groups or less concrete entities such as projects or decision processes (Creswell & Creswell, 2017). The proposed theory that this research project is investigating is a framework composed of a process and a suite of tools and aids.

After examining multiple AI application domains—ranging from movie recommender systems, game-playing models like AlphaGo, to automated vehicles—I deliberately narrowed my research focus to a single case study scenario: autonomous driving systems. The decision to concentrate on autonomous vehicle (AV) technologies was driven by several critical considerations. The AI application domain of AVs have technological complexity with real-world high stakes outcomes that hold broad

societal relevance due to the fact that vehicles are ubiquitous and impact a broad spectrum of stakeholders. The technological complexity of AVs is due to the fact that autonomous driving represents an intricate AI challenge requiring sophisticated integration of multiple AI models performing multiple tasks. For example, AI models using ML techniques such as neural networks, support vector machine (SVM), Bayesian networks, reinforcement learning (RL), etc. tackle the tasks of motion planning, vehicle localization, motion control, perception, detection of pedestrians, traffic signals, road markings, automated parking, etc. (Bachute & Subhedar, 2021). The AI domain of AVs have real-world high-stakes outcomes because they directly impact human safety, especially when considering the real-world cases of AVs involved in a series of car crashes. With increased adoption, AVs will hold broad societal relevance due to the fact that vehicles are universally encountered across society. Furthermore, the impact of AV technologies will extend to a broad spectrum of stakeholders throughout society, cutting across all professions and walks of life, as vehicles are a ubiquitous part of daily life. All of the above considerations provide a rationale for conducting focused research with a manageable and narrowly defined scope, involving a single case study, which has the potential to yield meaningful insights and make significant research contributions to the field of xAI.

2.2.3.1.1 Case Study Scenario: AV Systems Crashing into First Responder Vehicles

In order to test the proposed FGCR, I used its components including a process and a set of tools and aids to identify the varied explanatory information sought out by different stakeholders who are impacted by a real-world AI application in the case study.

For this thesis, the interviews involved a case study to investigate the AI application domain of autonomous vehicles (AVs). The case study examined a real-world scenario of AVs involved in a series of car crashes. Interview techniques were used to query different stakeholders in order to collect data about specific explanatory information that they would seek from AI systems regarding certain algorithmic decisions, actions, and outputs.

The case study scenario I chose was motivated by a confluence of critical factors: the burgeoning societal interest in the evolution of AV technology; extensive media coverage of a USA National Highway Traffic Safety Administration (NHTSA) investigation into a series of AV car crashes (Gitlin, 2022; Moreschi et al., 2023; Stoklosa, 2022); and the urgent need to address the profound societal impacts results from these high-stakes safety incidents. This scenario offered a unique opportunity to examine official government data and documentation (NHTSA investigation EA22-002) for complex AI outcomes such as AVs involved in high-stakes outcomes such as car crashes. Specifically, this study focused on a pattern of real-world car crashes involving Tesla vehicles equipped with the advanced driver assistance system (ADAS) called "Autopilot" which controls the steering, braking and acceleration functions of the AV without any human intervention and could, at any time, disengage and return control to the human driver.

My exploration focused on a specific pattern of accidents that triggered an official investigation by USA NHTSA: the occurrences of Tesla's Autopilot system deactivating less than a second before colliding with emergency vehicles (Gitlin, 2022; Moreschi et al., 2023; Stoklosa, 2022; NHTSA, 2022). In August 2021, USA NHTSA's Office of Defects Investigation (ODI) opened an investigation into the accumulation of car crash incidents involving Tesla AVs. According to NHTSA's investigation, EA22-002, between January 2018 and January 2022, Tesla's AVs, with Autopilot engaged, were involved in 16 crashes where they struck highly visible stationary in-road or roadside first responder vehicles (e.g., police, ambulance, fire trucks, road maintenance vehicles) that were attending to pre-existing collision scenes. And, according to NHTSA, ". . . on average, in these crashes, Autopilot aborted vehicle control less than one second prior to the first impact" (2022).

I identified and queried different stakeholders impacted by AV AI systems about the explanatory information they would seek about the decisions made (or not) and actions taken (or not) by Tesla's Autopilot. The recruitment efforts including the identification of types of stakeholders for inclusion in this study is detailed below in section 2.2.3.1.2. Note that the stakeholders chosen for participation in

this study did not include anyone involved in these 16 crashes, but rather, a representative sample of the general population that comes in contact with and/or is impacted by automobiles in general. The types of stakeholders targeted for inclusion in this study were intended to cover a broad spectrum of society including pedestrians, drivers and/or owners of AV vehicles, drivers of other non-AV vehicles on the road, police/first responders, government representatives, lawyers, judges, policy makers, insurance industry members, AI developers, etc. The intent was to elicit different stakeholders' opinions, thoughts and questions about the decisions, actions and outputs of this new AI technology that operates in our world as an agent and is anticipated to replace the widely adopted existing technology of human-operated automobiles.

The testing of the FGCR enabled an assessment of whether it generates specific requirements for different stakeholders who would seek different types of explanations for the scenario outlined in the case study. Because I was exploring the contextuality of explanations, I anticipated that different stakeholders would ask for varied explanatory information about the different functions and tasks performed by the AI system in one specific scenario chosen for the case study, an expectation borne out by the interview data. The testing enabled the identification of requirements in terms of SSH-based categories of explanation with a potential list of either existing xAI models or requirements to drive development of new xAI models that can fulfill those requirements. The experimental data from the testing aided with the definition and optimization of the design of the framework.

2.2.3.1.2 Stage 1's Data Collection and Analysis Processes

Because ethics review is required for any research involving human participants, prior to commencement of data collection, this project received ethics approval from the University of Ottawa's Office of Research Ethics and Integrity (File# H-01-25-10973) – Appendix A contains a copy of the approval letter. This project conducted the following data collection and analysis activities per the guidelines and requirements of the University of Ottawa's Research Ethics Board (REB):

- recruitment of participants

- details of procedures and protocols for gathering of participation data, including audio and video recordings of interviews
- obtaining free and informed consent from participants
- maintaining confidentiality and safeguarding of entrusted information

The data collection process involved two major activities: the recruitment efforts and the actual data collection involving interviews between the interviewer, me (Unnati Patel, the principal investigator (PI)) and the interviewee (participant). These qualitative interviews occurred either in-person or remotely. These interviews took the form of face-to-face, telephone, or video-calls depending upon the availability and preferences of particular interviewees.

Recruitment

The recruitment efforts involved purposeful sampling strategies to recruit participants. I used two specific purposeful sampling strategies to recruit participants: snowball sampling and convenience sampling. Snowball sampling involves asking informants or well-connected individuals (e.g. Transport Canada officials, first responders, police, lawyers, judges, teachers, researchers, etc.) to distribute the recruitment flyers to others with similar characteristics or relevant knowledge, who can then contact the researchers directly should they be interested in participating. Note that in order to make the initial contact to well-connected individuals, I used either cold-calling techniques by using publicly available contact information from organizations such as Transport Canada, associations for retired police officers and Member of Parliament of Canada to identify and reach out to them or reaching out to individuals that I knew in my personal network. Convenience sampling involves gathering information from participants who are easily accessible to the researcher such as pedestrians or car drivers.

The types of stakeholders targeted for inclusion in this study was intended to cover a broad spectrum of society. The inclusion criteria for recruitment included the following: Participants must be adult 18 years of age and older and English speakers with ability to speak, understand, and write fluently in English. Participants who represent a diverse range of stakeholders across different sectors of society who are affected by the deployment of AI systems such as automated vehicles (AVs) in our

communities. The different stakeholder roles that people hold in society could include any of the stakeholder categories in the list below.

Category of stakeholder:
a) Pedestrian: _____
b) Driver of AV vehicle
c) Driver of other non-AV vehicle on the road
d) Insurance Industry – specify your role _____
e) Police or First responder – specify your role _____
f) Government (e.g., Transport Canada, Government of Ontario Ministry of Transportation, etc.…)
g) Policy Maker – specify your role and organization _____
h) Crash investigator – specify your role and organization _____
i) Automotive Industry – specify your role and organization _____
j) Engineer -- specify your role and organization _____
k) AI developer – specify your role and organization _____
l) Researcher -- specify your role and organization _____
m) Other – specify: _____

I created a recruitment flyer and posted it on my LinkedIn page. I also distributed it to highly connected personal contacts. Both sampling strategies utilized a screening questionnaire to identify the category of stakeholder role in society that each participant holds as listed above. Inclusion of participants selected for the study were on a first come, first serve basis.

The recruitment target was ~20-35 participants depending on success in recruiting ~3-5 participants from as many of the categories of stakeholders identified above. To maximize diverse representation while reducing bias, this project required a balance in the number of participants in each category. For example, if I had been able to recruit 5 participants in the "pedestrians" category and only one in the "police" category, then the data would not be balanced. In the final tally, a total number of 26 participants were ultimately recruited for the project and within that total, I was able to recruit ~3-5 participants for most of the categories above with a minimum of 3 in each category. And within that total, I was able to recruit the following number of participants outlined per category:

- 5 engineers (non-AI expert)
- 3 AI developers
- 3 AV drivers
- 4 non-AV drivers
- 4 police officers
- 4 legal experts including 2 lawyers and 2 judges

- 3 others – 1 non-transport policy maker; 1 pedestrian and/or passenger; 1 insurance industry.

Note that despite cold-calling and networking efforts, unfortunately, I was unable to recruit any lawmakers, government transport agency policy makers, crash investigators or automotive industry professionals. See Appendix A for copies of all recruitment material including the recruitment flyer and recruitment letters.

Data Collection Protocol:

I executed the data collection process in accordance with the protocol detailed in Appendix A including the following tasks: recruiting and categorizing interview subjects; conducting interviews; and providing a transcript of the interview within 48 hours and requesting approval from the interview subject.

Questionnaire

The full questionnaire that I used to conduct the interviews is in Appendix A. The questionnaire for these interviews was comprised of generally open-ended questions that are few in number and intended to elicit views and opinions from the participants. Open-ended questions use verbs and prompts such as "describe", "clarify", "detail", "tell me more", "I need more detail", "could you clarify x", "could you explain your response more?", "what does x mean?", etc.

The following text of the scenario was recited verbatim to the participants and then the questions were posed to them:

> This case study scenario involves a real-life case within the AI application of automated vehicles (AVs). It involves the occurrences of actual car crashes involving one particular AV brand, Tesla and its advanced driver assistance system (ADAS), "Autopilot". Tesla's Autopilot system controls the steering, braking and acceleration functions of the AV without any assistance from the human driver. Furthermore, note that Autopilot could, at any time, disengage and hand over controls to the human driver.
>
> According to USA NHTSA's Office of Defects Investigation (ODI), EA22-002, between January 2018 and January 2022, Tesla's AVs, with Autopilot engaged, were involved in 16 crashes where they struck highly visible stationary in-road or roadside first responder vehicles (e.g., police, ambulance, fire trucks, road maintenance vehicles) that were attending to pre-existing collision scenes. Furthermore, ". . . on average, in these crashes, Autopilot aborted vehicle control less than one second prior to the first impact".

The main question that guided the investigation was:

> Based on the scenario described above, you are seeking explanatory information about these car crashes from "Autopilot", the AI system that controls the steering, braking and acceleration functions of the AV (i.e., motion control functions).
>
> What specific questions would you ask of and/or what types of information would you seek from the AI system about the decisions made and actions taken by Autopilot (an agential entity) that is performing the human driving tasks of motion control of the AVs? i.e., controlling the steering, braking and acceleration functions of the AV.

Furthermore, note that the full questionnaire includes text that must be read verbatim to help put interviewees at ease in order to engage with them in a dialogue that elicits information while assisting me with taking on a neutral observational role. The text included a predetermined set of nudging prompts comprised of carefully worded phrases formulated with open-ended verbs. These nudging prompts aimed at minimizing the amount of possible researcher bias while ensuring the interviewee was focused on the topic.

I recorded every interview by audio or video. The transcripts were then generated automatically by the video conferencing applications MS Teams and Zoom. And in the case of audio recordings, the transcripts were generated by using the MS Word transcribe function. The output of all of these sources was transcribed text that became the final data set as described in further detail in the next section.

Data Analysis Process:

The first step in the data analysis process involved editing of the automatically generated transcripts to ensure accuracy of the content and confidentiality of the participants. To ensure confidentiality I deleted any and all comments and personal opinions expressed by the interviewee and/or the interviewer that were not within the scope of the research, namely, the topic of explanation. Once the pared down transcript was approved by the interviewee, I parsed the transcript further by a) removing all references to real people and places and then b) extracting only a subsection of that as it pertains to the topic of explanatory information that each stakeholder would be seeking. This final pared

down transcript underwent full data analysis to test out the FGCR and is the source of quotes in this thesis to illustrate key concepts.

Manual Analysis of Transcribed Text

I manually analyzed each stakeholder's transcript by first identifying specific quotes pertaining to explanatory information sought out by them. I focused on phrases and sentences that contained amongst others, "why" and "how" type questions as well as "I want to know" types of statements. To ensure rigor, consistency and accuracy in my data analysis, I developed and applied the individual stakeholder data analysis template with drop-down menus for categories (presented in Chapter 5, section 5.1.1.3.2). Note that the results of the analyses of the individual stakeholders also enabled me to identify and extract aggregate data patterns for further analysis. Sections 5.2-5.3 contain the results of the analysis of the gathered data from the interviews.

Augmenting Manual Analysis of Transcribed Text with Use of Software

To ensure rigor and reliability of the manual qualitative data analysis, I chose to adopt a strategy of computational augmentation. Upon evaluation of traditional qualitative data analysis software tools such as NVivo alongside the capabilities of Large Language Models (LLMs), I opted to use a readily available LLM: Google's Gemini (specifically, **Gemini 2.5 Pro Preview 05-06 gemini-2.5-pro-preview-05-06)**. I selected an LLM based on its potential to process large volumes of qualitative textual data with ease; discern the context and subtleties of language; identify complex conceptual relationships; and automate the coding and categorization process thereby allowing me to focus on higher-level analysis and interpretation.

A comprehensive account of the selection criteria, the specific materials processed and the iterative prompt engineering protocols utilized is provided in Appendix C. While the initial objective was to leverage the LLM as a supplementary analyst which could perform cross-validation of my manual analysis and improve efficiency and accuracy, the practical application revealed significant

limitations. The LLM produced spotty results and had limitations with respect to its precision, and accuracy for repetitive, high-stakes analytical tasks.

Consequently, I considered it a collaborative effort between myself and another analyst, albeit an unreliable one, but one that produced some results. In the end, I decided to harness the benefits of human-computer collaboration by using the LLM as a cognitive assistant to support my qualitative data analysis. This approach enhanced the rigor and reliability of my analysis by providing fresh perspectives, offering alternative phrasings that supplemented my interpretation. The LLM did not reveal any significant new patterns beyond those I had already identified, which could reflect a corroboration of the current state of my taxonomy of SSH-based explanation, which, as stated in section 3.1 below, remains a working foundation subject to further refinement.

2.3.2 Stage 2 of Proof-of-Concept Testing of the FGCR: Application of Real-World Data to Generate xAI Model Requirements

During stage 2 of the FGCR's proof-of-concept testing, I used the data obtained from stage 1's interviews and applied it to the FGCR's components to assess whether it can generate specific xAI model requirements for different stakeholders who would seek different types of explanations for the case study scenario. This enabled me to identify the flaws in, finetune, optimize, and enhance the FGCR's design features. The results of this proof-of-concept testing are presented in sections 5.2-5.4.

3 Phase 1 – Literature Review

Phase 1 of this research project involved the collection and analysis of relevant multi-disciplinary literature from the social sciences, humanities, engineering, project management, and xAI. This ultimately informed the development of the framework for generating context-dependent requirements (FGCR) for human-centric xAI model development. The structured review of the relevant literature enabled me to identify the theoretical gaps and inconsistences and synthesize the ideas from the literature analyzed in order to produce the proposed theory. The insights within the intersection of these multiple disciplines were used to create and are integral components of the FGCR.

In this chapter, I will present a synthesis of Social Sciences and Humanities (SSH) knowledge of the concept of explanation in section 3.1; a synthesis of Artificial Intelligence (AI) and machine learning concepts and definitions in section 3.2; a synthesis of the current Explainable AI (xAI) landscape in section 3.3; and a synthesis of product development requirements management knowledge from the fields of engineering and project management.

3.1 Explanation in Social Sciences & Humanities

Scholars in the social sciences and humanities (SSH), specifically, cognitive science, psychology and philosophy, have studied how humans engage in acts of explanation. As discussed above in Chapter 1, intuitively, we know that explanations are highly dependent on context. And existing SSH research on how humans explain our own thoughts and actions to each other has reaffirmed this by providing multiple schemas of explanation. In this section, I will describe the prevailing characterizations of the concept of explanation by drawing from SSH literature.

Specifically, this research project focuses on the concept of everyday explanations. Studies of the concept of explanation can be grouped into two classes: Scientific Explanation and Everyday Explanation. Scientific explanation, which has been studied in the field of philosophy, seeks general laws that govern relations between classes of events (Mayes, 2023; Woodward & Ross, 2021). And,

everyday explanation, which has recently seen an increased focus of study by multiple disciplines within the SSHs such as psychology and the cognitive sciences, seeks to address the causal question of why, when, and how particular phenomena (e.g., things, events, facts, situations, properties, decisions, etc.) occurred (Hilton et al., 2005; Miller 2019).

Everyday explanations are inherently transactional, characterized by the exchange of information through communicative actions. They are multi-predicated relational acts where someone *(explainer)* explains (communicates or provides) something *(explanans)* to someone else and/or oneself (*explainee)* about a phenomenon to be explained *(explanandum)*. When it comes to the *explanans*, an extensive body of SSH research has focused on causal attribution, which is the extraction of a causal chain and presenting it to a person. And xAI, in keeping with this tradition, has utilized causal explanation in the majority of xAI models. However, as Miller (2019) stated, ". . . explanations are not just the presentation of associations and causes (*causal attribution)*, they are *contextual*." We can grant that causal explanations are essential factors in providing an explanation and answering the "why" question. However, simply extracting a causal chain, presenting it to a person (and in the case of xAI, a lay person) and expecting them to be able to interpret a causal chain is not necessarily a sufficient explanation. SSH research over the past few decades has recognized this shortcoming and has yielded many more schema of everyday explanatory approaches that delve into aspects of causal chains beyond the simple causal attribution approach.

As Keil (2006) stated, ". . . causality seems to have a privileged role in most explanations". Therefore, to illustrate the various SSH-based explanatory concepts and categories, this section will use a running hypothetical example involving a causal chain of events involving a driver named Bob. The scenario will be referred to throughout this section in italics as *"e.g., text description of relevant factors from the Bob example below"*. This hypothetical example is an amalgamation and modification of scenarios provided in Hilton et al. (2005) and Lewis (1986). Following is the description of a hypothetical example involving a driver named Bob: *e.g., Bob was held up at his last appointment and*

was therefore driving beyond the local road's mandated speed limit. Because of this violation, he was stopped by the police and issued a speeding ticket. This made him even more behind schedule and he took a shortcut. On the shortcut, he encountered shards of glass which caused a flat tire. Note that just prior to Bob embarking on his journey, a winter storm had covered the roads in icy patches. And the shards of glass that Bob encountered were the result of someone dropping and shattering a parcel of glass when they slipped and fell on the icy patches on the road. The time delay due to changing the flat tire put Bob in the time frame of encountering a group of senior citizens at a crosswalk. Also note that the crosswalk was situated at the end of a curve in the road, so was only visible to Bob after he navigated the curve. When he encountered the senior citizens at the crosswalk, he swerved to avoid them and skidded off the road. This ended in Bob's car crashing into a tree.

In seeking an explanation for Bob's car crash, the answer(s) to the explanatory why and/or how type question(s) will involve identifying the key factor(s) in the causal chain that would suffice as an adequate explanation for a particular *explainee* and/or group of *explainees*. The explanation isn't just a matter of presenting this causal chain to *explainees* or assuming that there is only one and the same answer that is adequate for all *explainees*. Context matters. As is typically the case for most real-life *explananda*, there are many causes for the same *explanandum*, and depending upon context, the *explainee* will choose one as *the explanans* for them. As Lewis (1986) pointed out, an *explanandum* has multiple causes and they act jointly to make the *explanandum* ". . . inevitable, or at least highly probable, or at least much more probable than would otherwise have been. And the . . . [*explanandum*] depends on each" (p.214). Furthermore, *explainees* commonly speak of *the cause* of something, or "*the X*" despite knowing that there are many *X's* and each *X* is part of the causal history. When different *explainees* cite different *X's* as *the* cause, it is because that particular *X* is most salient for them. They ". . . may be looking for the most remarkable part, the most remediable or blameworthy part . . . [Essentially], some parts will be salient in some contexts, others in others" (p. 215). Therefore, to illustrate the various SSH-based explanatory concepts and categories (sections 3.1.1-3.1.4) and

demonstrate the importance of contextuality, I have chosen to use this one hypothetical example of a causal chain of multiple events.

3.1.1 SSH-Based Taxonomy of Explanation

There are many different ways to organize, group, catalogue and categorize the definitions and explications of the concept of explanation within the SSHs in general as well as within their individual disciplines. Two of the more relevant groupings that influenced this research project include the works of Miller (2019) and Kass and Leake (1987).

Miller's (2019) in-depth survey catalogued the vast body of SSH research on explanation to facilitate translation and application of SSH knowledge to the xAI field. Miller (2019) amalgamated this knowledge into the broad groupings of philosophical, psychological, cognitive process, and social interaction aspect foundations of explanation. Aspects of these four groupings will be explicated in detail in the following subsections of the proposed taxonomy. Here is a general view of each grouping: The grouping of philosophical explanation encompasses the ideas and concepts of the definition of explanation, types and levels of explanation, and the structure of explanation. Subgroupings of the definition of explanation include causal explanations; abductive reasoning; contrastive reasoning; types and levels of explanation including definitions provided by Aristotle's four causes; and structure of explanation including scientific explanation. The grouping of psychological explanation, specifically, the social attribution of explanation encompasses the ideas and concepts of how people attribute and explain behavior. Subgroupings of the social attribution of explanation include folk psychological concepts such as beliefs, desires, and intentions (BDIs); individual and group behaviors; and how norms and morals inform the motivations of actions and decisions taken by people. The grouping of cognitive processes of explanation encompasses the ideas and concepts of how people select and evaluate explanations. Sub-groupings of the cognitive processes involved in explanation include causal connections; types of reasoning (abductive, counterfactual, contrastive); selection of explanations; and evaluation of explanations. The grouping of social interaction explanation encompasses the ideas and

concepts of how people communicate explanation. Sub-groupings of social aspects of explanation include various conversational models of explanation and explanatory dialogue.

Kass & Leake (1987) approached the topic of explanation by viewing the world of anomalous events as types of phenomena to be explained. They provided a hierarchical categorization of explanation that is divided into three types of phenomena to be explained: intentional actions (e.g. deciding to drop out of school), material forces (e.g. an unexpected snowstorm) and social forces (e.g. an increase in the crime rate). Detailed explications of these concepts will be provided in the subsections below.

In the following sections (3.1.2-3.1.4), I will elucidate the ideas from these two exemplar papers as well as additional sources obtained through the usage of the snowballing technique. Backward snowballing is the technique of searching strategies applied to exemplar papers to identify additional references cited by them. To elucidate and clarify this knowledge, I have amalgamated these ideas into a proposed hierarchical classification of explanation types as shown in Figure 1's Social Sciences and Humanities Based Hierarchical Taxonomy of Explanation as well as a more detailed faceted taxonomy in Figure 2a & 2b's Social Sciences and Humanities Based Faceted Taxonomy of Explanation. As discussed in Chapter 4, these SSH-Based Taxonomies of Explanation are among the tools and aids prescribed to xAI engineers for their use during the development cycle. Appendix B contains the original versions of these taxonomies which were amended after Phase 3's proof-of-concept testing of the FGCR process (Chapter 5). In addition to the original versions of the taxonomies, Appendix B provides details of the changes made to them as a result of the FGCR's proof-of-concept testing. Note that these taxonomies are not exhaustive in either breadth or depth and are therefore intended as a starting point to capture what I consider to be the most relevant types of explanation that can be modified and added to in the future.

The top level of the taxonomy of explanation is divided into the two main categories of the "what" and the "how" of explanation which correspond to two main questions: "What type of

explanatory information are you seeking and/or providing?" and "How are you explaining?" The "what" can be thought of as a product and the "how" can be thought of as a process (Miller, 2019). The "what" is based on the majority of Kass & Leake's (1987) hierarchical categorization as well as portions of Miller's (2019) groupings and the "how" is composed of extractions from Miller's (2019) groupings.

Figure 1:

Social Sciences and Humanities (SSH)-Based Hierarchical Taxonomy of Explanation

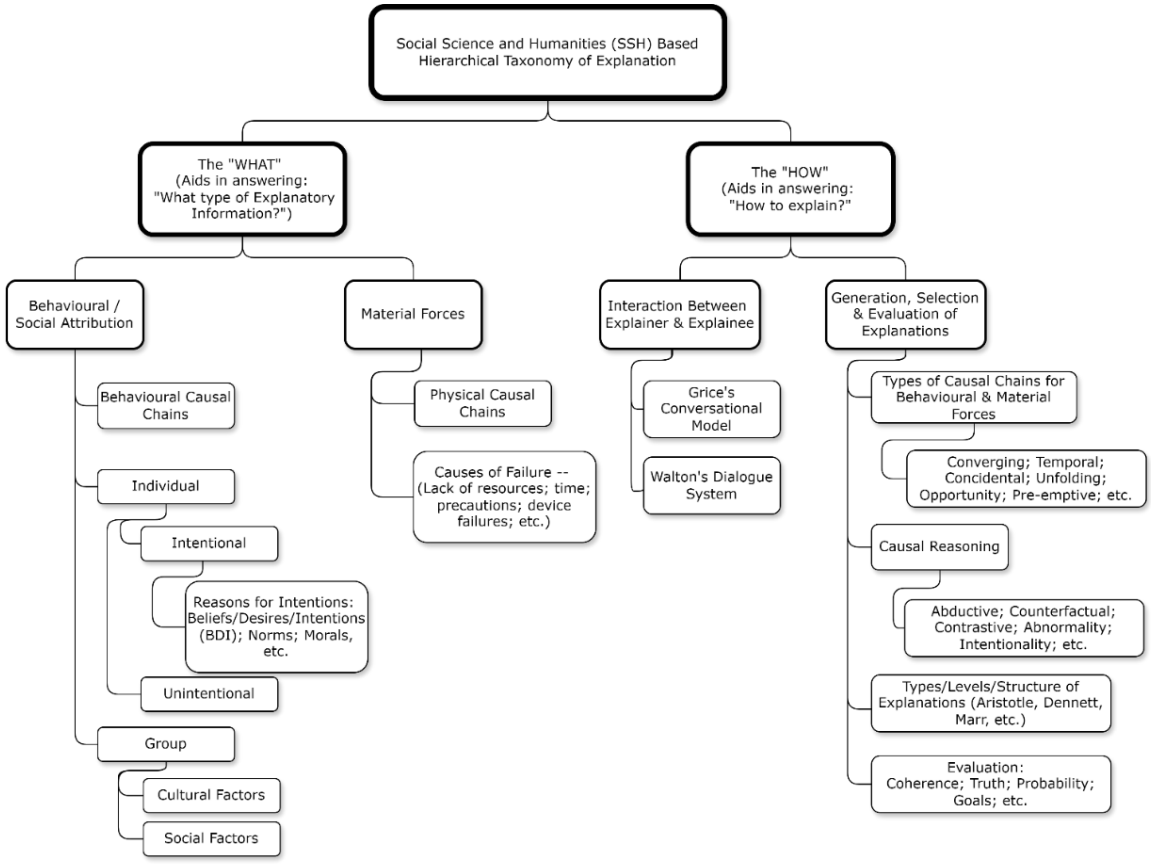


Figure 2a:

Social Sciences and Humanities (SSH)-Based Faceted Taxonomy of Explanation—The "WHAT"

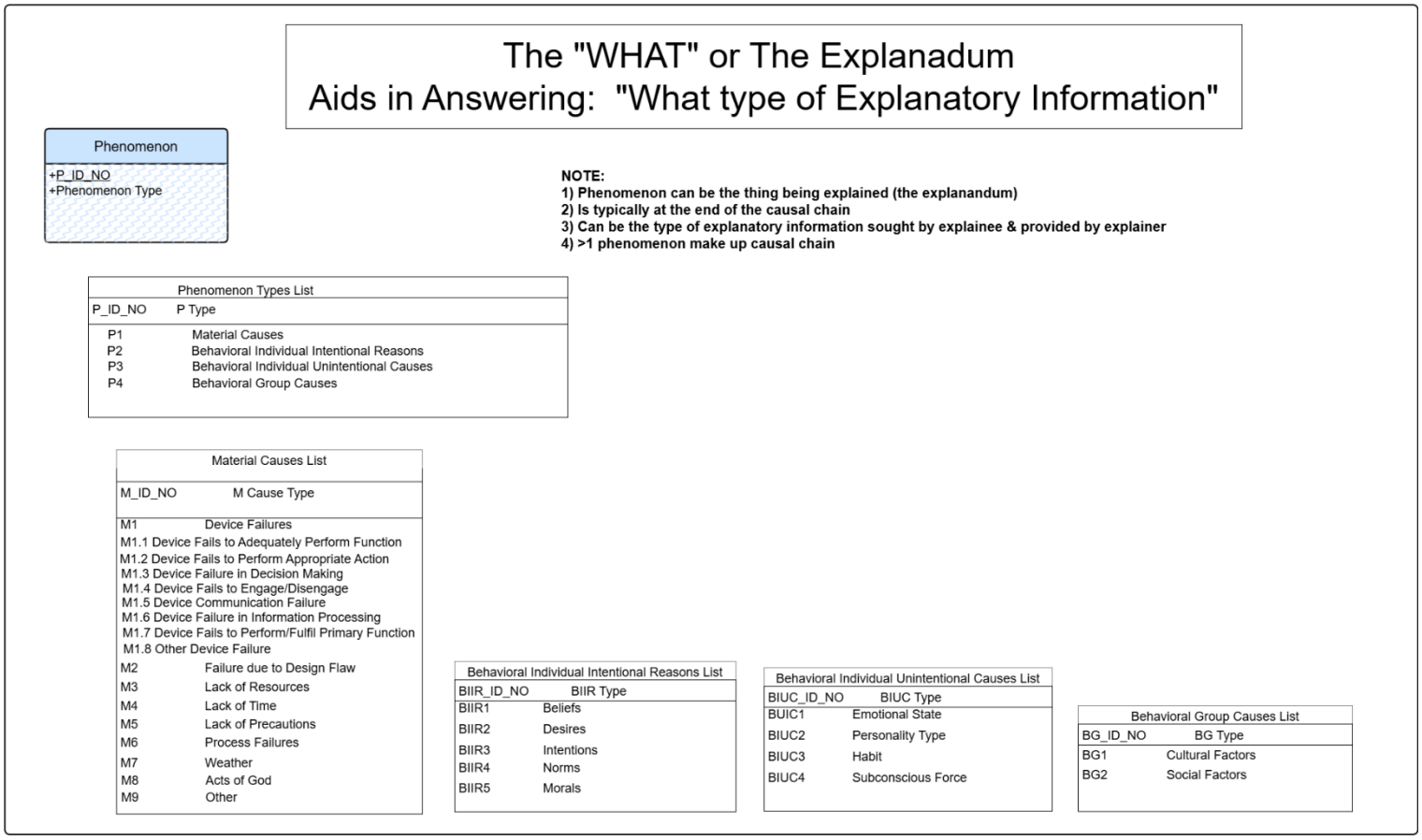
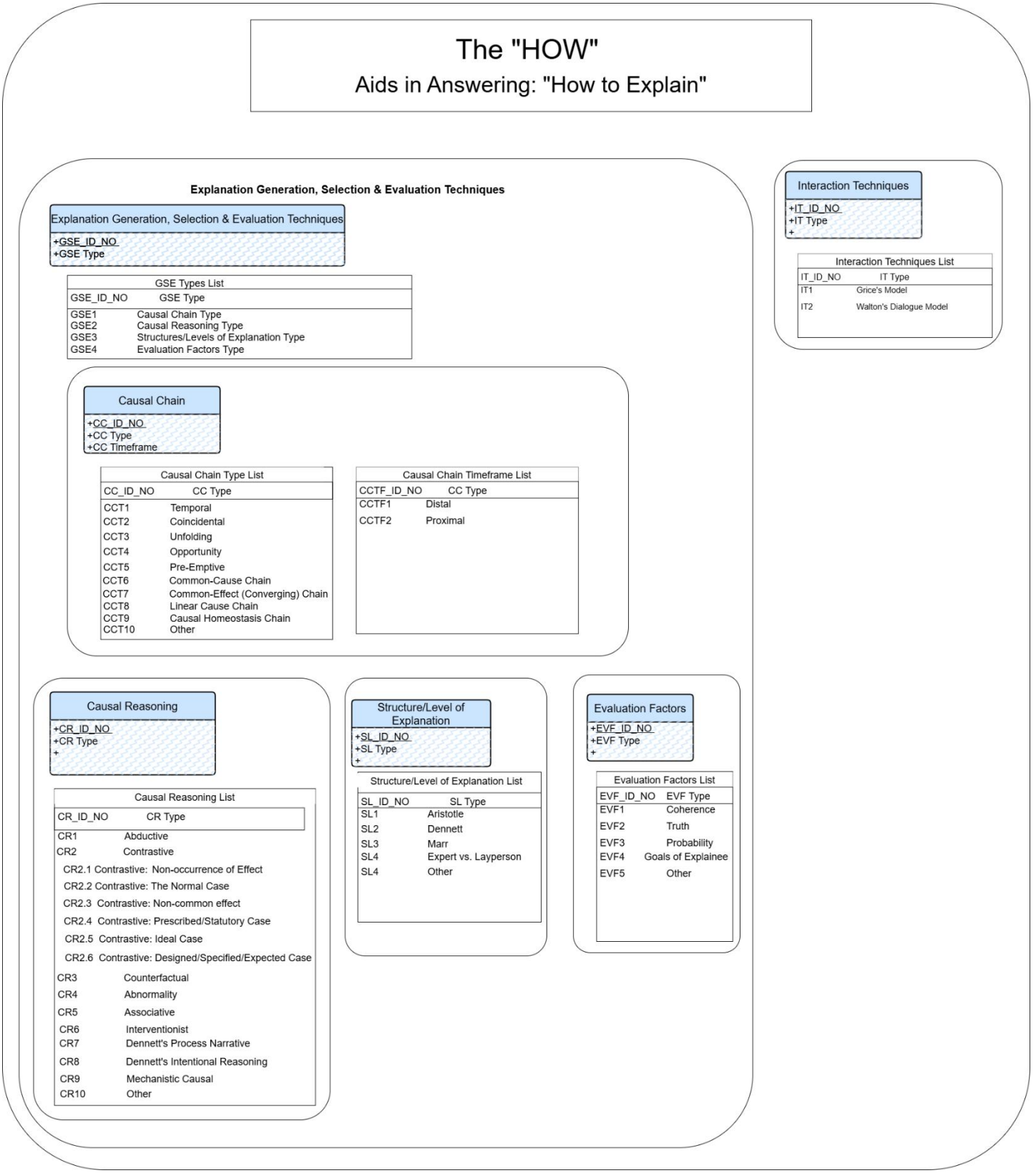


Figure 2b:

Social Sciences and Humanities (SSH)-Based Faceted Taxonomy of Explanation—The "HOW"



3.1.2 The "WHAT" of Explanation

The "what" category is divided into two types of phenomena to be explained: Behavioral and Material.

3.1.2.1 Behavioral Forces (Social Attribution)

The behavioral forces category involves phenomena that are brought about by humans. The behavioral forces are also referred to as the social attribution type of explanation which describes the causal attribution of social behavior. Social attribution defines how people attribute and explain their own or other people's behaviors. Behavioral forces are further subdivided into individual versus group behaviors. Individual behaviors are separated into intentional and unintentional behaviors.

Individual Intentional Behaviors

People use folk psychological notions to describe individual intentional behaviors. People use the notions of belief, desire, and intention (BDI) to understand, predict and explain individual human action (Miller, 2019). Intentional actions involve rational decision-making activities where an actor assesses the positive effects of a given action and weighs them against any negative side-effects (Kass & Leake, 1987). Based on philosophical and psychological theory of mind ideas, Malle (1999, 2006) provided a theory of how people attribute behavior of individuals (others and themselves) by focusing and expanding on the concepts of intentions and reasons as they relate to people's behaviors and their BDI mental states. Malle (1999) delineated between the usage of the terms "reasons" and "causes" to describe causal attribution of behaviors. He used the term "reasons" for intentional behaviors such as the mental states of BDI and the term "causes" to describe causal attribution of unintentional behaviors. Note that for consistency, in the next section's discussion on material forces that result in the phenomena being explained, I will use the term "cause" and not the term "reason". For an intentional action, we would be asked to give reasons regarding the belief, desire and/or intention of the actor. *e.g., Bob's decision to take a shortcut and speed*.

Kass and Leake (1987) also stipulated that intentional actions can be attributed to institutions as well as individuals. *e.g., Government reduces the speed limit on local roads to reduce number of fatal car crashes.*

Furthermore, moral judgments and norms can fall under behavioral factors for both individuals and institutions, and these can drive the reasons for action. *e.g., Bob believing that one should be punctual.*

Individual Unintentional Behaviors:

Unintentional behaviors involve an individual's personal states, emotional states, and subconscious forces such as habits, personality traits, etc. (Kass & Leake, 1987). *e.g., Bob may have a habit of speeding without even considering the associated risks or consequences.* Or*, Bob may be in an emotional state of nervousness about being on time for his next meeting which was a job interview. And, if under the influence of the above-mentioned mental states, although Bob is responsible for his actions, he did not act intentionally because he did not rationally weigh the option of speeding or not speeding.*

Group Behaviors

The category of group behaviors includes cultural and social factors that drive group behaviors (Kass & Leake, 1987). Social factors can provide explanations that have to do with stereotypical group behavior or behaviors that result from interactions of many independent actors whose actions are not coordinated. These actions are not intentionally driven and do not involve plans or goals. *e.g., A group of senior citizens slowly making their way through a crosswalk.* Cultural factors provide explanations related to traditional group behavior for particular cultures. *e.g., Some cultures, out of respect for elders, make accommodations for their mobility issues in city planning such as ease-of-use street crossings and longer stop lights.* Or, *the business organization to which Bob is interviewing for a job, has a culture that places punctuality as a highly valued trait.*

3.1.2.2 Material Forces

The material forces category includes explanations that identify causes involving aspects of the material situation of the world (Kass & Leake, 1987). These are non-social attribution type of explanations. They are impersonal and describe situational causality. Material forces are further subdivided into descriptions of the material properties of the world and identification of aspects of the material situation. Material explanations that identify material properties of the world are focused on building causal chains that use biological and/or physics laws of the world. *e.g., Winter storm and formation of icy patches on road.* Material explanations that identify aspects of the material situation include device problems; lack of precautions by some actor; planning/execution failure by actor; deadline being missed; lack of a resource necessary for event to take place; etc. *e.g., In addition to Bob driving at high speed, the tires on Bob's car were worn and therefore, he skidded.*

3.1.3 The "HOW" of Explanation

The "how" category is divided into two: The conversational model of explanation and the generation, selection, and evaluation processes of explanation.

3.1.3.1 Conversational Model of Explanation

The conversational model of explanation fits into the viewpoint of everyday explanation as a transactional social process where an *explainer* explains *an explanans* to *an explainee* about *an explanandum*. This model is based on Hilton's (1990) and Grice's (1975) work. Hilton (1990) asserts that explanation is a social process of conveying explanatory information in a conversation between an *explainer* and an *explainee*. When engaged in a conversation, an *explainer* presents an explanation to an *explainee*, and they tend to follow the basic rules of conversation. Conversational models of explanation recognize that explanations are interactive conversations between an *explainer* and *explainee*. According to Grice's (1975) conversational model, there are four maxims of conversation: Quantity, Quality, Relation and Manner. Under the maxim of quantity of information to be provided, the rule is for the *explainer* to provide the appropriate amount of information as is required for the purposes of

conversation. Two maxims are specified under quantity as: ". . . make your contribution as informative as is required . . . [and] do not make your contribution more informative than is required." Under the maxim of quality, the rule is for the explainer to try to provide information that is true. He specified two maxims under quality as: ". . . do not say what you believe to be false . . . [and] do not say that for which you lack adequate evidence." Under the maxim of relation, a single maxim is succinctly specified as: "Be relevant". Under the maxim of manner, which he described as ". . . how what is said is to be said", he specified one super-maxim as: "Be perspicuous". Various maxims under this super-maxim include: ". . . Avoid obscurity of expression, avoid ambiguity, be brief (avoid unnecessary prolixity), be orderly, etc." For an ideal conversation, these can be summarized by the adage: "Only say what you believe [quality]; only say as much as much as is necessary [quantity]; only say what is relevant [relation]; and say it in a nice way [manner]" (Miller, 2019). Another conversational model of explanation is Walton's (2011) dialectical system of explanation that involves three stages of dialogue which include an opening stage, an explanation stage and a closing state. Walton's (2011) system emphasizes an iterative model of conversation where the transfer of understanding, or lack thereof, is viewed as a crucial criterion for closure of an explanatory dialogue.

3.1.3.2 Explanation Generation, Selection and Evaluation

People use various cognitive processes to generate explanations by identifying causes of phenomena; select explanations by narrowing down the identified causes of phenomena to a small subset; and evaluate explanations by judging the quality of an explanation (Miller, 2019). The first two processes, explanation generation and selection are in the purview of the *explainer*. And the explanation evaluation process to determine the adequacy of the explanation is in the purview of the *explainee*.

Following are the main concepts and cognitive processes that people use to generate, select, and evaluate explanations:

Types of Causal Chains

As Hilton et al, (2005) stated, we humans have a ". . . propensity to generate causal chains when seeking to understand events . . ." (p. 44) Many different "what's" from the taxonomy above (such as behavioral forces or material forces that are based on biological or physical laws or causes of failures, etc.) are involved in building causal chains. Furthermore, because there are many different phenomena to be explained in many different contexts, there are many different types of causal chains. Two relevant descriptions of causal relations are provided by Keil (2006) and Hilton et al. (2005).

Keil (2006) identified four types of causal relations: Common-cause, common effect, linear causal chain, and causal homeostasis. For common cause explanations, a single cause is seen as having a branching hierarchy of consequences. The most obvious generic example is a virus spreading throughout a population. Or in the hypothetical case involving Bob's car crash example, *it could be* t*he winter storm causing icy patches across the terrain that Bob navigates and this subsequently leads to multiple branching consequences such as someone slipping on ice and dropping glass on the roadway prior to Bob's arrival on that roadway; senior citizens walking more slowly at the crosswalk; Bob's car skidding when he swerved on the ice, etc.* Common effect explanations involve cases where causes converge to create an event. *e.g., Bob being held up; glass on road left by someone else; policeman on duty on the particular road at the time that Bob was speeding all converge to place Bob at the site of the car crash, the intersection with senior citizens at the same time.* Linear causal chains are a special case of common cause and common effect explanations, and they are rare because there is one unique serial chain from a single initial cause through a series of steps to a single effect. e.*g., the part of the causal chain where Bob swerving to avoid the senior citizens leads to skidding off the road and crashing into a tree.* Causal homeostasis explanations explain how interlocking set of causes and effects results in a set of properties that endure over time. A generic example is a set of feathers, hollow bones, and nest building that all reinforce the presence of each other in birds over time.

Hilton et al. (2005) identified five kinds of causal chains that focus on the relationships between distal versus proximal events: Temporal, coincidental, unfolding, opportunity, and pre-emptive. For temporal causal chains, the distal does not constrain proximal. Events are temporally switchable, (i.e., events A and B together cause C, and the order of A and B is irrelevant). *e.g., Bob could have received the speeding ticket after he took the shortcut and got a flat tire still putting him in the time frame of encountering the group of senior citizens at the crosswalk.* For coincidental causal chains, the distal does constrain the proximal. The causal relationship holds in this particular case but not in general and events are switchable in time in general but not in this particular case, (i.e., A causes B, B causes C, etc. in this particular case but the general relationship does not hold). *e.g., Bob swerving to avoid the senior citizens causes the crash, but swerving does not generally result in car crash.* For unfolding causal chains, events are not switchable in time either in this case or in general, (i.e. A causes B and B causes C, etc. in this particular case but also in general). *e.g., the winter storm causes a temperature drop and precipitation which results in icy patches on the road.* For opportunity causal chains, the distal enables proximal events, they are not switchable in time, neither in this case, nor in general, (i.e., A enables B, B causes C). *e.g., Bob leaving late enables encountering the ensuing obstacles (being stopped by police, flat tire, senior citizens at crosswalk, swerving) the last of which cause the car crash.* For pre-emptive causal chains, the distal precedes proximal. The events are switchable in time in this case, and can be switched in general, (i.e., A is independent of B. In this case, B causes C, but A would have caused C if B did not occur. In general, A causes C, B also causes C). A generic example includes the murdering of a victim by poisoning him thereby sealing his fate but before he dies of poison, he is stabbed to death by another person.

Causal Reasoning

Abductive Reasoning

Abductive reasoning is closely related to explanation. Various philosophers and psychologists including Peirce (1903) and Harman (1965) have described abductive reasoning as a process that aids

people in determining what they consider to be the best explanation. Abductive reasoning is the process of observing some event(s), inferring causes that explain those events by generating hypotheses about them, judging the plausibility of the hypotheses, and selecting the best hypothesis as the explanation (Miller, 2019).

Suppose you have a premise (e.g. event, phenomenon, hypothesis, or body of knowledge) denoted as "a" that leads to a conclusion (e.g. general principle, phenomenon or event) denoted as "b" (i.e., "a" → "b"). In abductive reasoning, we derive a hypothesis "a" to explain a phenomenon "b". Essentially, the reasoning goes from the effect to the cause where we infer "a" as an explanation of "b". This is in contrast to deductive reasoning where we derive "b" from "a", and here, "b" is a formal logical consequence of "a" in that given the truth of premises "a", a valid deduction guarantees the truth of conclusion "b". And, in inductive reasoning, we infer some general principle "b" from a body of knowledge "a" where "b" does not necessarily follow from "a".

Counterfactual Reasoning

Counterfactual reasoning involves asking "What if" or "How to" questions such as "What is needed to (or how can someone) change" P (the fact, outcome or phenomenon that requires an explanation)?" People perform mental acts of simulation of counterfactual cases to derive a good explanation. As Kahneman and Tversky's (1982) research showed, when people answer questions about past events, they complete "if only" sentences to consider counterfactual cases. They look at the mutability of events, the degree to which the event can be "undone", to judge causality. In fact, per Kahneman and Tversky (1982), people, in order to judge causality, tend to focus on mutating abnormal events more than other types. *e.g. What if Bob had not encountered a group of senior citizens at a crosswalk which was situated at the end of a curve in the road which was only visible to Bob after he navigated the curve?*

Contrastive Reasoning

Contrastive reasoning in explanation goes beyond the causal approach—which asks the "why" question in a vacuum—by framing the "why" question against a specific contrasting alternative: It asks, "Why P (the fact that requires explanation) rather than or not Q (the foil)?" It is based on works by Hesslow (1988), Lewis (1986) and Lipton (1990), which assert that contrastive explanations are easier for humans to grasp and process than full causal attribution because explanations are not simply about causal attribution. Also, as Hilton (1990) outlined, there are different types of contrast cases depending upon the type of causal questions. People use counterfactual reasoning that answer different types of causal questions which align with different types of contrast cases. Table 1 from Hilton (1990) provides a typology of causes, contrast cases and explanations:

Table 1:

Typology of Causes, Contrast Cases and Explanations

Type of Cause	Type of Contrast Case	Type of Implied Question
Millian sum of necessary conditions	Non-occurrence of effect	"Why X rather than *not X*?"
Abnormal condition	The normal case	"Why X rather than *the default value for X*?"
Differentiating factor	Non-common effect	"Why X rather than *Y*?"
Moral or legal fault	Prescribed or statutory case	"Why X rather than *what ought to be the case*?"
Design fault or bug	Ideal case	"Why X rather than *the ideal value for X*?"

In the non-occurrence of effect contrast case, the foil that is the negation or absence of the fact. *e.g. Why did Bob drive fast through the curve on the road rather than slow down?* In the normal contrast case, the foil is the default value for the fact. *e.g. Why did the city put a crosswalk at the end of a curve on the road rather than on a straight portion of the road where visibility is higher?* In the non-common effect contrast case, the foil is a differentiating factor from the fact. *e.g. Why did Bob take the shortcut that put him on this curved road with a crosswalk at the end of the curve rather than take the straight route?* In the prescribed or statutory contrast case, the foil is what ought to be the fact. *e.g. Why did Bob speed rather than drive at the legal speed limit?* In the ideal contrast case, the foil is the ideal value for the fact. *e.g. Why did the city not put flashing lights before the curve indicating/warning drivers of the crosswalk at the end of the curve?*

As a result of the proof-of-concept testing of the FGCR (see Chapter 5), I amended the above-mentioned causal reasoning list that was synthesized from SSH literature: In addition to the typology of contrast case list in Table 1 above, I introduced a more nuanced "design fault or bug" type of cause and labeled it as "Design/Specified/Expected" contrast case with the associated type of implied question phrased as "Why X rather than *the designed/specified/expected value* for X?" For this contrast case, the foil is the designed/specified/expected value for the fact. *e.g. Why did Bob drive on worn tires rather than follow the specified maintenance schedule for his car tires?*

Furthermore, during proof-of-concept testing, I observed several patterns that are worth mentioning: With respect to phrasing contrastive, differentiating or comparative questions, participants rarely employed the academic phrase "rather than", which is used widely in the SSH literature I reviewed. Instead, people opted for a diverse set of alternatives such as "instead of", "versus, often abbreviated as 'vs.'", "against", "against", "compared to", "in contrast to", "in place of", " in lieu of", "as opposed to", "an alternative to/is", "conversely", "however", "side by side to", "on the flip side", "on the other hand", etc. Also note that in cases of contrastive non-occurrence of effect, the foil was rarely stated explicitly. People typically posed contrastive questions with an implied foil in terms such as "Why didn't it/they [action verb such as do] X?" To summarize these patterns from the data, when people questioned a system or technology failure related to performing a function, they frequently adopted negative constructions with foils, typically phrasing their queries as: "Why didn't it/they [action verb such as do] X instead of [action verb such as do] X?". Since the FGCR aims to bridge the gap between formal SSH-based taxonomies and everyday explanatory queries, these findings are critical for xAI developers to ensure that requirements activities capture a broad range of the human linguistic spectrum.

Reasoning Required for Different Classes of Explanatory Questions

Although explanations typically involve asking and answering "Why?" questions, they may also involve other types of questions such as "How?", "What?", "What if?", "What for?", "How come?", etc.

These questions can be grouped into classes which require different types of reasoning to answer them. Miller (2019) proposed a schema for explanatory questions based on Pearl and MacKenzie's *Ladder of Causation* (2018). Table 2 below, from Miller (2019) provides a summary of the classes of explanatory questions and the reasoning required to answer them.

Table 2:

Classes of Explanatory Questions and Reasoning Required to Answer

Question	Reasoning	Description
What?	Associative	Reason about which unobserved events could have occurred given the observed events
How?	Interventionist	Simulate a change in the situation to see if the event still happens
Why?	Counterfactual	Simulating alternative causes to see whether the event still happens

A "What?" question involves associative reasoning to determine which unobserved events (the "what") could have happened from observations about events that did happen. It focuses on the acts of seeing or observing by asking, "What could observing event X tell me about the likelihood of observing event Y?" or hypothesizing about what event Y could occur given my observation of event X. A generic example would be "What does observing symptom X (coughing) tell me about the likelihood of having disease Y (lung cancer)?" Or in Bob's case, asking, *"What does seeing a crosswalk at the end of a curve in the road tell me about the probability of pedestrians being hit by cars or cars more often than not?"* A "How" question involves interventionist reasoning to determine if an event would still happen under a different situation. It focuses on changing the environment (the "what is") by hypothesizing about the effects of doing or intervening. In interventionist reasoning, we would ask, "How can we achieve Y or make Y happen?" or hypothesize about what event Y could occur if we do X. A generic example would be "How can we prevent lung cancer?" which would involve hypothesizing about whether we can achieve this by banning cigarettes. Or, in Bob's case, asking, *"How can we prevent car crashes at that crosswalk?" and hypothesizing about achieving this by moving the crosswalk location or putting flashing lights before the curve in the road to warn about the crosswalk ahead.* A "Why?" question involves counterfactual reasoning to undo events that happened and simulate other possible

events. It focuses on acts of imagination, retrospection and understanding and asking, "What if I had done X?" to hypothesize about whether outcome Y would still occur. A generic example would be asking "What if I had not smoked for the last 20 years?" to hypothesize about whether getting lung cancer would still have occurred. And in Bob's case, asking, *"What if Bob had driven on well-maintained tires through the curved road?" where the simulation would involve an alternative cause of Bob driving fast on well-maintained tires instead of worn tires through the curved road and swerving to see if the car still skidded and crashed.*

In addition to the above-mentioned explanatory questions, Dennett (2017) proposed that there are two other types of "Why?" questions: "How come?" and "What for?" According to Dennett (2017), "How come?" questions ask for a process narrative and "What for?" questions ask for a reason which implies some intentional thought behind the cause. For example,

> "Why are you handing me your camera?" asks *what* are you doing this *for*? . . . [and]. . . "Why does ice float?" asks *how come*: what it is about the way ice forms that makes it lower density than liquid water?

In my interpretation of Dennett's definition of "How come?" questions, there is an intentional or designed end goal aspect to them due to the inclusion of the word "come". To illustrate this aspect of the "How come?" question, I point to an example given by Dennett which involves the question "Do you know the reason why ball bearings are spherical?" which is answered by a process narrative that states ". . . the ball bearings were made on a lathe of sorts, which spun the metal. . . ". In my interpretation, implied in questions such as these and their answers, is an intended or planned end goal aspect. It isn't simply a process narrative, but a process narrative for the phenomenon that was the intended objective of the causal chain that led to it.

Although Dennett's (2017) "How come" questions require process narrative reasoning, I note that there is a need for another class of reasoning pertaining to "How" questions, one that involves *explananda* that are unintended or unplanned outcomes. I propose a new class of explanatory reasoning

under an umbrella of simple "How" questions that do not have any predetermined or designed objective in mind. These simple "How" questions could encompass both intentional and unintentional outcomes. This type of "How?" question simply asks for a process or mechanistic explanation of the causal chain and the necessary salient components of the causal chain leading up to the *explanandum*. Sometimes people just ask simple "How" questions without any speculative thoughts including pre-determined ideas; comparative contrast cases; counterfactual reasoning about mutability of events; simulations of changes in the causal chains; alternative causes; etc. I am labeling this type of causal reasoning as "Mechanistic Causal Reasoning".

This proposed mechanistic causing reasoning category complements causal inference by involving the role of mechanisms or processes that underlies an event and is based on Ahn et al. (1995)'s work. It focuses on the plain and simple "How?" type questions. This type of reasoning involves discovering the process that underlies the relationship between cause and effect. Mechanistic causal reasoning seeks some component or components of an event which is thought to have causal force or necessity. *e.g. the salient components in the causal chain include Bob's tardiness and ineffectual planning → Bob making the decision to take the shortcut → Bob getting a speeding ticket and encountering a flat tire → Bob driving fast around a curved road → Bob encountering a group of senior citizens at a crosswalk at the end of the curve → Bob swerving fast on worn tires → ultimately results in the car crash.*

Although the proposed mechanistic causal reasoning has similarities to the interventionist reasoning described above and Dennett's (2017) process narrative, it is different. Unlike interventionist reasoning, this isn't about seeking alternative causes and/or hypothesizing alternative events, it is about seeking a causal chain composed of necessary components in the causal chain linking cause to effect. And unlike Dennett's (2017) use of the second word in "How come" to delineate a process narrative which has a predetermined intentional objective in mind for the outcome, this is about a straightforward process narrative without any predetermined or speculative thoughts about causes. Note that the need

for this new class of reasoning pertaining to simple "How" questions that seek explanations arose as a result of FGCR's proof-of-concept testing and data analysis (Chapter 5). Details of the addition of this new category are outlined in Appendix B.

Framing Explanation

There are different ways to frame explanations by providing stances or modes of construal to yield different insights including works by Aristotle (1984), Dennett (1989), and Marr (1982). For example, Aristotle identified four categories of causes that could provide answers to "why-questions: material, formal, efficient, and final. In *Physics II.3*, Aristotle states that knowledge requires grasping the 'why' or the cause (194b18-19) (Aristotle, et al., 1984). And, he outlines four causes of change: (1) material, the ". . . that out of which as a constituent thing comes to be" (194b24); (2) formal, ". . . the account of what the being would be" (194b27); (3) efficient, ". . . the primary source of the change . . . that which changes something of that which is changed" (194b30-32); and (4) final, ". . . the end" (194b33) (Aristotle & Charlton, 1984).

e.g., Framing explanations for the tires on Bob's car would yield the following: The material cause identifies the "rubber" as the substance or material of which the car tire is made. The formal cause identifies the "being round or roundness" as the form or properties of the car tire that makes it what it is. The efficient cause identifies the "tire manufacturer" as the primary source of the change and provides mechanistic explanations about what causes the car tire to change. The final cause identifies the "movement of the vehicle" as the "end", goal, purpose, functional or teleological explanation of the car tire.

3.1.4 The How and the What of Explanation

As demonstrated above, there are many different SSH-based schema of explanation. And depending upon context, people pick and choose any one or a combination of types of explanation out of the many identified above in the taxonomy.

In order to generate, select and evaluate the most adequate explanation, people will typically engage in the transitive process of communicating the *explanans* between the *explainer* and *explainee*. This process can use the conversational models of explanation which could be based on Grice's (1975) maxims or Walton's (2011) iterative dialectical system among others. The process of communicating the *explanans* between the *explainer* and *explainee* can be iterative in order to ensure that an adequate explanation is provided to the *explainee*. The context for an explanation covers not only the phenomenon to be explained but also the situation and the *explainee*. And because the *explainee* will determine the adequacy of an explanation, it is crucial that the process prioritizes the needs of the *explainee*.

Because people seem to have a disposition to focus on causality when it comes to explanations, they typically tend to look at key causal connections to the phenomenon being explained when generating explanations, (Keil, 2006, p.7; Hilton et al., 2005, p.44). Furthermore, because most causal chains can become too complex and large to grasp, people typically do not provide all causes for a phenomenon but instead select a subset of causes from the generated causal connections. These two processes of generating and selecting explanations, which are in the purview of the *explainer*, typically encompass the following pattern: Identification of the "what" from the myriads of items listed in the taxonomy above. The use of causal reasoning processes (e.g. abductive, counterfactual simulation, contrastive, etc.) to generate various types of causal chains depending upon the context, (Hilton, et al., 2005; Keil, 2006). The criteria for selecting the most appropriate or best explanation predominantly involves contrasts as the primary way people select explanations by looking at the difference between fact and foil (Hesslow, 1988; Hilton, 1990; Lewis 1986; Lipton 1990). People may also look at intentionality, abnormality, framing, etc. when generating and selecting the best explanation.

The third process, explanation evaluation, falls within the purview of the *explainee*. The most important criteria people use to evaluate explanations are coherence with prior beliefs, truth, probability, simplicity, and generalizability (Miller, 2019). People prefer simpler, more coherent and general

explanations than complex ones. As Thagard (1989) asserted, coherence is the primary criteria for explanation where people prefer explanations that are consistent with prior beliefs and simpler explanations.

Essentially, the type of explanation generated, selected, evaluated, and communicated will depend upon the context of the phenomenon being explained. Furthermore, more than one type of explanation or a combination of different types of explanation may be applicable to a given phenomenon. The exegesis and application of the SSH-based taxonomy to Bob's car crash example throughout this chapter demonstrate that multiple valid and not necessarily mutually exclusive explanatory accounts may exist for any given phenomenon. And ultimately, the determination of the adequacy of the explanation(s) will depend upon the *explainee*. By placing the *explainee* at center of process, we can see that an adequate explanation about the cause of the car crash will vary from *explainee* to *explainee*: *e.g., For the Police or judge, the explanation about the cause of the car crash could be Bob not only driving over the speed limit but also considering road conditions, driving recklessly. For Carol, Bob's wife or his psychiatrist, it could be Bob's habit of being habitually late and incurring five other speeding tickets in as many months which may be a result of his childhood experiences of living amongst procrastinators. For the senior citizens, it could be Bob's reckless driving plus various factors such as the city planners who placed the crosswalk in a less visible spot at the end of the curve in a road and the lack of timely salting of the road after a winter storm. For the car manufacturer, it could be Bob's lack of tire maintenance and choosing to drive on worn tires. For Bob's future employer, it could be his poor time management and planning skills where he committed to back-to-back meetings with inadequate margin for extra time in between. For a bystander not involved in any of these events, it could be the winter storm that caused icy patches in multiple places thereby adding more time to Bob's commute and the ensuing time crunch resulting in Bob choosing to speed.*

Taken in aggregate, the recurring applications of Bob's car crash example to the taxonomic categories collectively illustrate that no matter the phenomenon to be explained, depending upon the context (situation and *explainee*) there will rarely be one universal adequate explanation.

3.2 Artificial Intelligence and Machine Learning

Because the purpose of xAI is to increase the transparency in AI systems that are currently opaque, I will situate the discussion of xAI within a basic summary of the field of AI. Therefore, in this section, I will amalgamate ideas from the works of Alpaydin (2014), Mohammed et al. (2017), Ng et al. (2023), and Theobald (2017) to provide basic background information about the complex subject matter of AI systems.

Machine learning (ML) is a branch of AI at the intersection of computer science and statistics that enables computers to identify patterns; learn; perform specific functions including human tasks, actions and/or decisions; and provide accurate results. ML is a process of programming computers to learn from data, by allowing them to analyze example data or past experiences and using that information to improve their performance over time by optimizing a specific criterion. A traditional computer program is an algorithm composed of a sequence of deterministic instructions that a computer carries out to transform inputs to outputs. In comparison, ML programs apply statistical modeling to construct a model which is an algorithmic equation that transforms inputs to outputs without direct deterministic instructions. ML programs take input data and use it to train a model by adjusting modifiable parameters through an iterative optimization process that minimizes the prediction error and produces the desired output for a particular task. Some ML models are transparent and understandable while others are as opaque as black boxes whose workings are not intuitively explainable (Alpaydin, 2014; Mohammed et al., 2017).

An example of a task that humans perform easily, regularly and take for granted is facial recognition. Even though humans can perform this task, we do not know exactly how the brain

transforms the visual inputs to recognise faces, therefore, we cannot write traditional deterministic computer algorithms that approach human-level performance for this task. The complexity of that task – which requires the ability to identify faces from varying angles, lighting, expressions, and ages—makes a rule-based approach impractical. However, due to the availability of massive amounts of data (pictures of faces) and computing power, we can use ML to learn a model that performs this task. To achieve this, input data of pictures of faces is split into training and test data. The training data is used to develop a model by repetitively and incrementally updating and optimizing modifiable parameters to recognize patterns in the distinctive features and the way they are combined to define faces. After the model has been successfully developed and its accuracy is determined to be satisfactory, the model is tested using test data which it has never seen before (Alpaydin, 2014). There are many different types of ML techniques and the choice of model or combination of models selected will depend on the task, action, decision, or action being undertaken by the ML technique. There are three overarching categories of ML programs: supervised learning (SL), unsupervised learning (USL), and reinforcement learning (RL) (Mohammed et al., 2017; Theobald, 2017).

The objective of SL is to learn patterns and infer a function from labeled training data. SL infers the relationship between variables and known outputs. The data is considered "labeled" because both the input and the corresponding correct output values are known and provided in the training dataset by the "supervisor". SL can be further broken down by the type of problem tackled, namely, a classification problem which predicts discrete categories or a regression problem which forecasts outcomes. A classification model identifies the category to which an item belongs, and a regression model predicts a continuous output (Mohammed et al., 2017; Theobald, 2017). There are many types of SL techniques including but not limited to regression analysis, decision trees, k-nearest neighbors, artificial neural networks (ANNs), and support vector machines (SVMs). For this paper, only regression analysis and ANNs will be discussed.

Unsupervised learning (USL) programs use unlabeled data and lack "supervisors" in that not all variables and data patterns are classified or labeled, and the correct outputs are not known in advance. USL techniques uncover the hidden structure in unlabeled data.

Reinforcement learning (RL) programs use observations from interactions with the environment to take actions and learn to produce outputs that maximize cumulative rewards and minimize risks to achieve the objective(s) of the task at hand. Due to its complexity, RL can be explained through an analogy: suppose you have a robot in a maze, and its objective is to find a way through the maze and reach a specified goal location. The robot is allowed to make one of four compass direction moves (left, right, forward, backward) that do not result in hitting a wall. A reinforcement learning algorithm would solve the problem by learning the "best" sequence of actions that has the maximum cumulative reward. In RL, there is an agent which is the decision maker that is placed in an environment. In the case of the robot in a maze, the agent is the robot, and the environment is the maze. At any time, the environment is in a certain state that is one of a set of possible states. In the case of the robot in a maze, the position of the robot in the maze is a state at any time. The agent is allowed a set of actions possible in any given state. In the case of the robot in a maze, the set of possible actions are the possible directions the robot can take without hitting any wall. Once the agent chooses an action, the state changes. The solution or output requires a sequence of actions and feedback is given in the form of rewards and generally only given when the complete sequence of actions is carried out and completed. The reward defines the problem and is necessary for the learning agent. RL is also known as "learning with a critic" vs. SL which is learning with a teacher. The critic does not tell the agent what to do, never informs in advance, only tells the agent how well it has been doing in the past, and feedback is scarce (Alpaydin, 2014; Theobald, 2017).

Neither USL nor RL is the focus of this thesis and will not be discussed further. The focus of this thesis will be tackling the black-box problem associated with SL techniques of ANNs. ANNs have

been chosen because they are commonly used in the driving and motion control tasks of AVs which are the target AI system for the case study research of this project.

In SL, two of the most common techniques are regression and classification. Input data for ML techniques comes in many forms including structured and unstructured. Structured data is used for SL techniques and is defined and labeled in a table. For the remainder of the discussion on learning techniques, the notation for structured data is outlined in Table 3 below:

Table 3:

Example of Structured Data Matrix

Matrix (**X**) for SL Techniques				
	Vector X_c			Vector Y
	Variable 1=$X_{1:}$	Variable 2=X_2:	Variable n =X_c:	Output (y):
Observation 1 (r=1)	x_{11}	x_{12}	x_{rc}	y_1
Observation 2 (r=2)	x_{21}			
Observation 3				
Observation m	x_{r1}	x_{r2}	x_{rc}	y_r

Data is structured in a table format known as a matrix. A matrix is denoted "**X" (bold, upper-case)** with "r" number of rows and "c" number of columns. Each column is a vector. Columns represent input vectors "X" (upper case) and an output vector "Y" (upper case). In each column, there is a feature, attribute, or variable and an output. And in each row is a single observation or example of a given variable. Vector "X_c" would represent input scalar values *"x_c"* (*italicized lower-case)* of the c^{th} variable. Vector "Y" would represent the output values. Each row represents an observation "r", the r^{th} element of any variable X would be the scalar value "x_r", and the r^{th} element of c^{th} variable in **X** would be "x_{rc}**"** (*italicized lower-case)*. Essentially, the dataset is composed of scalar values "*x"* (*italicized lower-case*) and "*y"* (*italicized lower-case*).

The learning program deciphers the patterns in the data and creates a model that can predict the output for new datasets. The model denoted by "g(*)", is an approximation of the actual output (Y) and it is created by the machine learning program. The loss (cost or error) function, denoted by "$\mathscr{L}$(*)", is the difference between the desired or actual output and the approximation g(*).

An SL program that tackles a linear regression problem would include tasks that require the prediction of continuous numerical value outputs. It models and identifies the relationship between input and output variables. An example of a linear regression problem would be the task of predicting the market value of a used car by analyzing the relationship between various attributes of a car and its selling price based on historical data. For this example, the Vectors $X_1 \ldots X_c$ would represent feature variables such as Variable 1=X_1=Year of Car; Variable 2=X_2=Make of Car, Variable n=X_c=Color of Car … and Vector Y represents the output values of y=price of car. In this case, g(X|W) is the model where *X* represents the input data and W represents the parameters to learn from the data.

A linear equation for a single input would be in the form:

$$g(x) = w_1 x + w_0$$

where *x* represents the scalar values and w_1 and w_0 represent the parameters that optimize the equation g(*x*), an approximation of the actual output values. g(*x*) is iteratively optimized to match the actual output values. (Shown graphically in Figure 3 from Theobald (2017)) And, a linear equation in matrix format that analyzes all of the input data would be:

$$g(X|W) = WX + W_0$$

Once the model has been created and parameters have been identified for each variable, optimization will be undertaken. To optimize the parameters, the loss function needs to be minimized. The loss function $\mathscr{L}(Y, g(X|W))$ computes the difference between the desired output (Y) and the model's approximation value, g(X|W) and represents the errors in the outputs. The optimization procedure uses methods such as least squares criterion for linear models that minimizes the sum of the squared prediction errors, or gradient-based algorithms, etc., to minimize the loss function $\mathscr{L}(*)$.

An SL program that tackles a classification problem to predict discrete classes would address questions such as "Is it "A" or "B?", "Is it positive or negative?", "Is it a customer or not a customer?", or "Is it a cat or not a cat?", etc. Logistic equations are used by SL programs to answer classification questions, and one specific equation is the sigmoid function that is used for binary classification to

predict two discrete classes. An example of a binary classification problem would be the task of identifying a cat versus non-cat in an input image. The goal is to train the SL model to predict whether the output is either 1 if it is a cat image and 0 if it is a non-cat image. The input, an image, is transformed into pixel data represented by the red, green and blue color channel values of the image. And this input data is in matrix format to be analyzed by the program. Unlike linear regression functions which yield numerical outputs that identify the relationship between variables and are continuous in value, logistic regression functions yield discrete classes or categorical values of 0 or 1. One common example of a logistic regression function is the sigmoid function (as shown in Figure 4 from Ng et al., (2023)). The sigmoid function produces an S-shaped curve that maps any number into a numerical value between 0 and 1 and does so without reaching those exact limits (Ng et al., 2023; Theobald, 2017).

The sigmoid function is of the form: $g(x) = 1 / (1 + e^{-x})$ where x is the scalar numerical value of inputs and e is Euler's constant, 2.718. And a logistic regression sigmoid function in matrix format that analyzes all of the input data would be $g(X|W) = 1 / (1 + e^{-(WX + W_0)})$

Figure 3:

Linear Regression Function

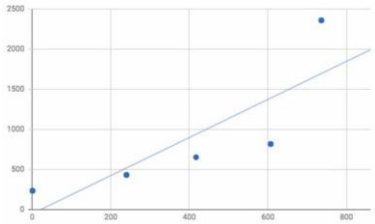
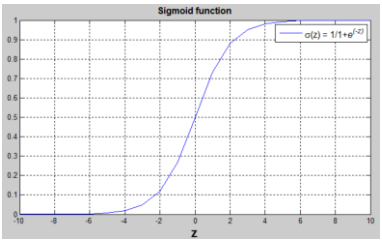


Figure 4:

Sigmoid Function



Similar to linear regression, the goal of logistic regression is to minimize the error between its predictions and training data. Once the model has been created and parameters have been identified for each variable, optimization will be undertaken. To optimize the parameters, the errors have to be minimized. The loss function $\mathscr{L}(Y, g(X|W)$ computes the difference between the desired output (Y) and

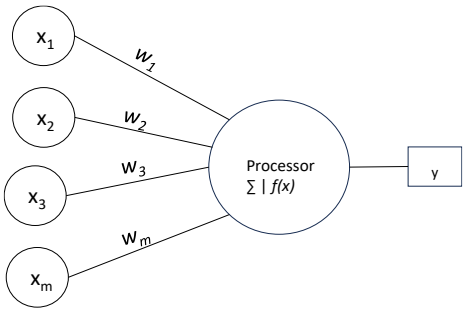
the model's approximation value, g(X|W). And the optimization procedure uses methods to minimize the loss function $\mathscr{L}(*)$ (Ng et al., 2023; Theobald, 2017).

3.2.1 Artificial Neural Networks and Deep Learning

Artificial Neural Networks (ANNs) are a machine learning technique that processes data through layers of analysis (Theobald, 2017). ANNs take their inspiration from the human brain which consists of billions of interconnected nerve cells called neurons. In ANNs, each neuron receives data from other neurons, processes the inputs and passes the outputs to other neurons. A perceptron is the simplest kind of ANN (Mohammed et al., 2017). A perceptron consists of a single neuron that can receive multiple inputs and produces a single output – as illustrated in Figure 5, reproduced from Mohammed et al. (2017). Each connection has a numeric weight or algorithmic parameter (w_x) that can be modified in order to optimize the model to produce the desired output for a particular task. If the sum of all of the weighted inputs ($\sum = x_1*w_1 + x_2*w_2 + x_3*w_3 + \ldots x_m*w_m$) satisfies a threshold, which is the activation function (f), it will produce an output. However, if the sum does not meet the set threshold, the neuron will not activate.

Figure 5:

Perceptron

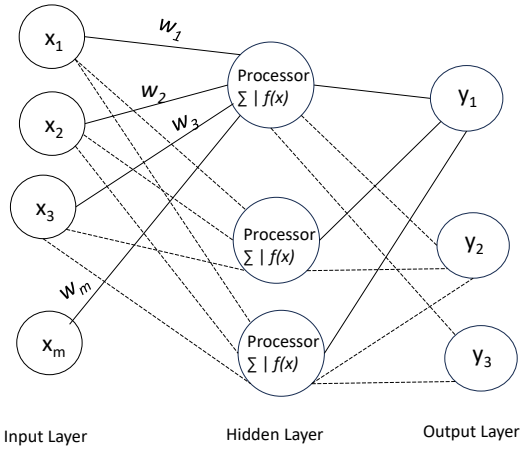


Some of the various activation functions that can be used with the perceptron include the step, sign, linear and sigmoid functions. The step function is of the form: $f(x) = 1$ if $x>0$ and $f(x)=0$ if $x<0$ The sign function is of the form: $f(x) = 1$ if $x>0$ and $f(x)=-1$ if $x<0$ The linear function is of the form $f(x) = x$ The sigmoid function is of the form: $f(x) = 1 / (1 + e^{-x})$ (Mohammed et al., 2017; Theobald, 2017).

Note that an ANN consists of multiple layers of multiple perceptrons that are interconnected in multiple ways. Figure 6, reproduced from Theobald (2017), illustrates a basic ANN with one input layer, one output layer and one hidden layer.

Figure 6:

Layers of an ANN



In an ANN, the neurons are stacked up in layers and start with a broad base with the first layer consisting of raw input data (e.g. numeric values, text, images, sound, etc.) which are divided and fed into neurons. Each neuron then sends information to the next layer of neurons which analyze and process the data. The hidden layer analyses, processes and streamlines the data until the final result is produced in the output layer.

Using SL techniques, an ANN is trained by comparing the model's predicted output to the known correct output and measuring the difference between the two. This is the loss (cost or error), and the purpose of training is to reduce the loss until the model's prediction closely matches the correct output. This is achieved by incrementally adjusting the weights of the network until the lowest possible loss is obtained. This process of training is performed by an algorithm called backpropagation, and it adjusts the weights which computes gradients from the output layer back to the input layer.

The above describes basic ANNs which are adequate for simple data processing. However, with more input data and more complex patterns to analyze, there will be a requirement to create more

advanced ANNs which will include linking various types of neurons with different types of activation functions (e.g. sigmoid, linear, non-linear, step, etc.) and a higher number of hidden layers to manage more input neurons. An ANN with more than 3-8 neuronal layers is considered deep learning (DL) (Mohammed et al., 2017; Theobald, 2017).

The middle layers of an ANN are considered hidden because like some human processing activities, such as the facial recognition example discussed above, they covertly process the data in ways that are not easily interpretable. Our human brain is conscious of the inputs and outputs but not always of the nuanced deterministic details of how we process and transform the information between the two. Essentially, ANNs tend to operate as black boxes because tracing the structure and examining the network's weights and connections reveals little to no insight into how specific inputs lead to specific outputs. Unlike the human capacity to introspect, query, analyze, and communicate our thoughts and actions, black box AI systems do not possess such capabilities and mechanisms to explain their internal reasoning processes.

3.3 Explainable AI (xAI) Landscape

As outlined above, there are many different types of AI applications and models depending on the task. Also, depending upon the context, we humans participate in many different types of explanations. Since the purpose of xAI is to provide explanations about AI systems to humans, both factors will drive the need for many different types of xAI solutions. In this section, I will provide an overview of the current xAI landscape that encompasses the many different explanatory models available.

3.3.1 Overview of xAI Landscape

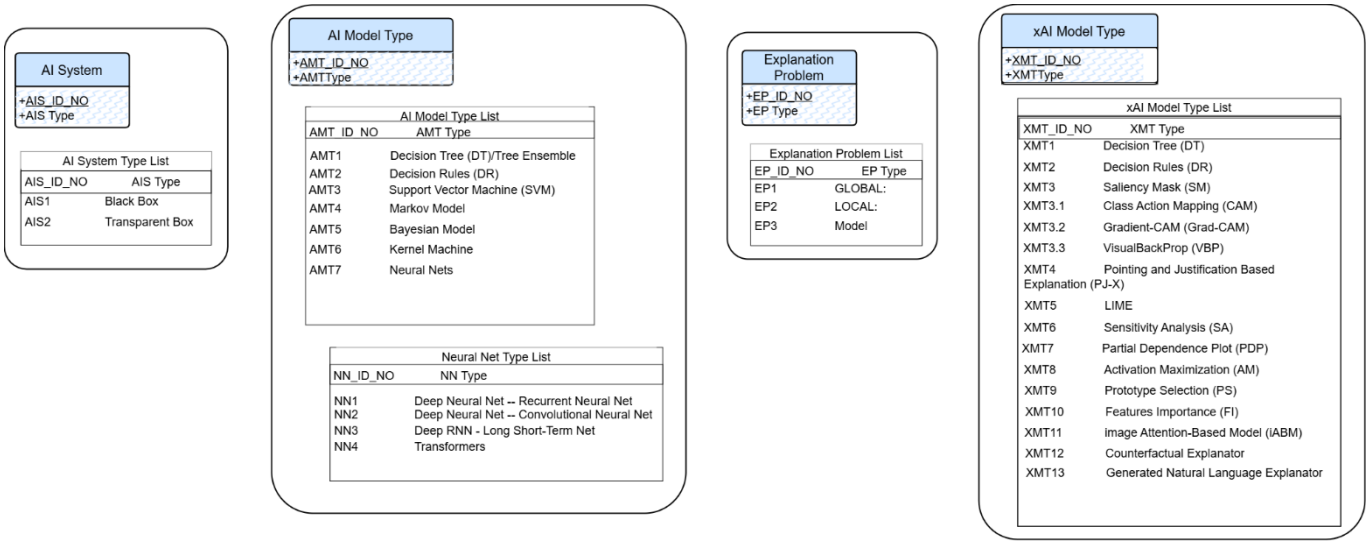
In order to provide an overview of current state-of-the-art xAI approaches, I have proposed the following classification of xAI as outlined in Figure 7's Faceted Taxonomy of the xAI Landscape. Appendix B contains the original versions of the taxonomy which were amended after Phase 3's proof-

of-concept testing of the FGCR process (Chapter 5). This classification of the xAI landscape is based upon Adadi and Berrada's (2018), Burkart and Huber's (2021), and especially Guidotti et al.'s (2018) surveys of xAI methods.

Several points to note with Figure 7: It is populated with the types of xAI models applicable to ANNs in particular as they are the focus of this research project. Therefore, it is not exhaustive in either breadth or depth and is intended as a starting point. It can be modified and added to in the future by xAI developers as required. For example, other types of AI models would expand this taxonomy depending upon the AI models that are the target for explanations. Furthermore, I have chosen to present the taxonomy of the xAI landscape in faceted format because, as shown in Appendix B, a hierarchical taxonomy was unmanageable and unwieldy and therefore, not efficient or user-friendly. For example, when other types of AI models were added to the list, the size (width and length) increased, and the taxonomy simply became badly organized. Also, since the various classifications within the xAI landscape can be structured and ordered in a variety of combinations, it makes a predetermined hierarchical taxonomy with frozen levels of classification unwieldy.

Figure 7:

Faceted Taxonomy of the xAI Landscape (Focus on Artificial Neural Networks)



The category of AI system includes black box and transparent box. A black box AI system is opaque, and therefore, requires a post-hoc xAI surrogate model where interpretability is added after the AI model has been developed. A post-hoc xAI surrogate model is a simple model that explains a complex black box model. In contrast, a transparent box AI system is an ante-hoc intrinsically interpretable "white box" model where the interpretability is built in from the beginning of the model creation and training. An interpretable-by-design transparent box is an ML model that does not require additional interpretability mechanisms because they are created with the intention of improving the interpretability for humans. Examples of transparent box ML techniques include decision trees, decision rules, and linear models (Burkart & Huber, 2021).

The category of AI Model assists with the identification of the AI system that requires an explanation. It involves classifying AI models or ML algorithms and comprises of a list of the types of AI models and lists of their corresponding sub-types. For example, types of black boxes include artificial neural networks (ANN), support vector machines (SVM), tree ensembles (TE), etc. And types of transparent boxes include trees (DT), decision rules (DR), linear models, etc. Each of these can be further subdivided such as neural net subtypes of DNNs, RNNs, etc.

The category of the explanation problem assists with the identification of the types of explanation problems and includes three choices: Model Explanation, Outcome Explanation or Model Inspection. Model explanation problems require the xAI model to provide a global explanation of the whole workings of the underlying AI black box model. The xAI model should be interpretable and transparent. It should be able to mimic the behavior of the black box and also be understandable by humans (Guidotti et al., 2018). Outcome explanation problems require the xAI to provide an explanation for a specific outcome, decision, or single prediction of the underlying AI black box. They do not require an explanation of the whole workings of the underlying the black box but only the reason for the prediction on a specific instance of interest. They are therefore, considered as providing a local explanation (Adadi & Berrada, 2018; Guidotti et al., 2018). Model inspection problems require the xAI

model to provide a representation (visual or textual) for some specific property of the underlying black box model or of its predictions, decisions or outcomes. Examples of properties of interest could include sensitivity to changes in features and identification of components of black boxes such as neurons in ANNs that are responsible for specific outcomes (Guidotti et al., 2018).

The category of xAI Model assists with the identification of the type of xAI model. Since there are many different types of xAI models, I have chosen to highlight only a few. The choices of explanators (xAI models) include, but are not limited to: Decision Tree (DT), Decision Rule (DR), Feature Importance (FI), Saliency Mask (SM), Class Activation Mapping (CAM), Sensitivity Analysis (SA), Partial Dependence Plot (PDP), Activation Maximization (AM), Prototype Selection (PS), counterfactual explanator, Pointing and Justification-based Explanation (PJ-X) model, image and Attention-Based (iABM) model, and Local Interpretable Model-agnostic Explanations (LIME). (See Figure 7)

A decision tree (DT) is a hierarchical data structure that implements the divide-and-conquer strategy. It is a tree shaped structure that is composed of internal decision nodes and terminal leaves. Each decision node represents a test on features or attributes of the training data with discrete outcomes that are branches. Each subsequent node represents another test with more branches and ensuing nodes. The process begins at the root and ends at a leaf node representing a class label. The paths from the root to the leaves represent the classification rules. A decision tree can be linearized into a set of decision rules with if-then form (Alpaydin, 2014; Guidotti et al., 2018). In addition to being a transparent box ML technique, DTs can assist with solving all three explanation problems (model explanation, model inspection, and outcome explanation) for the underlying AI black box model. For NNs, model explanation solutions include explanation via single-tree (DT) which involve using several different ML techniques to query a NN and induce a decision tree that represents the most relevant features of the NN in an understandable and comprehensible format (Guidotti et al., 2018).

A decision rule (DR) is a function that maps an observation to an appropriate action. In contrast to decision trees, decision rules have a textual representation. The most common rules are "if-then" rules where the "if" clause is a combination of conditions on the input variables and can be formed by conjunctions, negations and disjunctions and the "then" clause specifies the outcome (Guidotti et al., 2018). Similar to DTs above, DRs serve as transparent box ML techniques as well as assist with solving all three explanation problems listed above for the underlying AI black box. For neural nets, model explanation solutions include explanation via rule extraction (DR) which involve using several different techniques to extract rules from NNs include xAI models that insert knowledge into NNs, extract rules from trained NN and use NNs to refine existing rules (Guidotti et al., 2018).

Feature importance (FI) explanators address both outcome explanation and model explanation problems for the underlying AI black box. An FI explanator quantifies the weight and magnitude of the input features used to make the predictions of a complex black box ML model (Adadi & Berrada, 2018; Guidotti et al., 2018). To quantify the importance of a complex ML model's feature, the increase of the ML model's prediction error is calculated after permuting the feature. Model error is increased by important features value permutations and model error is unchanged for unimportant features value permutations (Adadi & Berrada, 2018).

A saliency mask (SM) explanator is typically used for treating images or texts to point out what causes a certain outcome. SMs address both outcome explanation and model inspection problems for the underlying AI black box. SMs identify the saliency of the individual features of a complex ML model. An SM uses "masks" to visually highlight the determining aspects of the record analyzed. It is generally used to explain deep ANN's and can be viewed as a visual representation of FI (Guidotti et al., 2018). A sub-class of saliency mask explanators involves an approach that incorporates network activations into their visualizations and are named Class Activation Mapping (CAM). A CAM local explanation for a particular outcome highlights the discriminative active region of an image. A generalization of a CAM is a Grad-CAM which visualizes the linear combination of a late layer's

activations and label-specific weights or gradients. It provides visual explanations from deep networks via gradient-based localization for DNNs and analyzes image data type (Guidotti et al., 2018).

Another visualization tool is a VisualBackProp (VBP) explanator which addresses both an outcome explanation and model inspection problem for image classification models such as convoluted neural networks (CNNs). It provides a visual representation of the sets of pixels of the input image that contribute most to the output. The VBP method assume that the feature maps contain less and less irrelevant information to the output when moving deeper into the CNN. It was initially developed as a debugging tool for CNN-based systems for steering self-driving cars (Bojarski et al., 2017). It is an invaluable inspection tool that can be easily used for both training and inference aspect of model development (Guidotti et al., 2018).

A partial dependence plot (PDP) explanator is a graphical representation that helps in visualizing and understanding the relationship between one or more input variables and the outcome of a black box (Adadi & Berrada, 2018; Guidotti et al., 2018). PDPs address outcome explanation and model inspection problems for the underlying AI black box (Guidotti et al., 2018).

A sensitivity analysis (SA) model addresses model inspection problems for the underlying AI black box. An SA model evaluates the influences of uncertainties of inputs and/or weight perturbations on a black box's outcome. It is used for explanations related to ANN and DNN classification of images. SA produces an explanation of the variation in the function value rather than the function value itself. Its purpose is to develop tools for model inspection by finding and removing unimportant input attributes (Adadi & Berrada, 2018; Guidotti et al., 2018).

An activation maximization (AM) explanator addresses model inspection problems for the underlying AI black box. An AM explanator carries out inspections of ANNs and DNNs by observing which fundamental neurons are activated with respect to particular inputs. AMs look for input patterns that maximize the activation of a certain neuron in a certain layer (Guidotti et al., 2018).

A prototype selection (PS) explanator addresses outcome explanation problems for the underlying AI black box. A PS explanator provides an example very similar to the predicted output along with the predicted output. A prototype is representative of a set of similar instances or observed points in data or summarizes a subset of these with similar characteristics (Guidotti et al., 2018).

Counterfactual explanators address outcome explanation problems for the underlying black box. These explanators provide descriptions of the minimum conditions that would have led to an alternative decision. They focus on a single output and provide a contrasting adversarial example where the emphasis is on reversing the output or prediction rather than providing an explanation of the full logic of the overall algorithm (Adadi & Berrada, 2018).

Attention-based model explanators highlight the most promising parts of input features that lead to a certain output for a given task. They can summarize textual and/or visual image data. An image attention-based model (iABM) provides a description of the content of an image by using image interpretation techniques to illustrate how the description of the image was created. (Burkart & Huber, 2021; Adadi & Berrada, 2018) A multi-modal explanator is a pointing and justification-based explanation (PJ-X) for visual decision tasks. It provides explanations for an image by simultaneously highlighting the relevant parts and generating a textual justification (Burkart and Huber, 2021). These models address outcome explanation problems for the underlying black boxes.

Also note that xAI models can be classified as being either model-specific or model-agnostic. Model-specific xAI models are applicable only for a single type or class of AI algorithm and are therefore limited. By nature, transparent box intrinsically interpretable AI systems are examples of model-specific xAI models. Model-agnostic models are xAI models that can be used to any type of ML algorithms, in that they separate the prediction task from the explanatory task. Model-agnostic xAI models are usually post-hoc, can be either local or global and are generally used for ANN explanation tasks (Adadi & Berrada, 2018).

A local interpretable model-agnostic explanations (LIME) explanator addresses the outcome explanation problem for the underlying black box. A LIME model provides an explanation of predicted outputs by approximating the outputs near the output of interest. It selects representative instances near the prediction of interest with explanations that are textual or visual artifacts such as words in a text or patches in an image.

Because the focus of this paper is ANNs, Figure 7 is populated with the types of xAI models applicable to ANN's in particular. In order to provide a global model explanation for an ANN, DTs and DRs are the choice of xAI models. In the case of outcome explanation problems, which require local explanations, SMs, PSs, PJ-Xs, iABMs, counterfactual explanators, and agnostic models such as LIME are the choice of xAI models. In the case of model inspection problems, FIs, SAs, SMs, Grad-CAMs, VBPs, PDPs, and AMs are of interest.

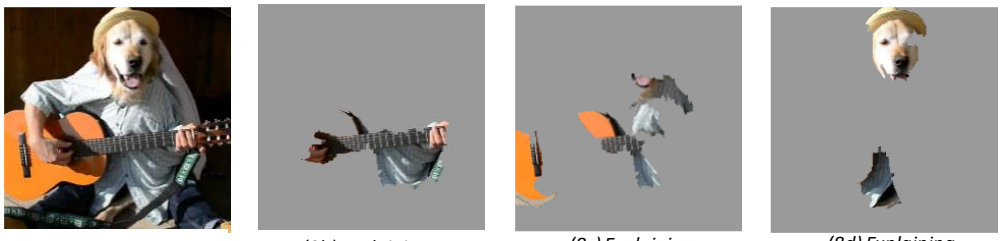
To exemplify the details of how an xAI model calculates its outputs and illustrate how these xAI models can be applied in practice, a detailed description of Ribeiro et al.'s (2016) Local Interpretable Model-agnostic Explanations (LIME) model is explicated below. As stated above, LIME is an agnostic post-hoc surrogate model that is used for ANN local explanations. LIME is a model that approximates a black box model locally in the neighbourhood of a prediction of interest and provides an interpretable and faithful explanation. It is an explanation technique that provides an explanation of the predictions of any classifier or regressor by approximating the outputs near the output of interest. A LIME model selects a set of representative instances near the prediction of interest with explanations that are textual or visual artifacts such as words in a text or patches in an image. LIME provides a qualitative understanding of the relationships between components of the instances and the prediction of a black box model (Ribeiro et al., 2016). An example of LIME's capabilities involves understanding the decisions of image classification systems. LIME finds regions of an image that are particularly influential to the final classification of an image. These are also called sensitivity maps, saliency maps or pixel attribution maps which are approaches that use gradient calculations to assign an importance

value to individual pixels. LIME is an ML technique and follows a similar general notation and model development process as outlined in section 3.2. The LIME model is denoted as "g" and can be presented to the user in the form of visual or textual artifacts. The AI model being explained, in our example, an ANN, is denoted as "*f*". The original prediction of an instance to be explained is denoted as "*x*" and "*x'*" denotes its interpretable representation. "$\mathscr{L}$" denotes the measure of how unfaithful g is in approximating *f* in the locality of "*x*". The objective is to minimize $\mathscr{L}$ in order to ensure interpretability and local fidelity while minimizing locality-aware loss without making any assumptions about *f* to ensure that the model is agnostic. $\mathscr{L}$ is approximated by drawing sample instances around *x'* by drawing nonzero elements of *x'* uniformly at random. And instances are sampled both in the vicinity of *x* and far away from *x*.

In the case of obtaining local explanations for image classifiers LIME highlights super-pixels that have a positive weight towards a specific class in order to provide information about the presence of that specific class. Figures 8a-8d from Ribeiro et al. (2016) demonstrate how an image classification prediction can be explained using LIME which highlights super-pixels.

Figure 8:

Explaining an Image Classification Prediction made by Google's Inception Neural Network



(8a) Original Image *(8b) Explaining Electric guitar* *(8c) Explaining Acoustic guitar* *(8d) Explaining Labrador*

The original image of a dog with a guitar (Figure 8a) is analyzed. Figure 8b highlights the pixels that explain the classification for electric guitar. Especially note that it highlights the fretboard to differentiate between an acoustic and electric guitar. Figure 8c highlights the pixels that explain the

classification for acoustic guitar. And Figure 8d highlights the pixels that explain the classification Labrador.

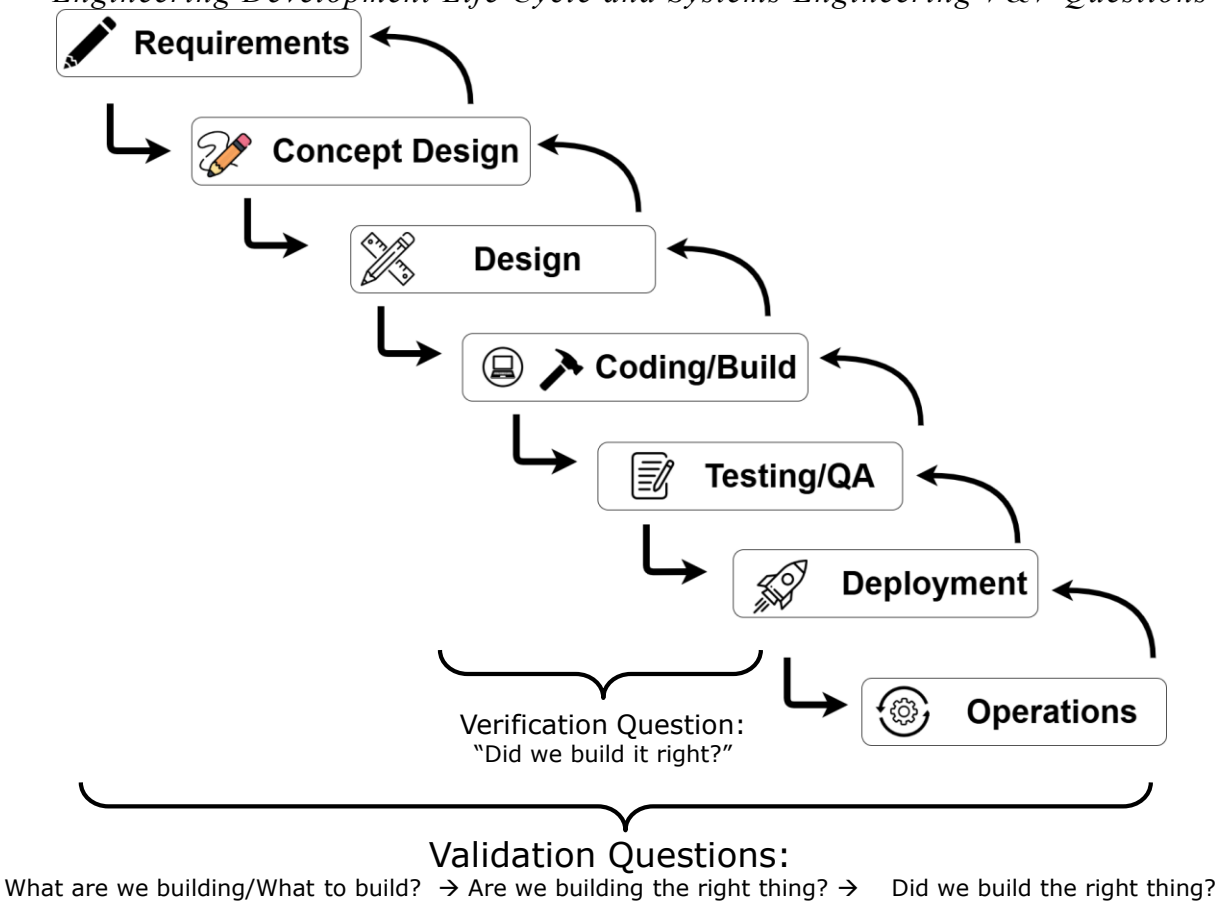
3.4 Engineering Development Cycles and Requirements Management Practices

To develop a suitable xAI solution, xAI developers must identify the contextual requirements that human-centric xAI solutions need to meet. The disciplines of engineering and project management have established detailed guidelines, tools, and techniques to support requirements management activities for the development of products. In this section, I will provide an overview of the existing state-of-the-art standards and practices in engineering and project management related to stakeholder requirements management.

As illustrated below in Figure 9, product development activities follow a typical engineering life cycle which begins with requirements definition and proceeds through concept, preliminary design, detailed design, build, test, certify and product in-service support phases. In addition to the waterfall-style linear model with iterative loops depiction of the lifecycle, note that superimposed are the key product systems engineering verification and validation (V&V) questions which will be discussed in detail in section 3.4.2 below. A product's success is directly influenced by the needs of the stakeholders. Stakeholders are entities (individuals or organizations) who have a legitimate interest in, may affect, be affected by, or perceive themselves to be affected by a decision, activity or outcome of a product or system. Typical stakeholders of AI systems include end-users; AI experts, engineers, regulatory bodies, operators of systems, customer organizational decision-makers, society at large, etc. (INCOSE, 2023, Section 2.3.5.2; Project Management Institute, 2021, Section 2.2.1; Bourque & Fairley, 2014). Diverse stakeholder consultations are a crucial and necessary step for ensuring that xAI models are built for all stakeholders, especially end-users, and not just for xAI developers. Stakeholder requirements definition and analysis processes facilitate the development of robust xAI solutions.

Figure 9

Engineering Development Life Cycle and Systems Engineering V&V Questions



3.4.1 Stakeholder Requirements Definition and Analysis Processes

The aim of stakeholder requirements definition process is to elicit, negotiate, determine, document, and manage stakeholder needs and requirements. The aim of the stakeholder requirements analysis process is to review, assess, prioritize, and balance all stakeholder requirements and to transform those requirements into a specification. A specification provides a functional and technical view of a system description of the features, functions, and characteristics of a product that meets the stakeholders' needs. The activities involved in these two processes include (but are not limited to) identification of stakeholders; elicitation of requirements; establishment of critical and desired system performance, measures of effectiveness and suitability; definition of functional boundary and performance, interfaces, design constraints; verification and validation of stakeholder requirements

throughout the product life cycle; etc. (INCOSE, 2023, Sections 2.3.5.2, 2.3.5.3, and 2.3.5.11; Bourque & Fairley, 2014, Section 1, pp.1-1 – 1-14; Project Management Institute, 2021, Section 5.2)

3.4.2 INCOSE's Product Verification and Validation (V&V) Processes

The International Council on Systems Engineering (INCOSE) outlined detailed trans-disciplinary principles and practices for product development (INCOSE, 2023). Of particular importance to this research project are INCOSE's principles and definitions of product verification and validation processes. Product verification and validation (V&V) activities are conducted within the engineering product development cycle. Although similar in nature, V&V objectives are fundamentally different. The verification process proves that an end product conforms to its requirements and ensures that the product correctly solves the problem. Verification answers the critical question "Did we build it right?" In contrast, validation answers the critical question "Are we building the right thing for relevant stakeholders?" The validation process provides confidence that the product is able to accomplish its intended use, goals, and objectives; meets stakeholder requirements; and solves the right problem in the intended environment. Recursive and iterative V&V activities are crucial for a robust engineering life cycle (Hirshorn et al., 2017; INCOSE, 2023, Section 2.3.5.11; SEBOK, 2024; Bourque & Fairley, 2014). Figure 9 illustrates the typical target points for V&V questions within the product development lifecycle.

A basic illustration of the validation process that answers the question "Are we building the right product?" is the case of designing a chair. Beginning with the definition of a chair, the engineering process identifies parameters, characteristics, and functional requirements such as "has a seat; is supported by leg-like structures; is a height of x cm above the ground; is something to sit on; etc." The engineering design activities could yield various products that meet these requirements, such as a plain wooden chair with a seat mounted on four legs; an upholstered recliner; a rocking chair; a wheelchair; an adjustable office chair; Ikea's Poang cantilevered chair. An example of the validation process applied to the robotics field would be a pick and place robot. The requirements would include the ability to locate,

lift and retrieve an object from one location, carry it to and placing it in another location. The engineering design activities could and have yielded various robot designs such as Amazon's flat pods that have wheels and slide under stacked shelf units or Boston Dynamics' versatile robot Stretch with a mobile base, robotic arm, and adaptive gripper or a simple stationary device with robotic arms located at one workstation (NOVA PBS, 2018; Boston Dynamics, 2022). As demonstrated, a simple definition of a chair or a pick and place robot results in a multiplicity of designs and because of this, recursive and iterative validation activities are crucial for ensuring that the right product is built.

Within a robust human-centric engineering lifecycle, validation cannot be treated as a single, static milestone deferred to the end of development. Imposed directly upon the engineering lifecycle structure illustrated in Figure 9 above, are the formal systems engineering verification and validation (V&V) definitions established by INCOSE (2023). While verification strictly checks whether the system was "built right", validation addresses a broader, more critical question, "Was the right system built?" For robust product development, validation cannot be treated as a single, isolated event left for the backend of the lifecycle. Instead, validation must occur continuously throughout the entire lifecycle dynamically shifting its focus to evaluate the product at its current stage of maturity. This continuous validation process is operationalized by asking the specific validation questions shown at the bottom of Figure 9 as the development cycle progresses. In the early upstream phases, during stakeholder requirements definition, the validation inquiry focuses on answering the question, "What is the right thing to build?" Progressing through the design and build phases, the validation inquiry focuses on the question, "Are we building the right thing?". And culminating with the testing/QA and deployment phases with answering the question "Did we build the right thing?" By systematically asking these progressive validation questions, developers ensure that the evolving end-product remains firmly tethered to human-centric requirements from inception to execution.

It should be noted that while the linear (with iterative loops) model structured in Figure 9 utilizes a traditional waterfall-style depiction of the systems engineering lifecycle, this representation is not

exclusive. Modern software and systems engineering frequently leverage alternative product development lifecycles (e.g. Agile, Scrum, etc.), to accommodate rapid iteration and continuous deployment. However, regardless of the overarching execution framework, the foundational logic of mapping progressive validation inquiries onto the development path is applicable across different lifecycle models. This mapping grants developers the flexibility to consult or bypass stakeholder feedback loops as appropriate for their specific operational velocity. Rather than requiring constant, disruptive stakeholder consultation at every micro-iteration, the framework establishes a critical baseline: developers must, at a minimum, engage stakeholders upstream to answer the question, "What is the right thing to build?" during the specification phase, and downstream during testing and deployment to definitively answer the question, "Did we build the right thing?"

Ultimately, because the focus of this research project is concerned with determining a suitable xAI model to develop, and in particular with identifying what xAI model to develop, robust product development activities consistent with INCOSE's (2023) validation mandate of determining whether the right thing is being built for the relevant stakeholder are crucial. And therefore, diverse and structured stakeholder consultations are a necessary and foundational part of this process.

3.4.3 The Emerging Intersection of Requirements Engineering and xAI

The state-of-the-art in requirements engineering management outlined in sections 3.4.1 and 3.4.2 above reflects the benchmark standard practices established by INCOSE (2023) and Bourque and Fairley (2014). These are the foundations upon which the FGCR (detailed below in Chapter 4) is constructed and against which it should be assessed. It is worth noting, however, that a parallel and younger body of research has emerged that specifically addresses the intersection of Requirements Engineering (RE) and Explainable AI (xAI). This work applies RE concepts directly to the problem of making AI systems explainable.

The maturity and trajectory of this intersection is reflected in the growth of the International Workshop on Artificial Intelligence and Requirements Engineering (AIRE), now in its twelfth year as of

2025, co-located annually with the IEEE International Requirements Engineering Conference (Ghavanati et al., 2025). The workshop's continued relevance confirms both the importance and the continued open-endedness of the challenges involved in applying RE discipline to AI development. Notably, within the broader IEEE RE Conference ecosystem, two contributions stand out: Köhl et al.'s (2019) treatment of explainability as a non-functional requirement, and Habiba et al.'s (2022) proposal of structured elicitation steps that are explicitly oriented around the question of explainability for whom and to what degree.

While these contributions are meaningful steps toward incorporating RE discipline into xAI development, they have not yet achieved the operational maturity and institutional consolidation represented by INCOSE (2023) and Bourque and Fairley (2014). Importantly for this research project, they do not yet provide the fully operationalised, stakeholder-centred, context-sensitive requirements process that the FGCR is designed to deliver. A detailed critical analysis of these shortcomings — and a direct assessment of how the FGCR addresses them — is provided in Chapter 5, sections 5.4.4 and 5.4.5, where empirical data from the proof-of-concept testing provides the grounding for that evaluation.

The purpose of this brief note is not to dismiss the emerging RE-for-xAI literature, but to situate it appropriately: as a developing field that confirms the relevance of the problem this thesis addresses, while the established engineering standards of INCOSE and SWEBOK remain the appropriate benchmarks for the FGCR's design and evaluation.

4 Phase 2 – Framework for Generating Context-Dependent Requirements for xAI Model Development

Phase 2 of this research project involved the creation of the proposed theory – a Framework for Generating Context-Dependent Requirements (FGCR) for human-centric xAI model development. The proposed FGCR will aid xAI developers with identifying the contextual requirements that human-centric xAI solutions need to meet. It serves as a guide comprising of a structured process and a suite of supporting tools and aids that can enable xAI developers to determine and identify a suitable xAI solution.

In this chapter, I will present the components of the FGCR: Section 4.1 will describe the FGCR process itself, while section 4.2 will provide details of the supporting tools including the taxonomies. Section 4.3 will describe the structural evolution of the FGCR from its initial foundational architecture to its final refined version presented in sections 4.1 and 4.2.

The FGCR is designed to facilitate xAI developers with the requirements management activities involved in generating context-driven requirements that feed into the requirements specification for development of human-centric xAI models. The FGCR is intended to complement, augment, and assist with the execution of the tasks associated with two key processes within xAI development cycles: the product validation process and the stakeholder requirements definition and analysis processes. The purpose of the product validation process as specified by INCOSE's (2023) *Systems Engineering Handbook* is to provide evidence that "the 'right' product has been or is being made according to the stakeholder needs and requirements" (INCOSE, 2023, Section 2.3.5.11). And the stakeholder requirements definition and analysis processes involve the elicitation, definition and validation of the stakeholder needs and transforming them into product (xAI model) requirements (INCOSE, 2023; Project Management Institute, 2021; Bourque & Fairley, 2014).

The aim of the proposed FGCR process is to enhance the engineering stakeholder requirements definition and product validation practices established in INCOSE's (2023) *Systems Engineering*

Handbook and Bourque and Fairley's (2014) *Guide to the Software Engineering Body of Knowledge (SWEBOK)* by extending their application to the emerging field of xAI. I am providing an approach and complementary strategies that build upon the existing engineering practices and frameworks rather than replace them, thereby enriching the overall understanding and application of these essential resources to the development of human-centric xAI solutions.

In order to capture the needs of *explainees* into a set of requirements that feed into the development cycle of human-centric xAI models, the FGCR structure is composed of a combination of new and existing knowledge: (1) A new process that outlines iterative activities to progressively elaborate the needs of the *explainees* based upon context (section 4.1's Figure 10). (2) New aids in the form of taxonomies that serve as methods of classifying existing knowledge of SSH-based explanatory classes and current landscape of xAI explanatory models (see sections 3.1 and 3.3 for figures of taxonomies and section 4.2.1). (3) Existing stakeholder requirements definition and analysis tools and techniques from the fields of engineering and project management (section 4.2.2).

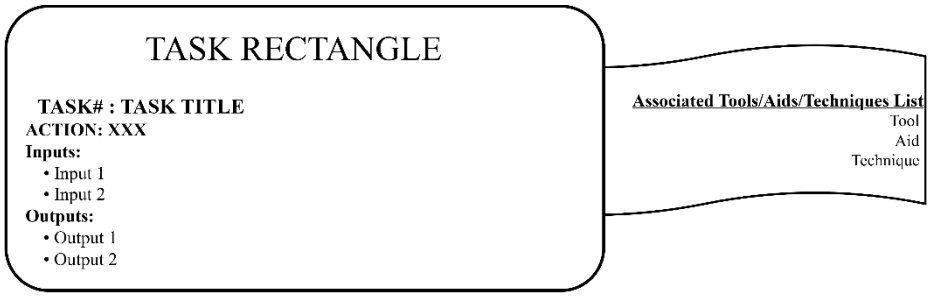
4.1 FGCR Process

The FGCR process involves iterative tasks that progressively elaborate the needs of the *explainees* based upon context and capture them into a set of requirements that feed into the xAI development cycle. Processes are sets of interrelated tasks that are performed to create a result or produce an output (International Organization for Standardization [ISO], 2015). A process transforms inputs to outputs through a series of tasks that apply tools, aids, and techniques. These tasks, can be sequential and/or iterative, require specific inputs, such as data, materials, or information, and utilize tools and techniques to achieve the desired outcome. By combining these elements, a process generates a valuable output, which can be a product, service, or result.

As outlined in section 2.2.2, I developed the FGCR's process by extracting and synthesizing descriptive requirements management knowledge from the fields of engineering and project

management (INCOSE, 2023; Project Management Institute, 2021; Bourque & Fairley, 2014) into a process map.

The FGCR process is illustrated below in Figure 10. The process details step-by-step tasks, inputs and outputs of those tasks as well as various tools, techniques and resources such as the proposed taxonomies. I devised a notation where tasks are displayed in rectangular boxes, and the suggested tools, aids and techniques that can be used by xAI developers for those tasks are displayed in squiggly boxes alongside them:



As mapped out in Figure 10, the key tasks of the FGCR process to be performed include:

Task#1: Identification and Recruitment of Stakeholders/*Explainees*

- The main activities of this task are:
  i) Brainstorm to identify target stakeholders requiring explanations
  ii) Recruit target stakeholders
- The inputs to this task are members of multi-disciplinary xAI development teams that will serve as *explainers*.
- The output of this task is a list of stakeholders/*explainees*.

Task#2: Stakeholders/*Explainee* Consultation for Identification of *Explanans*

- The main activities of this task are:
  i) Conduct interaction between *explainer(s)* and *explainee(s)*/stakeholder(s) to pose questions along the lines of "What type of explanatory information are you seeking a) for the output of this AI system; and/or b) about this AI system(s) with respect to this situation?
  ii) Identify the AI application, model and actions performed by the AI.
- The inputs to this task are:
  - the list of stakeholders/*explainees*
  - *explainers* from the membership of multi-disciplinary xAI teams.
- The outputs of this task are:
  - List of questions asked by *explainees* that represent the *explanans* (explanation) sought
  - Classification of the explanatory information sought by *explainees* (using the SSH-based taxonomy)

  - Groupings of explanatory information by AI function or actions/decisions performed
  - Identification of AI application, AI model and/or AI function that requires explanation (using the taxonomy of the xAI landscape)

Task#3: Generation of Detailed xAI Model Requirements

- The main activity of this task is to generate xAI model requirements that would satisfy the explanatory needs of stakeholders
- The inputs to this task are:
  - Members of multi-disciplinary xAI teams
  - Explanatory information sought by explainees/stakeholders as per the outputs of Task#2 above
- The output of this task is a detailed summary of the inputs as a set of xAI requirements

Task#4: Identification of Potential xAI Models that Could Satisfy the Explanatory Needs of Stakeholders/*Explainees*

- The main activity of this task is to use the requirements (outputs of Task#3), to identify existing xAI model(s) that could provide explanations
- The inputs to this task are the outputs of tasks #2 and #3
- The outputs to this task are:
  - Determination of the type of explanation problem per the taxonomy of the xAI Landscape
  - List of potential xAI models that could provide explanations

Task#5: Usage of Existing xAI Models to Answer Stakeholders/*Explainees'* Questions via Consultation

- The main activity of this task is to conduct stakeholder consultation in order to provide explanatory information from existing xAI model to stakeholder. This task is focused on validation of explanations by consulting with stakeholders per Task#2 above.
- The inputs to this task are potential xAI models that could provide explanations from the output of Task#4 above.
- The output is a satisfactory/not satisfactory determination based on response from stakeholders/explainees about the adequacy of explanatory information provided to them.

If Task#4 or Task#5 do not meet the needs of stakeholders/*explainees*, then

Task#6: Generation of Detailed xAI Model Requirements for Inclusion in a Specification to Modify Existing or Create New xAI Model(s).

- The main activity of this task is to generate xAI model requirements that would satisfy explanatory needs of stakeholders
- The inputs include xAI model requirements based on feedback from stakeholders/*explainees* including:
  - List of explanations that can serve as requirements for xAI developers
  - List of AI applications, models, and/or functions performed that require explanations
- The outputs are xAI model requirements for inclusion in the development specification.

Note that several of these steps require iterative activities that involving consultations with stakeholders/*explainees*.

To ensure that stakeholder explanatory needs directly drive the technical development of explainable systems, the FGCR maps directly onto the systems engineering life cycle, establishing a critical conceptual link between the framework's operational tasks and traditional engineering development lifecycles. As illustrated in Figure 11 (Focus of FGCR's Tasks in Engineering Development Lifecycle), the FGCR process (detailed in Figure 10) interfaces explicitly with the standard systems engineering development lifecycle originally outlined in Figure 9 (Section 3.4). By design, the execution of the FGCR occurs primarily 'upstream' at the front end of the engineering development cycle. Rather than treating explainability as an afterthought, Tasks 1 through 5 of the FGCR function during the initial stakeholder requirements elicitation, definition, and analysis phases. This front-end concentration ensures that human-centric explanatory needs are codified into formal technical specifications before architectural choices or algorithmic designs are locked in and the system is built. However, while the core synthesis of requirements is front-loaded upstream, validation is treated as a continuous, iterative thread throughout the entire life cycle. The framework transitions 'downstream' to the backend of the engineering development cycle following the execution of Task 6, which generates a new or updated xAI model. Once the xAI model(s) and/or solution(s) are produced, a critical backend validation phase is triggered to ensure that the operationalized xAI model satisfies the context-dependent criteria and user-specific requirements elicited during the front-end phase. This dual-focus architecture establishes a closed-loop engineering cycle, anchoring subjective human-centric needs firmly into the xAI development process.

Figure 10:

Framework for Generating Context-Dependent Requirements (FGCR) Process

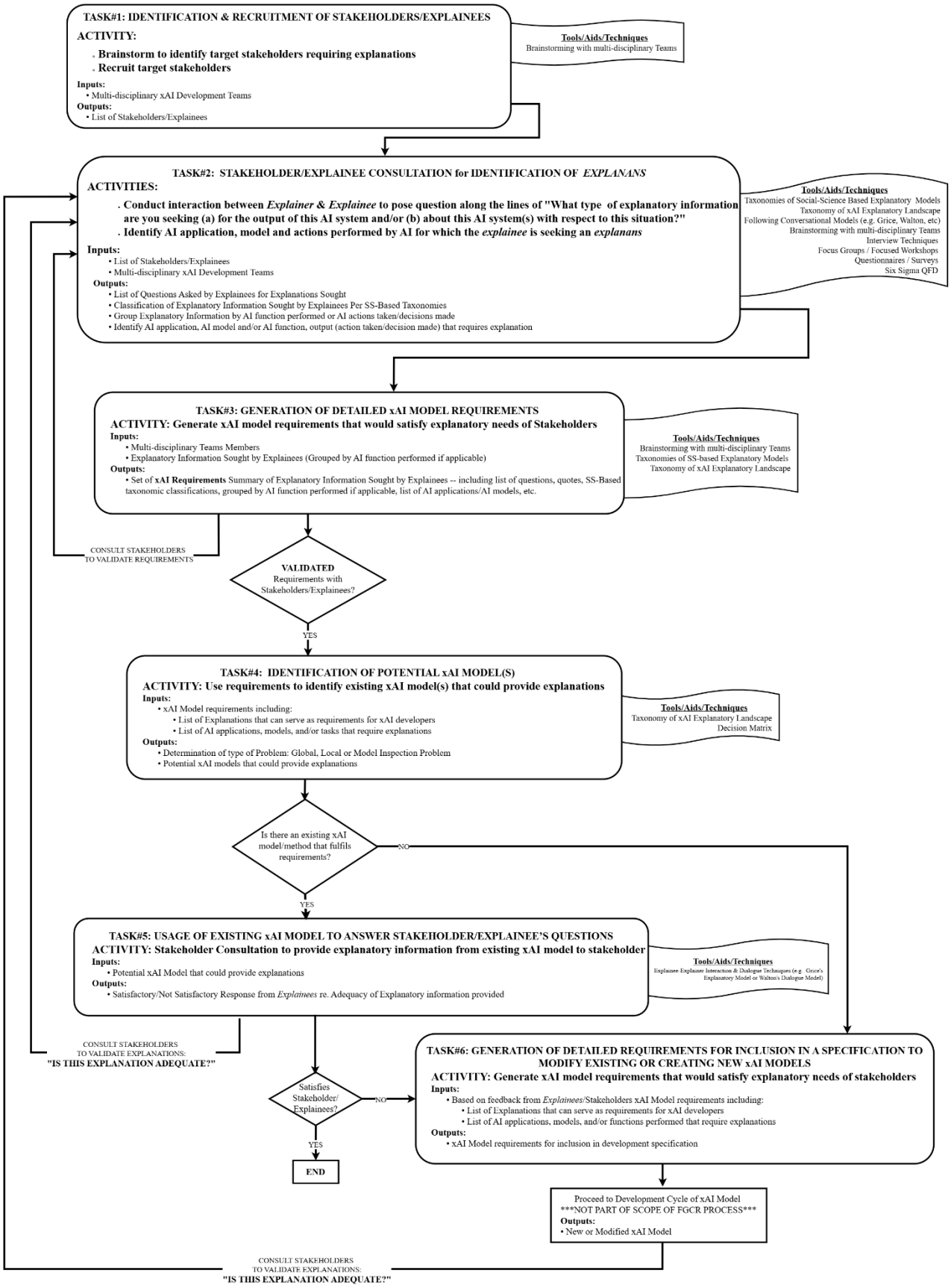
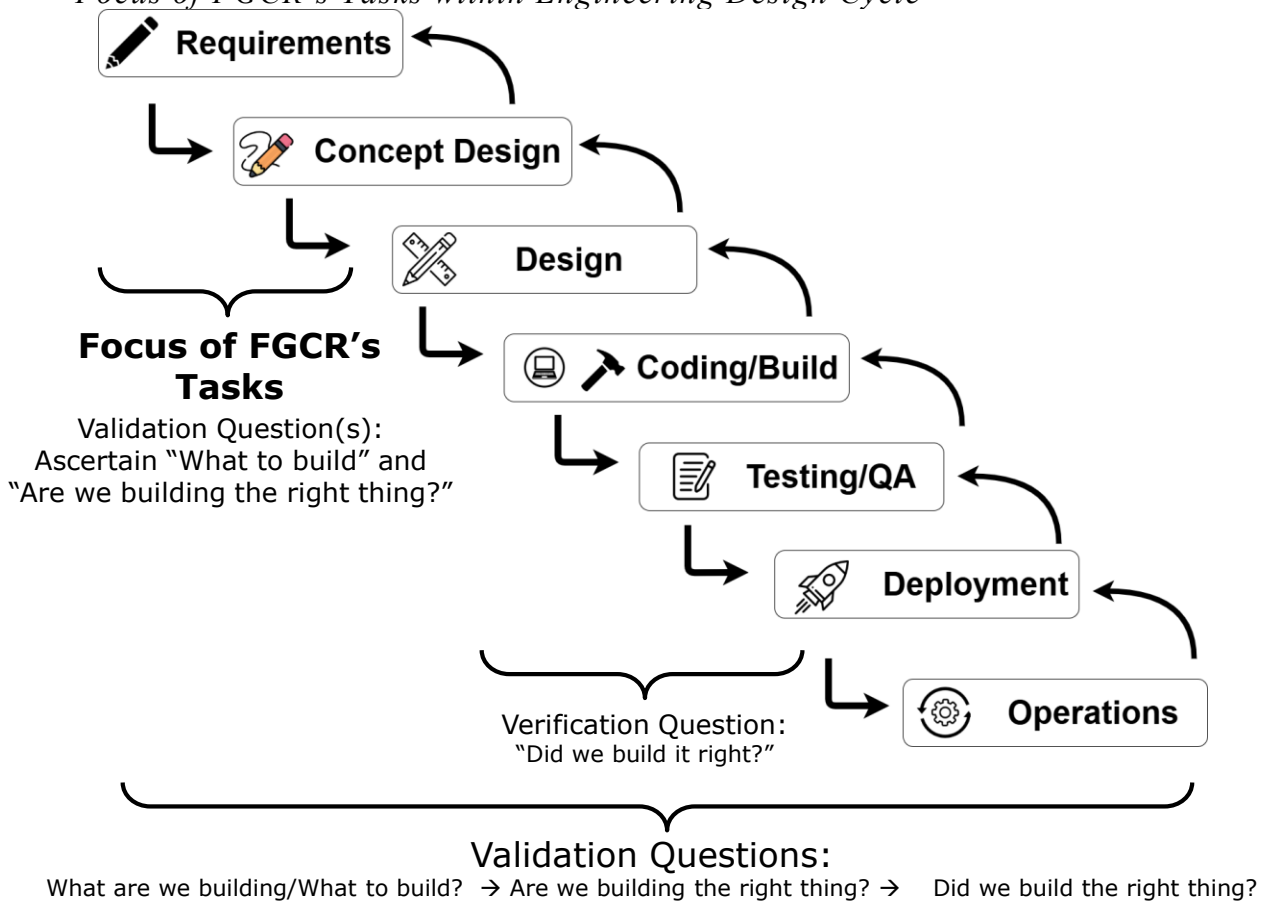


Figure 11

Focus of FGCR's Tasks within Engineering Design Cycle



The FGCR process and associated tools, aids and techniques in Figure 10 represent a refined version of the original detailed in Appendix B. I used the original FGCR process during proof-of-concept testing of the FGCR to gather stakeholder needs and generate detailed xAI model requirements. Based on the results of this testing, as discussed in Chapter 5, I fine-tuned and optimized the components of the FGCR. Appendix B documents the original FGCR components, alongside details of the amendments made as a result of Phase 3's proof-of-concept testing. And section 4.3 provides a summary of the structural evolution of the FGCR from a research design perspective.

4.2 FGCR Tools and Aids

Tools and aids of the FGCR include the following: aids such as the aforementioned taxonomies – the SSH-based Explanatory (section 3.1) and xAI Landscape taxonomies (section 3.3); and various stakeholder requirements definition and analysis tools (section 4.2.2) such as interview techniques, Six Sigma QFD, etc. Note that these tools and aids can be used by xAI developers as they see fit and are listed in Figure 10's squiggly boxes alongside the rectangular task boxes.

4.2.1 Aids (Taxonomies) for Identification of Types of SSH-Based Explanation and xAI Models

Because of the large body of knowledge within the fields of AI, xAI and the SSHs, I chose to amalgamate the varied and complex information into taxonomic forms. When faced with the problem of providing descriptions of complex inter-related things of interest, the challenge is organizing, structuring, and communicating the information for multiple parties in as efficient, easy to interpret, and useful way. Taxonomies assist with addressing this problem by serving as aids for classifying and communicating complex, inter-related information.

The taxonomies I developed will assist xAI developers with the tasks of efficiently identifying the types of SSH-based and current xAI landscape of explanatory models: SSH-Based Hierarchical and Faceted Taxonomies of Explanation (outlined in section 3.1's Figures 1, 2a and 2b) and Faceted Taxonomy of xAI Landscape (outlined in section 3.3's Figure 7). The detailed definitions and descriptions for the categories and sub-categories are provided in Chapter 3.

Two classification approaches are used across these taxonomies: hierarchical and faceted. A hierarchical classification approach takes a single predetermined hierarchical order and structure. In contrast, a faceted classification approach allows for the use of multiple classification criteria and enables greater freedom in organizing information. It can use a hierarchical structure in one or more of its facets while allowing for a combination of facets to broaden coverage and accommodate more varied and complex structural approaches in which information may need to be organized (Usman et al., 2017).

For SSH-based explanation classes, I developed both hierarchical and faceted taxonomic structures (section 3.1's Figures 1, 2a and 2b) so that xAI developers may choose whichever classification approach they prefer. For the taxonomy of the xAI Landscape (section 3.3's Figure 7), I chose to produce only a faceted taxonomy. As shown in Appendix B, the hierarchical version proved unmanageable. As additional AI model types were added, the taxonomy became unwieldy in both width and length and poorly organized. The faceted structure's flexibility to combine and reorder classifications makes it far better suited for this purpose. It should also be noted that this taxonomy is populated with the xAI model types applicable to ANN's specifically, as these are the focus of this research project. Therefore, it is not exhaustive in either breadth or depth and is intended as a starting point. It can be modified and added to in the future by xAI developers as required. For example, incorporating other types of AI models would expand its scope depending upon the AI models targeted for explanations.

Both taxonomies represent a refined and optimized version of the originals detailed in Appendix B. After using the original taxonomies during Phase 3's proof-of-concept testing of the FGCR, I was able to finetune and improve them. Appendix B contains the original versions alongside a description of the amendments made as a result of the testing. Furthermore, note that both taxonomies (SSH-based Taxonomy of Explanation and xAI landscape Taxonomy) are purposefully scoped and provisional. They are designed to evolve as xAI developers build upon them in response to the advancing state of the knowledge in both fields.

4.2.2 Stakeholder Requirements Definition and Analysis Process Tools

The purpose of stakeholder requirements definition and analysis processes is to elicit, identify, and prioritize stakeholder needs, and then to refine and balance these needs and requirements, ultimately translating them into a formal specification. In this section, I am providing an overview of the existing state of the art codified engineering and project management standards and practices pertaining stakeholder requirements definition and analysis tools.

Some of the stakeholder requirements definition and analysis tools that I am recommending to xAI developers include but are not limited to interviews, focus groups, questionnaires, and surveys, facilitated workshops, brainstorming, nominal group techniques, idea/mind mapping; multi-criteria decision analysis; group decision-making techniques; decision matrices, requirements classification, and Six Sigma Quality Function Deployment. (Project Management Institute, 2021, Section 5.2; Bourque & Fairley, 2014, Sections 3-4)

A crucial part of the stakeholder requirements definition and analysis process involves the activities of eliciting or deriving stakeholder needs and requirements. Because requirements are rarely ready-made, xAI engineers must actively elicit information from stakeholders about their needs. These needs, once identified, can then be translated and distilled into requirements that feed into the specifications for xAI models. There are multiple elicitation techniques and tools available to facilitate the identification and elicitation of the needs of stakeholders including interviews; focus groups; facilitated workshops; brainstorming; questionnaires and surveys (Project Management Institute, 2021 Section 5.2, Bourque & Fairley, 2014, Section 3.2).

Interviews can take a formal or informal approach to eliciting requirements information from stakeholders by talking to them directly. Interviews are performed by asking prepared and/or spontaneous questions and recording the responses. They can be conducted one-on-one or involve multiple parties of interviewers and interviewees. Focus groups involve a trained moderator who guides a group of stakeholders and subject matter experts through an interactive discussion about expectations and attitudes about a proposed product. Questionnaires and surveys involve the usage of written sets of questions that are designed to quickly accumulate information from a large set of varied audiences who are geographically dispersed. Facilitated workshops are interactive sessions that bring key stakeholders together to quickly define cross-functional product requirements and reconcile stakeholder differences. Brainstorming is a technique used by individual or groups of multiple stakeholders to generate and collect multiple ideas.

Once stakeholder requirements are defined, they need to be analyzed and transformed into a specification that will be used throughout the xAI model development cycle. The analysis of requirements involves detecting and resolving conflicts between requirements; conducting trade-offs; defining the scope of the model being developed; and elaborating top-level requirements into detailed requirements. There are multiple techniques for requirements analysis including nominal group techniques; idea/mind mapping; group decision-making techniques; decision matrix; requirements classification techniques such as affinity diagrams; requirements traceability matrix; and Six Sigma Quality Function Deployment which can also function as a requirements definition tool.

Nominal group techniques enhance brainstorming with a voting process used to rank the most useful ideas for further analysis. Idea/minding mapping techniques enhance brainstorming by adding a voting process that ranks the most useful ideas for prioritization. Group decision making techniques are methods to reach a group decision which are used to generate, classify, and prioritize product requirements including the methods of unanimity, majority, plurality, and/or dictatorship. A decision matrix is a tool that is composed of rows and columns that form a matrix containing data pertaining to key factors of different choices. It is a systematic analysis of evaluating and selecting the best option between different choices. A requirements traceability matrix is a grid that links product requirements from their origin to the deliverables that satisfy them and provides a means to track requirements throughout the development cycle and provides a structure for managing scope. Requirements classification techniques such as affinity diagrams allow large numbers of ideas to be classified into groups for review and analysis. Requirements can be classified on a number of dimensions including whether a requirement is functional or nonfunctional; is derived from higher-level requirements or an emergent property; is on the product or process; is volatile or stable as in whether it changes during the life cycle. Another dimension for classification involves the priority and scope of the requirement.

Six Sigma Quality Function Deployment (QFD) is a technique that falls under both the requirements definition and analysis processes. It takes a structured approach for determining and

defining critical stakeholder requirements and translating them into characteristics, attributes and features of a product that are documented in a detailed specification. Interviews, focus groups and/or facilitated workshops can use QFD techniques to elicit key product requirements and result in more robust specifications.

Note that this is not an exhaustive list of requirements definition and analysis tools and techniques and I have provided these as suggestions and options for xAI developers to use depending on their preferences.

As will be outlined below in Chapter 5, for the proof-of-concept testing of the FGCR, I used the elicitation technique of interviews for requirements elicitation and definition. And then I used classification techniques involving the new taxonomies for requirements analysis in order to transform them into xAI model requirements.

4.3 Summary of Structural Evolution of the FGCR

To fully appreciate the refinement of the FGCR presented in this chapter, it is necessary to contextualize its development against its original foundational architecture. The framework did not emerge as a rigid, static construct; rather, it underwent an expected, iterative design cycle where original baseline mechanisms were tracked, challenged, and refined through a proof-of-concept demonstration. A comprehensive repository containing the full visual maps, structural parameters, and technical schemas of these initial baselines is preserved in Appendix B (Original FGCR Components). This section provides an extended abstract and analytical synthesis of that material to map the evolutionary trajectory of the framework from a research design perspective.

The original iteration of the FGCR process (detailed explicitly in Figure 12, Appendix B) was structured around an initial baseline methodology for translating human explanatory needs into xAI solutions. In this initial deployment version of the FGCR, the taxonomies served as the primary structuring mechanisms to help organize stakeholder inquiries into discrete, bounded Social Sciences

and Humanities (SSH)-based categories. These early taxonomies included the *Original SSH-Based Hierarchical Taxonomy* (Figure 13, Appendix B), the *Original SSH-Based Faceted Taxonomy of Explanation—The 'WHAT' and 'HOW'* (Figures 14 and 15, Appendix B), and the initial structural classifications of the computational landscape via the *Original Faceted and Hierarchical Taxonomies of the xAI Landscape* (Figures 16 and 17, Appendix B).

Consistent with the underlying research design established in Chapter 2, these baseline structures were always intended to be evolutionary. As stated in Chapter 2 and demonstrated through the empirical deployment in Chapter 5, the core purpose of a proof-of-concept demonstration is to explicitly test theoretical boundaries and make necessary modifications to the theory as empirical reality dictates. Real-world stakeholder interactions highlighted that human explanations are fundamentally fluid and context-variable, rarely conforming perfectly to rigid, academic, hierarchical classifications.

From a research design perspective, this evolutionary shift directly operationalizes Maxwell's (2012) Interactive Model of Qualitative Research Design introduced in Chapter 2. Rather than viewing this structural modification as a linear correction or a failure of the initial model, the refinement of the taxonomies exemplifies Maxwell's philosophy that research design components—namely goals, conceptual frameworks, research questions, methods, and validity—exist in a non-linear, interconnected, and dialectic relationship. The empirical data collected via user interactions (methods) actively informed and reshaped the utility of the initial taxonomies and FGCR process (conceptual framework), triggering an interactive re-alignment across the system to preserve the descriptive and interpretive integrity of the final framework.

Consistent with the scope outlined in Chapter 6, this iterative refinement is not leveraged to make a definitive or absolute validation claim regarding the framework's output. Instead, this evolutionary process serves as a practical demonstration of the framework's feasibility and functionality. It proves that the FGCR can dynamically adapt from the rigid categorization structures preserved in Appendix B to the highly fluid, user-centric flexible instruments presented in sections 4.1 and 4.2. The

vital lesson extracted from tracking this evolution is that xAI developers must treat these taxonomies not as absolute constraints, but as flexible aids. Maintaining the record of the original baselines in Appendix B serves a dual purpose: it provides a clear engineering audit trail of the research design methodology and offers a comparative benchmark that justifies the flexible, human-centric design choices embedded within the finalized FGCR.

5 Phase 3 – Proof-of-Concept Testing of the Framework

Phase 3 of this research project involved proof-of-concept testing of the proposed theory, a Framework for Generating Context-Dependent Requirements (FGCR) for xAI development. As outlined in Chapter 4 above, the proposed FGCR serves as a guide comprising of a structured process and a suite of supporting tools and aids that is designed to facilitate the identification of a contextually suitable xAI solution. It can assist xAI developers with executing requirements management activities involved in generating context-driven requirements that feed into the requirements specification for development of human-centric xAI models. As mentioned in section 2.2.3 above, the scope of testing was limited to proof-of-concept testing which enables an assessment of a theory's feasibility and functionality before proceeding to its full development. (INCOSE, 2023; SEBok, 2024).

In this chapter, I will present the results of the proof-of-concept testing of the FGCR using real-world stakeholder data. This chapter is organized to first provide an overview of the stages and scope of proof-of-concept testing and in order to situate the detailed presentation of the results, a summary of four different contextual configurations that were used to interrogate and present the findings of the testing. The subsequent sections of this chapter will provide details of the data collection and analysis protocol (section 5.1); an in-depth analysis of the results of testing (sections 5.2-5.3); and will conclude with a discussion of the findings and their implications (section 5.4).

I assessed the FGCR's feasibility and functionality in two sequential stages. During the first stage, I gathered real world data about the explanatory needs of stakeholders pertaining to the outputs of an AI system. To do so, I employed the qualitative research design of inquiry involving case studies research and used interviews as the primary data collection method. As outlined in section 2.2.3, I chose the AI domain of AVs, more specifically, a case study that explored a real-world scenario of a series of car crashes involving AVs. This allowed me to determine whether, when subjected to real-world data, the FGCR's components can effectively generate specific xAI model requirements for different

stakeholders who would seek varied explanations. The testing also provided valuable insights that supported my efforts to finetune, optimize and enhance the FGCR's design features. The final version of the FGCR is presented in Chapter 4 above and the original version is presented in Appendix B.

Due to the lack of access to real-world proprietary AI models involved in the driving tasks performed by the AVs that were discussed in the case study scenario, the scope of proof-of-concept testing was limited to the first four tasks described in Chapter 4's Figure 10. Accordingly, a summary the results of proof-of-concept testing of the following FGCR tasks will be presented in this chapter:

- FGCR Task#1: Identification of Stakeholders/*Explainees*
- FGCR Task#2: Stakeholder/*Explainee* Consultation for Identification of *Explanans*
- FGCR Task#3: Generation of Detailed xAI Model Requirements
- FGCR Task#4: Identification of Potential xAI Models that Could Provide Explanation(s)

During stage 1, the data gathering activities inherent to this case study coincided with the first two tasks of the FGCR. Specifically, the standard process of identifying and consulting stakeholders within the case study provided the necessary data for Task #1, Identification and Recruitment of Stakeholders/*Explainees*, and Task #2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*. This enabled me to collect real-world data that was then used in stage 2 to assess whether the FGCR can generate specific xAI model requirements. The activities in stage 1 of the testing enabled me to ascertain whether:

(a) FGCR's Task#2 can be employed to identify, and classify the specific types of explanations sought out by explainees by using the taxonomies of SSH-based explanation classes
(b) FGCR's Task#2 can be employed to identify the AI application, or AI models that require an explanation by using the taxonomy of the xAI landscape, or, if specific information about the AI model is not available, it can identify the actions taken and/or decisions made by the AI system

During stage 2, to assess the feasibility and functionality of the FGCR process and components, I used the analyzed data from stage 1 to first ascertain whether I could use the FGCR to generate xAI requirements for the case study and then whether I could use it to effectively support the identification and specification of suitable xAI solutions. To evaluate the theory, the FGCR, I used it to ascertain whether:

(c) FGCR's Task#3 can be employed by the xAI developers to generate detailed xAI model requirements based on the outputs of (a) and (b)
(d) FGCR's Task#4 can be employed to determine and identify if there exist xAI models that can meet the explanatory needs of the different stakeholders by using the taxonomy of the xAI landscape

The remaining two FGCR tasks could not be evaluated because the lack of access to actual AI models prevented the use of existing xAI models to generate explanations that could be validated by stakeholders. Therefore, Task#5, Use Existing xAI Model to Answer Stakeholder Questions, could not be evaluated. In addition, within the limited scope of this research project, the outputs of Task#3 provided a preliminary assessment of proof-of-concept of the FGCR, thereby rendering the more comprehensive Task#6, Generation of Detailed xAI Model Requirements for Inclusion in a Specification to Modify Existing or Create New xAI Model(s), beyond the current research project's defined scope and resources. Therefore, the testing of FGCR Tasks #5 and #6 remain the purview of future research efforts within the scope of real-world acceptance testing as discussed in section 5.4 below.

In order to assess and evaluate the feasibility and versatility of the FGCR while illustrating the pivotal role that contextuality plays in explanations, I have chosen to interrogate and present the findings of the testing through multiple contextual perspectives. To assess the FGCR's ability to handle varied contextual configurations effectively, I subjected it to contextual conditions that are composed of multiple combinations of factors including multiple AI models performing multiple AI functions and multiple *explainees* seeking explanations for the same phenomenon (a series of car crashes).

I will discuss four different contextual configurations of the analysed data. These configurations assessed the workability of the FGCR, namely, whether it can serve as a guide for identifying a contextually appropriate xAI solution regardless of AI domain, AI application, AI model, actions performed by AI systems and *explainee(s)* seeking an explanation. The configurations also illustrate the pivotal role that contextuality not only plays in explanations but also drives the need for a multiplicity of explanations even for the same phenomenon which in turn drives the need for tailoring xAI solutions. I used the following contextual configurations to evaluate the functionality and versatility of the FGCR:

In order to probe the need for tailoring xAI solutions to the stakeholders impacted by AI systems, the first two analyses approached the data from the vantage point of *explainee(s)*.

i. A disaggregated analysis of one individual stakeholder (Stakeholder#1 in sections 5.2.3, 5.3.1-5.3.2).
   - This involves the contextual configuration of the same phenomenon that needs an explanation, the same *explainee* with the added factors of the diverse and unique perspective of the *explainee*, and the complexity of driving a vehicle involving multiple AI actions such as steering, braking, etc.
   - This analysis will evaluate whether the results of this contextual configuration support the claim in Chapter 1 that the same *explainee* will seek multiple explanations for the same phenomenon.

ii. A disaggregated analysis of another individual stakeholder (Stakeholder#25 in sections 5.2.4, 5.3.3-5.3.4).
   - This involves the contextual configuration as listed above in (i) along with the added factor of a separate and different *explainee*.
   - This analysis will evaluate whether the results of this contextual configuration support the claims in Chapter 1 that not only will the same *explainee* seek multiple explanations but also, when compared to another stakeholder's analysis, different *explainees* will seek different (and sometimes multiple) explanations for the same phenomenon.

In addition to the above, two more analyses were undertaken to illustrate that even when approaching the data from the vantage point of the AI's actions (or particular AI models which perform those actions) rather than the *explainees*, the need for tailoring xAI solutions still centers on the *explainees* (stakeholders) impacted by AI systems.

iii. An analysis of one AI action (perception in sections 5.2.5, 5.3.5-5.3.6).
   - This involves the contextual configuration of the same phenomenon that needs an explanation, multiple *explainees* and the same action performed by the AI.
   - This analysis will evaluate whether the results of this contextual configuration support the claim in Chapter 1 that even for the same AI action, due to diverse needs of *explainees,* different and multiple explanations are needed for the same phenomenon.

iv. An analysis of another AI action (system disengagement in sections 5.2.6, 5.3.7-5.3.8).
   - This involves the contextual configuration as outlined in (iii) above but for a different AI action.
   - This analysis will evaluate the ability of the FGCR to tackle the objective of identifying an appropriate xAI solution regardless of contextual combination of factors (different AI models performing different AI actions, different and multiple *explainees,* etc.).

These four analyzes tested the robustness and generalizability of the FGCR's ability to handle varied contextual configurations effectively and generate requirements for and/or identify a suitable xAI solution.

5.1 Stage 1 of FGCR Testing: Data Collection via Case Studies

Stage 1 of testing the FGCR, involved the first two FGCR tasks: FGCR Task#1, Identification and Recruitment of Stakeholders/*Explainees* and FGCR Task#2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*. In this section, I will provide details of the data collection and analysis protocol for stage 1 of testing the FGCR which includes the execution of FGCR Task#1, Identification and Recruitment of Stakeholders/*Explainees*.

I employed the qualitative research design of inquiry involving case studies and used participatory and empirical research methods such as interviews (section 2.2.3). Because most stakeholders of AI systems are not AI experts, the need for human-centric layperson termed explanations is great and engagement of diverse stakeholder consultations which include specific *explainees* that are the target audiences for explanations is crucial. Therefore, I conducted interviews with multiple stakeholders impacted by the target AI technology of AVs within the case study (outlined in below in section 5.1.1 and Appendix A).

I undertook and completed the following data collection and analysis activities:

- obtaining ethics approval prior to commencement of data collection
- recruiting of participants
- obtaining free and informed consent from participants
- conducting interviews
- transcribing interviews
- analyzing the transcribed text

Stage 1's activities involving the data collection and analysis protocol are detailed in section 5.1.1 and the results of executing FGCR Tasks #1 and #2 are presented in section 5.2. The outputs of stage 1 testing are:

- a summary of questions posed by stakeholders
- classification the explanatory information sought by stakeholders per the SSH-based taxonomy – this will be presented in the format of the data analysis template
- grouping of explanatory information by AI function, outputs as in decisions made/actions taken by the AI
- identification of AI application, AI model, or AI actions taken or decisions made by the AI system that requires an explanation

5.1.1 Data Collection & Analysis Protocol

5.1.1.1 Case Studies Research

In order to test the proposed FGCR, I used its components, a process and a set of resources, to identify the varied explanatory information sought out by different stakeholders who are impacted by a real-world AI application. I employed a single case study of a real-world AI application domain and elicited information through interviews. As outlined in section 2.2.3.1 above, I chose a case study that permitted a focused approach that enabled the rigorous testing of the FGCR while managing the scope of this thesis and generating data that could make significant research contributions to the field of xAI. I chose the AI application of AVs for three reasons: (i) AVs use multiple AI models that perform diverse sets of actions (such as motion planning, motion control, vehicle localization, perception, detection of pedestrians, traffic signs, obstacles, etc.), enabling and facilitating an assessment of the FGCR across different AI functionalities; (ii) AVs operate in the real world with high-stakes outcomes where safety of human lives is at risk (such as AVs involved in car crashes); and (iii) AVs are broadly deployed and have an impact on a broad spectrum of stakeholders throughout society who are from all walks of life and professions and are impacted daily by these things because automobiles are ubiquitous in our lives.

5.1.1.1.1 Case Study Scenario: AV Systems Crashing into First Responder Vehicles

As I described in full in Section 2.2.3.1, the case study scenario involves a series of crashes in which Tesla vehicles with Autopilot engaged struck stationary first responder vehicles that triggered a NHTSA investigation (EA22-002). My analysis focused on this a specific pattern: the occurrences of Tesla's Autopilot system deactivating less than a second before colliding with emergency vehicles. The interview techniques were used to query different stakeholders in order to collect data about specific explanatory information that they would seek regarding these real-world AI failures.

5.1.1.2 Data Collection

I began with the execution of the FGCR Task#1, Identification and Recruitment of Stakeholders/*Explainees*. Since automobiles are deployed broadly and have an impact on a wide range of stakeholders throughout society, I set out to elicit the opinions of diverse sets of stakeholders about the decisions, actions, and outputs of this new AI technology that operates in our world as an agent and is anticipated to replace the widely adopted existing technology of human-operated vehicles. Therefore, I undertook and completed the data collection and analysis activities per the guidelines and requirements of University of Ottawa's Research Ethics Board (detailed in section 2.2.3.1.2 and Appendix A).

The inclusion criteria and stakeholder categories applied in this study are detailed in section 2.2.3.1.2 and reproduced in full in Appendix A. In brief, participants were required to be adults (18 years of age or older) who are fluent English speakers and who represent a diverse cross-section of society affected by AV deployment — spanning categories including pedestrians, AV and non-AV drivers, police and first responders, legal experts, engineers, AI developers, and others. A screening questionnaire was used to verify each participant's stakeholder category, and inclusion was on a first-come, first-served basis within each category.

The recruitment target was ~20-35 participants with a sub-target of ~3-5 participants from as many of the categories of stakeholders identified above. To maximize diverse representation while reducing bias, this project required a balance in the number of participants in each category. For example, if I had been able to recruit 5 participants in the "pedestrians" category and only one in the "police" category, then the data would not have been balanced. Ultimately, a total number of 26 participants were recruited for the project. And within that total, I was able to recruit the following number of participants outlined per category:

- 5 engineers (non-AI expert)
- 3 AI developers
- 3 AV drivers
- 4 non-AV drivers
- 4 police officers

- 4 legal experts including 2 lawyers and 2 judges
- 3 others – 1 non-transport policy maker; 1 pedestrian and/or passenger; 1 insurance industry employee

Despite cold-calling and networking efforts, unfortunately, I was unable to recruit any lawmakers, government transport agency policy makers, crash investigators or automotive industry professionals. Acknowledging that I encountered limitations in recruiting these categories of stakeholders, the final sample, while not exhaustive, exhibited meaningful heterogeneity across stakeholder categories and provided a sufficiently diverse range of perspectives to enable me to address my research objectives. See Appendix A for copies of all recruitment material including the recruitment flyer and recruitment letters.

5.1.1.2.1 Interviews

After recruiting stakeholders, I conducted interviews to collect real-world data that would be used in stage 2 to assess whether the FGCR can generate specific xAI model requirements. Therefore, I executed the first part of Task#2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*, namely, interacting with the *explainees* to elicit the explanatory information they would seek about the case study scenario presented to them.

The full questionnaire, including the verbatim scenario text and all nudging prompts, is provided in Appendix A and described in detail in section 2.2.3.1.2.

For the reader's convenient reference, the verbatim scenario and primary question recited to participants are reproduced below:

> This case study scenario involves a real-life case within the AI application of automated vehicles (AVs). It involves the occurrences of actual car crashes involving one particular AV brand, Tesla and its advanced driver assistance system (ADAS), "Autopilot". Tesla's Autopilot system controls the steering, braking and acceleration functions of the AV without any assistance from the human driver. Furthermore, note that Autopilot could, at any time, disengage and hand over controls to the human driver.
>
> According to USA NHTSA's Office of Defects Investigation (ODI), EA22-002, between January 2018 and January 2022, Tesla's AVs, with Autopilot engaged, were involved in 16 crashes where they struck highly visible stationary in-road or roadside first responder vehicles (e.g., police, ambulance, fire trucks, road maintenance vehicles) that were attending to pre-existing

collision scenes. Furthermore, ". . . on average, in these crashes, Autopilot aborted vehicle control less than one second prior to the first impact".

The main question that guided the investigation was:

> Based on the scenario described above, you are seeking explanatory information about these car crashes from "Autopilot", the AI system that controls the steering, braking and acceleration functions of the AV (i.e., motion control functions).
>
> What specific questions would you ask of and/or what types of information would you seek from the AI system about the decisions made and actions taken by Autopilot (an agential entity) that is performing the human driving tasks of motion control of the AVs? i.e., controlling the steering, braking and acceleration functions of the AV.

I recorded every interview by audio or video. The transcripts were then generated automatically by the video conferencing applications MS Teams and Zoom. And in the case of audio recordings, the transcripts were generated by using the MS Word transcribe function. The output of all of these sources was transcribed text that once edited, became the final data set as described in further detail in the next section.

5.1.1.3 Data Analysis Process:

The first step in the data analysis process involved editing of the automatically generated transcripts to ensure accuracy of the content and confidentiality of the participants.

5.1.1.3.1 Manual Analysis of Transcribed Text

I manually analyzed each stakeholder's transcript by first identifying specific quotes pertaining to explanatory information sought out by them. I focused on phrases and sentences that contained amongst others, "why" and "how" type questions as well as "I want to know" types of statements. To ensure rigor, consistency and accuracy in my data analysis, I developed and applied the individual stakeholder data analysis template outlined in Table 4 below with drop-down menus for each category as listed in Table 5. Note that the results of the analyses of the individual stakeholders also enabled me to identify and extract aggregate data patterns for further analysis.

5.1.1.3.2 Template for Data Analysis:

To analyze the transcripts of all of the stakeholders, I created and used Table 4 Individual stakeholder data analysis template which has drop-down menus for each category as listed in Table 5 and is available in excel format in the Harvard Dataverse repository: https://doi.org/10.7910/DVN/WW5NZD. For each of the 26 stakeholder transcripts, I completed an individual analysis using this blank template form.

Table 4

Individual Stakeholder Data Analysis Table Template

STAKEHOLDER (SH# & Description):																							
ACTIONS Executed by AI Models/ Algorithms Category	THE WHAT OF EXPLANATION											THE HOW OF EXPLANATION											
	Quote(s)	P_ID_NO & TYPE	M_ID_NO & TYPE	BIIR_ID _NO & TYPE	BIUC_ID_ NO & TYPE	BG_ID_ NO & TYPE	Comments	Quote(s)	GSE_ID_NO & TYPE	CC_ID_ NO & TYPE	CCTF_ID_ NO & TYPE	CR_ID_ NO & TYPE	SL_ID_ NO & TYPE	EVF_ID_ NO & TYPE	Comments
									GSE1=Causal Chain Type						
									GSE2=Causal Reasoning Type						
									GSE3=Structures/Levels of Explanation Type						
									GSE4=Evaluation Factors Type						
									GSE5=Other						
									GSE1=Causal Chain Type						
									GSE2=Causal Reasoning Type						
									GSE3=Structures/Levels of Explanation Type						
									GSE4=Evaluation Factors Type						
									GSE5=Other						

Table 5:

Individual Stakeholder Data Analysis Table Drop Down Lists for Template

Interaction Techniques List	**Explanation Generation, Selection & Evaluation Techniques List**
IT_ID_NO & TYPE	**GSE_ID_NO & TYPE**
IT1 = Grice's Model	GSE1=Causal Chain Type
IT2 = Walton's Dialogue Model	GSE2=Causal Reasoning Type
IT3 = OTHER	GSE3=Structures/Levels of Explanation Type
	GSE4=Evaluation Factors Type
Phenomenon Types List	GSE5=Other
P_ID_NO & TYPE	
P1=Material Causes	**Causal Chain Types List**
P2=Behavioural Individual Intentional Reasons	**CC_ID_NO & TYPE**
P3=Behavioural Individual Unintentional Causes	CCT1=Temporal
P4=Behavioural Group Causes	CCT2=Coincidental
	CCT3=Unfolding
Material Causes List	CCT4=Opportunity
M_ID_NO & TYPE	CCT5=Pre-Emptive
M1= Device Failures	CCT6= Common-Cause Chain
M1.1 Device Fails to Adequately Perform Function	CCT7=Common-effect (Converging) Chain
M1.2 Device Fails to Perform Appropriate Action	CCT8=Linear Causal Chain
M1.3 Device Failure in Decision Making	CCT9=Causal Homeostasis Chain
M1.4 Device Fails to Engage/Disengage	CCT10=Other
M1.5 Device Communication Failure	
M1.6 Device Failure in Information Processing	
M1.7 Device Fails to Perform/Fulfil Primary Function	**Causal Chain Timeframe List**
M1.8 Other Device Failure	**CCTF_ID_NO & TYPE**
M2=Failure due to Design Flaw	CCTF1=Distal
M3=Lack of Resources	CCTF2=Proximal
M4=Lack of Time	CCTF3=Other
M5=Lack of Precautions	N/A
M6=Process Failures	
M7=Weather	**Causal Reasoning List**
M8=Acts of God, etc.	**CR_ID_NO & TYPE**
M9=Other	CR1=Abductive
N/A	CR2=Contrastive
	CR2.1=Contrastive Non-Occurrence of Effect
Behavioural Individual Intentional Reasons List	CR2.2=Contrastive The Normal Case
BIIR_ID_NO & TYPE	CR2.3=Contrastive Non-Common Effect
BIIR1=Beliefs	CR2.4=Contrastive Prescribed/Statutory Case
BIIR2= Desires	CR2.5=Contrastive Ideal Case
BIIR3=Intentions	CR2.6=Contrastive Designed/Specified/Expected Case
BIIR4=Norms	CR3=Counterfactual
BIIR5=Morals	CR4=Abnormality
BIIR6=Other	CR5=Associative
N/A	CR6=Interventionist
	CR7=Dennett's Process Narrative
Behavioural Individual Unitentional Causes List	CR8=Dennett's Intentional Reasoning
BIUC_ID_NO & TYPE	CR9=Mechanistic Causal
BIUC1=Emotional State	CR10=Other
BIUC2=Personality Type	N/A
BIUC3=Habit	
BIUC4=Subconscious Force	**Structure/Level of Explanation List**
BIUC5=Other	**SL_ID_NO & TYPE**
N/A	SL1=Aristotle
	SL2=Dennett
Behavioural Group Causes List	SL3=Marr
BG_ID_NO & TYPE	SL4=Expert vs. Layperson
BG1= Cultural Factors	SL5=Other
BG2=Social Factors	**Evaluation Factors List**
BG3=Other	**EVF_ID_NO & TYPE**
N/A	EVF1=Coherence
	EVF2=Truth
	EVF3=Probability
	EVF4=Goals of Explainee
	EVF5=Other

The main columns of the table included: “ACTIONS Executed by AI Models/Algorithms”; “Quotes”; “THE WHAT OF EXPLANATION”; and “THE HOW OF EXPLANATION”. The latter two columns were further delineated with sub-columns that represented the main categories per the SSH-based taxonomy and had drop-down menus in each cell for the user to easily perform the categorization task. And each of the main rows represented a particular action taken or decision made by the AI model and the associated stakeholder’s query which is to be quoted verbatim that pertains to that function.

To fill in the template, I extracted individual quotes of explanatory information sought out from the transcripts. I then identified the context of the questions and quotes in order to group the quotes into the major categories of actions performed by the AI model. The explanatory information sought by the stakeholders pertaining to each action and/or decision were then categorized by the two major groupings of the “what” and the “how” of explanation and their respective subset categories per the taxonomies of SSH-based explanations outlined in section 3.1 above.

Once that was completed for each individual stakeholder, in order to identify any patterns of explanatory information sought either by all stakeholders or various subsets of stakeholders, I carried out an aggregate analysis by action performed by the AI model.

5.1.1.3.3 Augmenting Manual Analysis of Transcribed Text with Use of Software:

As outlined in section 2.2.3.1.2 above and Appendix C, I augmented the manual analysis of transcribed text with the use of Google’s Gemini large language model (specifically Gemini 2.5 Pro Preview 05-06 gemini-2.5-pro-preview-05-06).

Initially, I anticipated that the LLM could be leveraged for qualitative data analysis expecting it to serve as a supplementary analyst which could perform cross-validation of my manual analysis and improve efficiency and accuracy. However, the implementation revealed a disparity between the theoretical promise of LLMs and their reliability in a research-intensive context. As detailed in the

performance evaluation in Appendix C, the LLM exhibited 'spotty" results, particularly regarding its ability to repetitively execute prescribed tasks with the necessary precision and accuracy.

Ultimately, I adopted a human-computer collaboration approach, using the LLM as a cognitive assistant to support my analysis. By utilizing the LLM as a cognitive assistant to rephrase results and offer distinct linguistic perspectives, the rigor and reliability of the manual analysis was enhanced. Notably, the LLM did not surface any significant new patterns or themes that had been overlooked during the manual analysis. Apart from the analytical limitations of the LLM, this outcome is consistent with the coverage of the SSH-based taxonomy of explanation proposed in this thesis, could serve as a corroboration of rather than a confirmation of the completeness of the current state of the taxonomy of SSH-based explanation, which, as stated in section 3.1 above, remains a working baseline subject to further refinement.

5.2 Results from Stage 1 of FGCR Testing

During stage 1 of testing the FGCR, I tested the functionality of the first two FGCR tasks: FGCR Task#1, Identification and Recruitment of Stakeholders/*Explainees* and FGCR Task#2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*.

The sections below that describe the results of proof-of-concept testing of FGCR Tasks #1 and #2 include:

- A summary of participants' stakeholder categories and key characteristics that represent their perspectives or viewpoints such as profession or road user status (section 5.2.1)
- An aggregate-analysis of stakeholder interviews categorized by actions executed by the AV's AI systems (section 5.2.2)
- Two disaggregated analyses for FGCR Task#2 of two individual stakeholders' (Stakeholder #1 and #25) interviews for multiple AI actions (sections 5.2.3 and 5.2.4)
- Two separate analyses of two different AI actions (perception and system disengagement) for FGCR Task#2 – note that each analysis will be disaggregated by AI action but aggregated by multiple stakeholders' interviews (sections 5.2.5 and 5.2.6). i.e. explanatory information sought by multiple stakeholders (aggregated) for two individual and separate AI actions (disaggregated).

Because the AI application of AVs is composed of multiple AI models performing different actions, I analysed the interview data by organizing stakeholder queries into groups of actions performed by these AI models. And these groupings enabled the execution of the next steps in the FGCR's process, namely, identifying the type of xAI model that could fulfill multiple explanatory needs of stakeholders. The groupings emerged from the patterns I observed about questions asked by the participants as they pertained to the types of actions taken and/or decisions made by the AI. The groupings included:

- Perception/Detection
  - Queries about failure to detect obstacles ahead
- Motion Control (Steering and Braking)
  - Queries about decision making & taking appropriate action failures such as not steering away or braking before the crash
- System Disengagement (Timing, Process)
  - Queries about reasons for system decision to disengage from autonomous mode seconds before the crash
- Communication/Warning to Human
  - Queries about reasons for not warning the human driver prior to the crash and/or prior to the decision to disengage

Furthermore, because the majority of stakeholders asked questions about the design(s) of the AI models, I identified a separate activity pertaining to the actions of the AI designers:

- Failure in design/training of the AI model
  - Queries about identifying what was overlooked in the design process and training as well as testing of the system prior to deployment in the field

It is worth noting that, among the outputs of stage 1 testing, specifically the identification of the AI application, AI model, or AI action(s) within the AI system requiring explanation, the analysis of the various FGCR tasks will reference only the AI actions. This is due to the lack of access to proprietary AI models involved in the case study scenario. Therefore, when it comes to assessing FGCR's Task#4 of determining and identifying potential xAI models, I will assume that all of the above functions are performed by one type of black box AI model – Neural Net. This will enable me to satisfy the main objective of testing the FGCR and demonstrating its feasibility while managing the scope of the thesis.

To facilitate the evaluation of the FGCR in stage 2's testing, I first analyzed the transcripts of all 26 stakeholders individually using Table 4's Individual Stakeholder Data Analysis Table Template. It is essential to emphasize that because the objective of this research project is to demonstrate and evaluate the feasibility of the FGCR process, the full analysis of all 26 interviews by each task performed by the multiple AI models operating the AV will not be presented in this thesis.

Instead, this full analysis of all data gathered is available for use and re-analysis by independent parties as part of future research efforts. All data collected and analyzed has been anonymized as outlined in section 2.2.3.1.2 above and will be available in the Harvard Dataverse repository: https://doi.org/10.7910/DVN/WW5NZD. This data includes fully anonymized transcripts for all 26 interviews as well as the analysis of each transcript in the format of Table 4's Individual Stakeholder Data Analysis Table Template. This data will be available up to 5 years after the successful completion of this PhD thesis defense.

To demonstrate the feasibility of the FGCR and to illustrate the pivotal role that contextuality plays in explanations and drives the need for a multiplicity of explanations even for the same phenomenon (series of car crashes), I have chosen to interrogate and present the analysis through four contextual configurations. Two from the vantage point of the *explainee(s)*, Stakeholder #1_Engineer (Non-AI Expert) in section 5.2.3 and Stakeholder #25_Police Officer in section 5.2.4. And two from the vantage point of the AI models which perform AI actions, perception in section 5.2.5 and system disengagement in section 5.2.6.

Upon completion of all 26 individual stakeholder data analyses, I then proceeded to aggregate the individual analyses. I will present a summary of the AI action types for which each of the 26 stakeholders sought explanatory information in section 5.2.2 below. Note that the two analyses from the vantage point of the AI models which perform AI actions are aggregated by multiple stakeholders' interviews about two separate AI actions.

5.2.1 Analysis of FGCR Task#1 Identification and Recruitment of Stakeholders – Participants' Stakeholder Categories and Key Characteristics

The analysis of FGCR Task#1, Identification and Recruitment of Stakeholders, is summarized in Table 6 below. Based on the order in which the interviews were conducted, each stakeholder was given a numerical designation denoted by "SH#". Note that SH# was assigned based on the order in which stakeholders were recruited and interviewed, with the first participants receiving the lowest numbers. A total of 26 participants were recruited.

Also note that, each participant stakeholder can be classified in two ways: First, all participants have a basic road user status as either a pedestrian, passenger, driver of AV, or a driver of other non-AV vehicle. However, for this study, each stakeholder was assigned a primary stakeholder category designation for the study. The primary stakeholder designation represented the perspective they held while providing insights to the study. The primary stakeholder designation typically represented their profession such as Engineer (Non-AI Expert); AI developer; Legal Expert; First Responder; Policy Maker; Insurance Industry. Where the category encompassed multiple professions (such as legal expert), I highlighted in red their particular profession within that category. And if the participant didn't fit into any professional category, they were classified solely by their road user status.

Also note that in the Table 6 below, I have provided information about each stakeholder's answers to two additional questions. Question 2a involved a rating their knowledge of AI systems. They were asked: How much do you know about AI systems? On a scale of 1 to 5 where:

1=no knowledge of AI systems technology
2=knowledge of different types of AI systems and have used some applications
3=top level abstract understanding of AI algorithms – types of applications/types of algorithms for different applications/functions (e.g., NN's, CNN's, RL's)
4=software engineering knowledge of various AI systems and applications
5=expert software engineering knowledge of different types of AI systems including being part of design team that has developed these systems.

And Question 4 pertained to whether and/or why they would want explanatory information about this case study scenario such as trust, transparency, other, etc.

Table 6:

Recruited Participant Designations & Key Characteristics

Stakeholder Designation (SH#)	Primary Stakeholder Category for Study	Road User Status	Question 2a	Question 4
SH1	Engineer (Non-AI Expert)	Driver of other non-AV vehicle on the road	2.5	Knowledge of how system works + trust/transparency
SH2	AI developer	Passenger	5	Improving AI models and algorithms
SH3	Engineer (Non-AI Expert)	Driver of other non-AV vehicle on the road	2	Trust; safety
SH4	Engineer (Non-AI Expert)	Driver of other non-AV vehicle on the road	2	Trust; safety
SH5	Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	2	Improving AI models and algorithms
SH6	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	2.5	Transparency and safety
SH7	Engineer (Non-AI Expert)	Pedestrian	3	Trust and safety
SH8	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	2	Transparency and then trust
SH9	Engineer (Non-AI Expert)	Driver of other non-AV vehicle on the road	3	Trust, understand the algorithm
SH10	Passenger	Passenger	1	Don't want explanation
SH11	Policy maker	Driver of other non-AV vehicle on the road	1	Safety
SH12	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	2.5	Improving AI models and algorithms
SH13	AI developer	Pedestrian	5	Transparency
SH14	Insurance Industry	Driver of other non-AV vehicle on the road	2	Transparency
SH15	Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	2	Safety
SH16	AI developer	Pedestrian	5	Transparency, safety, improving the algorithm
SH17	Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	1	Not really important to me
SH18	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Driver of other non-AV vehicle on the road	2	Safety, trust
SH19	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Driver of other non-AV vehicle on the road	1	Trust
SH20	Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	2	Indecipherable Answer
SH21	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Driver of other non-AV vehicle on the road	3	Transparency
SH22	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Driver of other non-AV vehicle on the road	2	Trust and Transparency
SH23	First Responder (e.g. Police, Ambulance, Fire, etc.)	Driver of other non-AV vehicle on the road	1.5	I don't really care about technology
SH24	First Responder (e.g. Police, Ambulance, Fire, etc.)	Driver of other non-AV vehicle on the road	2	Transparency
SH25	First Responder (e.g. Police, Ambulance, Fire, etc.)	Driver of other non-AV vehicle on the road	2	Trust
SH26	First Responder (e.g. Police, Ambulance, Fire, etc.)	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	2	Improving algorithm for safety reasons

5.2.2 Aggregate Analysis: Summary of Explanatory Information Sought by all Stakeholders by Types of Actions Taken/Decisions Made by AI

The results of the aggregate analysis summary of explanatory information sought by all stakeholders, organized by the types of actions taken and/or decisions made by the AV, are presented in Table 7 below.

The findings support the claim advanced in Chapter 1 that explanations are contextual, and multiple explanations will be sought for the same phenomenon – series of car crashes. This multiplicity of explanations stems from the combination of two factors: (i) the diverse nature of AI applications such as AVs which use multiple AI models executing multiple actions and, (ii) the contextuality of explanations and the needs of diverse *explainees.*

To analyze this multiplicity of explanations, I grouped each person's queries by the actions that the AV's AI models would perform. The groupings included:

- Perception/Detection
- Motion Control (Steering and Braking)
- System Disengagement (Timing, Process)
- Communication/Warning to Human

Also, a separate task pertaining to the actions of the AI designers which address the decisions made by AI designers rather than real-time system behavior was added to the list:

- Failure in Design/Training of AI algorithm/model

Table 7 provides an aggregate analysis of the types of actions executed by the AV for which each stakeholder sought explanatory information. This aggregate analysis demonstrates that even for the same phenomenon, multiple explanations were sought, driving the need for multiple xAI solutions. This multiplicity is due to the context-dependent nature of explanations, which is shaped by the diversity of multiple stakeholders and the distinct explanatory needs of each. Notably, even individual *explainees* sought multiple explanations that were contingent upon inquiries regarding the different actions executed by various AI models.

Table 7

Aggregate Analysis: Types of AI Actions Queried by Each Stakeholder

	SH CATEGORY	SH#	SH ASKED ABOUT ACTIONS TAKEN AND/OR DECISIONS MADE by AI Algorithm				
			Perception/ Detection	Motion Control (Steering and Braking) -- Decision Making & Taking Appropriate Action Failures	System Disengagement (Timing, Process)	Communication/ Warning to Human	Design/Training of the Algorithm (More Distal Actions of AI developers)
Engineer (Non-AI Expert)	Engineer (Non-AI Expert)	1	YES	YES	YES	YES	YES
	Engineer (Non-AI Expert)	3	YES	YES	YES	YES	NO
	Engineer (Non-AI Expert)	4	YES	YES	YES	YES	NO
	Engineer (Non-AI Expert)	7	NO	NO	YES	NO	YES
	Engineer (Non-AI Expert)	9	YES	NO	YES	NO	YES
AI developer	AI developer	2	YES	NO	NO	NO	YES
	AI developer	13	YES	YES	NO	NO	YES
	AI developer	16	YES	NO	NO	NO	YES
Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	12	YES	YES	YES	NO	NO
	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	6	YES	YES	YES	YES	NO
	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	8	YES	YES	NO	NO	YES
Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	5	YES	YES	YES	NO	NO
	Driver of other non-AV vehicle on the road	15	YES	YES	YES	NO	NO
	Driver of other non-AV vehicle on the road	17	NO	YES	YES	NO	NO
	Driver of other non-AV vehicle on the road	20	YES	YES	YES	NO	YES
OTHER	Policy maker	11	YES	YES	YES	NO	NO
OTHER	Pedestrian	10	YES	NO	YES	NO	YES
OTHER	Insurance Industry	14	YES	YES	NO	NO	YES
First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	23	YES	YES	NO	NO	NO
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	24	YES	NO	NO	NO	YES
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	25	YES	YES	NO	NO	YES
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	26	YES	YES	NO	NO	YES
Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	18	YES	YES	YES	NO	NO
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	19	YES	YES	YES	NO	YES
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	21	YES	YES	YES	YES	YES
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	22	YES	YES	YES	YES	YES

One prominent pattern emerged from the data analysis: 24 out of 26 stakeholders asked for explanatory information about the act of perception performed by the AV's AI systems. The main questions posed by these 24 stakeholders involved the issue of why the AV did not perceive the upcoming obstacles – they used phrases such as "detect", "sense", "see", etc. to seek explanatory information about the act of perception. While most stakeholders (23 out of 26) asked about more than one type of action, 3 stakeholders (SH#23, Police Officer and SH#2 and SH#16, AI Developers) asked only about the AI's act of perception along with queries about the distal tasks of the AI model(s)' design and training as performed by the AI designers.

Examining patterns across stakeholder categories showed that there were no discernable patterns between stakeholder category and the types of actions about which a particular stakeholder category member enquired. For example, members of the stakeholder category of police officer didn't ask about the same type of actions. The only pattern that emerged between stakeholder categories was the way questions were phrased. For example, engineers and AI developers tended to phrase their queries in terms of neutral root cause analysis whereas police officers, judges and lawyers tended to phrase their queries in terms of blame and/or responsible parties causing the failures.

5.2.3 Analysis of FGCR Task#2 – Stakeholder#1 Consultation for Identification of Explanans

To assess the feasibility of the FGCR, I am interrogating and presenting a disaggregated analysis of one individual stakeholder, Stakeholder #1_Engineer (Non-AI Expert). I will present the assessment of FGCR Task#2, Stakeholder Consultation for Identification of *Explanans*, in this section, and corresponding assessments of FGCR Task#3 in section 5.3.1 and FGCR Task#4 in section 5.3.2 for Stakeholder#1 (SH#1).

To illustrate the functionality of FGCR's Task#2, I am presenting an analysis of SH#1's transcript evaluating the contextual configuration of the same phenomenon that needs an explanation and the same *explainee.* Note that to support the discussion in this section, I am presenting only a

synthesized analysis of SH#1's transcript in Table 8a below. Table 8a contains a synthesized analysis highlighting the key phenomenon type designations (the "What") and causal chain and reasoning type designations (the "How"), along with commentary on the stakeholder's core query. The full analysis, including all classifications, verbatim evidence, and analytical commentary, appears in the corresponding Table 8b in Appendix D.

The results in this section also support the claim in Chapter 1 that explanations are contextual and that, even for the same *explainee,* multiple explanations are needed for the same phenomenon. The multiplicity of explanations sought by each individual stakeholder were driven by their inquiries into the different tasks performed by various AI models within the complex AI application of AVs as well as their unique and diverse perspectives.

I grouped this stakeholder's queries about the explanatory information sought by the types of AI actions to enable the next steps in the FGCR process – Tasks #3 and #4. This analysis will be used to generate detailed xAI model requirements for stage 2 of the testing (section 5.3) and will facilitate the identification of potential xAI models that could fulfill the stakeholder's explanatory needs for each specific AI action and/or decision.

Table 8a

Individual Stakeholder Data Analysis for Stakeholder#1_Engineer (Non-AI Expert)

STAKEHOLDER (SH# & Description):	SH1_Engineer (Non-AI Expert)							
	THE WHAT OF EXPLANATION			THE HOW OF EXPLANATION				
ACTIONS Executed by AI Models/ Algorithms Category	P_ID_NO & TYPE	M_ID_NO & TYPE	BIIR_ID_NO & TYPE	GSE_ID_NO & TYPE	CC_ID_NO & TYPE	CCTF_ID_NO & TYPE	CR_ID_NO & TYPE	Comments
PERCEPTION/ Detection	P1=Material Causes	M1.1 Device Fails to Adequately Perform Function	N/A	GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal		Could also be Distal -- questioning design
				GSE2=Causal Reasoning Type			CR2.6=Contrastive Designed/Specified/Expected Case	Multiple contrastive questions: [Why X vs. not X?] Contrast=could not see (X) vs. could see [Why X vs. Y?] Contrast=First responder vehicles (X) vs. other types of vehicles [Why X vs. not X?] [Why X vs. specified/designed/expacted value for X?"] Contrast=not avoiding obstacles (X) vs. designed/expected to avoid obstacles
Motion Control (Steering and Braking) -- Decision Making & Taking Appropriate Action Failures: i.e. Decision-making & Action (given potential perception) Failure	P1=Material Causes	M1.3 Device Failure in Decision Making	N/A	GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal		Focus is on decision to continue driving and not taking action to avoid Engineer wants to understand the step-by-step process (or breakdown in process) from perception to action. Most questions focus on immediate moments leading up to the crash -- why didn't it see, why didn't it react, why it disengaged late? Also asking Distal questions about how many other obstacles were avoided by this system in the past as opposed to these 16?
	P1=Material Causes	M1.2 Device Fails to Perform Appropriate Action		GSE2=Causal Reasoning Type			CR2.6=Contrastive Designed/Specified/Expected Case	Multiple Contrastive Reasoning: CR2.6=Actual inaction to avoid as opposed to expected action of avoiding. Could also be CR2.3 Non-common effect when asking distal questions about not avoiding these 16 vs. avoiding all other obstacles in the past ? This could also be construed as CR2.1 non-occurence of effect "why did not avoid vs. avoding obstacles? Could also be CR2.2 Normal the question highlights the abnormality of not avoindg and just giving up by not acting contrasting with normal/expected behaviore Also CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain
Communication/ Warning to Human	P1=Material Causes	M1.5 Device Communication Failure	N/A	GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal		Could also be characterized as CCTF1= Distal, namely, design time frame. Questions intention of designers & choices made by designers
	P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions	GSE2=Causal Reasoning Type			CR2.2=Contrastive The Normal Case	
System Disengagement (Timing, Process)	P1=Material Causes	M1.4 Device Fails to Engage/Disengage		GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal		
	P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions	GSE2=Causal Reasoning Type			CR4=Abnormality	CR4=Abnormality -- late (a second before crash) disengagement is abnormal and decision to abort when faced with a problem is also abnormal Also CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain
Design/Training of the Algorithm (More Distal Task Performed by AI developers)	P1=Material Causes	M2=Failure due to Design Flaw		GSE1=Causal Chain Type	CCT3=Unfolding	CCTF1=Distal		Questions design: How was this system trained before being put into the field?
				GSE2=Causal Reasoning Type			CR9=Mechanistic Causal	Overall, seeking CR9=mechanistic causal chain information for all AI functions performed. Could also be CR2.3= non-common effect, comparing training situation vs. AI learning/training

The interview with SH#1_Engineer (Non-AI Expert), yielded queries pertaining to all of the types of AI actions listed above in section 5.2 as well as the AI designers' actions. The following analysis uses the category and sub-type designations of the SSH-based faceted taxonomies (Figures 2a and 2b from section 3.1) to categorize the explanatory information sought by SH#1:

Action Performed by AI: Perception

For the act of perception, SH#1 asked the main question, "… how come … the AV couldn't see the … stationary … obstacle? … why it didn't detect these … these obstacles…"

Under the "what" category of explanation, the types of explanatory information sought included:

- phenomenon type P1=material causes
- sub-type M1.1=device fails to adequately perform the function of perception

Under the "how" category of explanation, this main question and its subsequent contextual comments were phrased in contrastive causal reasoning terms. SH#1 asked for multiple contrastive explanations:

- For contrastive explanation #1, SH#1 asked, "Why it didn't detect…".

  This posits an implicit contrastive type of CR2.1=non-occurrence of effect case of asking a "Why *X* vs. *not X"* type of question. SH#1 is asking for differentiating information about why it did not detect stationary obstacles versus the implied contrast of detecting them.

- For contrastive explanation #2, SH#1 asked, "…why first responder vehicles".

  This posits another implicit contrastive type of CR2.3=non-common effect case of asking a "Why *X* vs. *Y"* type of question. SH#1 is asking for differentiating information pertaining to why it did not detect first responder vehicles versus the implied contrast of other types of vehicles.

- For contrastive explanation #3, SH#1 asked, "… [why] the AV couldn't see this stationary … obstacle in general… because the whole system … is based on... identifying obstacles and of course avoid them...Which is the primary function and ADAS should have. . .".

  This posits contrastive type CR2.6=designed/expected case of asking a "Why *X* vs. *designed/expected value for X*" type of question. For this contrast, SH#1 is asking for differentiating information pertaining to why the AV couldn't see stationary obstacles versus what it was designed to do, namely, seeing stationary obstacles and ultimately avoiding them. This could also be construed as CR2.2=normal contrasting case where the differentiating factor is not seeing stationary obstacles versus the norm of seeing stationary objects.

Most of the questions for the act of perception were focused on an unfolding causal chain with a proximal timeframe focus. Portions of the questions could also be construed as questioning the design of the AI model, namely, the distal tasks performed by AI developers. Although I categorized the structures and levels of explanation type (Aristotle, Marr or Dennett) and the evaluation factors types (truth and coherence), these classifications proved to be too superficial for generating design requirements and I therefore omitted them from analysis in section 5.3 below.

Action Performed by AI: Motion Control (Steering and Braking)

For the actions of motion control (Steering and Braking), and particularly, the AV's failures in decision making and taking appropriate action with respect to motion control, SH#1 asked the main question,

> … why it didn't detect these ... these obstacles. And it went crashing [into them]… OK, I saw them. OK. Why didn't you avoid? ... many possibilities, 2 main ones of course is changing direction or …or braking. Why it doesn't happen?

This is a query not only about the task of perception but about the failure(s) associated within the decision and action chain.

Under the "what" category of explanation, the explanatory information sought were:

- phenomenon type P1=material causes
- sub-type M1.2=device fails to perform appropriate action (e.g. steering away and/or braking)
- sub-type M1.3=device failure in decision making

The focus of questioning is on the decision to continue driving and not taking actions to avoid the scene ahead (assuming that it detected the objects ahead).

Under the "how" category of explanation, the main question and its subsequent contextual comments were phrased in contrastive causal reasoning terms while focusing on an unfolding causal chain within a proximal timeframe. SH#1 wanted to understand the car's step-by-step process or breakdown in process starting with the task of perception through to its decisions and actions of continuing to drive despite obstacles up ahead.

SH#1 asked for multiple contrastive explanations:

- For contrastive explanation #1, the overall question of "why didn't you avoid? … [the obstacle]" can be construed as the CR2.6=designed/expected case of asking a "Why *X* vs. *designed/expected value for X*"; CR2.1=non-occurrence of effect case of asking a "Why *X* vs. *not X*"; or even CR2.2=normal case of asking "Why *X* vs. normal/default value for X" types of questions.

  For this contrast, SH#1 is asking for differentiating information pertaining the AV's inaction of not avoiding obstacles ahead versus the opposite, expected, or normal action of avoiding.

- For contrastive explanation #2, SH#1 asked, ". . . how many obstacles this system avoided before getting to this one that it ignored".

  This posits CR2.3=non-common effect case of asking a "Why *X* vs. *Y"* type of question. Here, SH#1 is asking questions about the AV's distal actions where it regularly avoided obstacles in its path. For this contrast, SH#1 is asking for differentiating information

pertaining to how the AV did not avoid obstacles in these 16 cases versus avoiding all other obstacles it encountered in the past.

As mentioned above, for these tasks, I omitted the structures and levels of explanation and evaluation factor categorizations from analysis in section 5.3 below.

Action(s) Performed by AI: Communication/Warning to Human & System Disengagement

For the failure in performing the actions of communication/warning to human and failures related to action of system disengagement, SH#1 asked the main question(s):

> why … does it [not] notify the driver before ... before aborting the aborting the autopilot? … I don't understand how a second before it realises that there is a problem and try to give control to the driver by aborting the… the control. … And why the conclusion of this interpretation was just to… to give up?

Under the "what of explanation categories, the explanatory information sought were:

- phenomenon type P1=material causes
- sub-type M1.4=device fails to engage/disengage
- M1.5=device communication failure.
- Portions of the quotes could also be construed as
  - P2=behavioral individual intentional reasons
  - subtype BIIR3-intentions about the AV choosing to disengage with only a second to go.

Under the "how" category of explanation, SH#1 queried these tasks in various contrastive terms. When SH#1 stated, ". . .normally it should be programmed to … just give control to the driver and notify the driver", this can be construed as CR2.6=designed/expected case of asking a "Why *X* vs. *designed/expected value for X*"; or even CR2.2=normal case of asking "Why *X* vs. *normal/default value for X*" types of questions. Here, SH#1 is asking for differentiating information pertaining to choosing to disengage without notification versus the abnormal and/or expected actions of simply "giving up" and aborting control rather than acting. While the questions focused on an unfolding causal chain within a proximal timeframe, SH#1 was also asking about distal timeframe with respect to the design of the AI model.

Action Performed by AI Designers: Design/Training of AI Model(s)

Overall, regardless of queries about the actions taken and/or decisions made by the AI, all explanatory information sought throughout the interview, SH#1 wanted to understand and asked questions about the design and training of the AI model(s) especially how and which components of the AI model(s) and/or algorithm(s) failed. SH#1 is seeking a mechanistic causal reasoning process narrative that identifies the necessary and salient components in the causal chain that led to the outcome(s).

This assessment of FGCR Task #2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*, has yielded a set of explanatory information sought by SH#1. The outputs of this task include questions asked by SH#1 that represent the *explanans* sought, which were grouped by action performed by the AI and classified using the SSH-based taxonomy. The *explanans* sought by SH#1 included 4 AI actions (perception, motion control (steering/braking), system disengagement, and communication/warning to human) as well as the actions performed by the AI designers of design/training of the AI models. These outputs of stage 1's data analysis will feed into the corresponding assessments of FGCR Task#3 in section 5.3.1 and FGCR Task#4 in section 5.3.2 for SH#1.

5.2.4 Analysis of FGCR Task#2 – Stakeholder#25 Consultation for Identification of Explanans

In this section, I am interrogating and presenting a disaggregated analysis of another individual stakeholder, Stakeholder#25_Police Officer. I will present the assessment of FGCR Task#2 in this section, and corresponding assessments of FGCR Task#3 in section 5.3.3 and FGCR Task#4 in section 5.3.4 for Stakeholder#25 (SH#25).

To illustrate the functionality of FGCR's Task#2, Stakeholder Consultation for Identification of *Explanans*, I am presenting an analysis of the transcript for SH#25 in Table 9a below. This analysis evaluates the contextual configuration of the same phenomenon (series of car crashes) that needs an

explanation, the same individual *explainee*, along with the added factor of enabling a comparison to section 5.2.3's SH#1, a separate and different *explainee.* Note that to support the discussion in this section, I am presenting only a synthesized analysis of SH#25's transcript in Table 9a below. Table 9a contains a synthesized analysis highlighting the key phenomenon type designations (the "What") and causal chain and reasoning type designations (the "How"), along with commentary on the stakeholder's core query. The full analysis, including all classifications, verbatim evidence, and analytical commentary, appears in the corresponding Table 9b in Appendix D.

The results in this section also support the claim in Chapter 1 that explanations are context-driven: even for the same *explainee,* multiple explanations are needed and, when compared to another stakeholder (section 5.2.3's SH#1), different *explainees* will seek different (and sometimes multiple) explanations for the same phenomenon. The multiplicity of explanations sought by each individual stakeholder was driven by their unique and diverse perspectives which informed their inquiries into the different tasks performed by various AI models within the complex AI application of AVs.

This analysis, which reflects the explanatory information sought by SH#25, will be used to generate detailed xAI model requirements for stage 2 of the testing (section 5.3) and facilitate the identification of potential xAI models that could fulfill the stakeholder's explanatory needs for each specific AI action and/or decision.

Table 9a

Individual Stakeholder Data Analysis for Stakeholder#25_Police Officer

STAKEHOLDER (SH# & Description):	SH25_Police Officer							
	THE WHAT OF EXPLANATION			THE HOW OF EXPLANATION				
ACTIONS Executed by AI Models/ Algorithms Category	P_ID_NO & TYPE	M_ID_NO & TYPE	BIIR_ID_NO & TYPE	GSE_ID_NO & TYPE	CC_ID_NO & TYPE	CCTF_ID_NO & TYPE	CR_ID_NO & TYPE	Comments
PERCEPTION/ Detection	P1=Material Causes	M1.1 Device Fails to Adequately Perform Function		GSE1=Causal Chain Type	CCT3=Unfolding	CCTF1=Distal		Questions on capabilities of AI algorithm and whether it was designed properly
	P1=Material Causes	M2=Failure due to Design Flaw		GSE2=Causal Reasoning Type		CCTF2=Proximal	CR2.3=Contrastive Non-Common Effect	Could be CR2.2 Norm = ability to recognize flashing lights and sirens as opposed to other normal conditions on road Also compares to humans having tendency to drive towards lights where their eye focuses on Also, CR2.3 is about contrasting to very specific way that first responders park at 45 degree angle at active collision scene and blocking more than one lane-- is AI programmed to recognize such objects? Also, questioning physical hardware's capability (camera) to see far enough ahead could be construed as CR4=abnormality or CR1=abductive Questioning design choices made by designers of AV's perception system
	P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions	GSE3=Structures/Levels of Explanation Type				
Design/Training of the Algorithm (More Distal Task Performed by AI developers)	P1=Material Causes	M2=Failure due to Design Flaw		GSE1=Causal Chain Type	CCT6= Common-Cause Chain	CCTF1=Distal		
	P1=Material Causes	M1.6 Device Failure in Information Processing		GSE2=Causal Reasoning Type			CR2.3=Contrastive Non-Common Effect	Could also be CR4=abnormality and CR1 Abductive looking at multiple hypotheses especially with respect to how first responders park their vehicles at collision scene
	P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions					
Motion Control (Steering and Braking) -- Decision Making & Taking Appropriate Action Failures	P1=Material Causes	M1.3 Device Failure in Decision Making		GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal		
	P1=Material Causes	M2=Failure due to Design Flaw		GSE2=Causal Reasoning Type			CR2.2=Contrastive The Normal Case	CR2.2: Contrasting to normal case of pulling over and stopping or steering away as Teslas have done many times before. Could also be CR2.6 contrasting to designed/expected/specified case of previous actions by Tesla Could also be CR2.1 contrasting to not stopping.
	P1=Material Causes	M1.6 Device Failure in Information Processing						
	P1=Material Causes	M1.2 Device Fails to Perform Appropriate Action						
	P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions					

The interview with SH#25_Engineer (Non-AI Expert), yielded queries pertaining to the AI's actions of perception and motion control (steering and braking) and to the AI designers' actions of design/training of the algorithm. The following analysis uses the category and sub-type designations of the SSH-based faceted taxonomies (Figures 2a and 2b from section 3.1) to categorize the explanatory information sought by SH#25:

Action Performed by AI: Perception

For the act of perception, SH#25 made the following queries (quoted verbatim):

… is the AI programmed to see lights and sirens and what that means to them, to the programming

… an AI computer can't comprehend if a fire truck is in a 45-degree angle in lane one, lane two, lane three and then it kind of just shuts off 'cause …it's like, oh, I've never seen this

… what about the camera system? Like you know the … you know how the lens or like the … how far it can project …

Under the "what" category of explanation, the types of explanatory information sought included:

- phenomenon type P1=material causes
- sub-type M1.1=device fails to adequately perform the function (of perception)
- sub-type M2=failure due to design flaw
- phenomenon type P2=Behavioral Individual Intentional Reasons (pertaining to design choices made by designers)
- sub-type BIIR3=intentions

Under the "how" category of explanation, the questions and their subsequent contextual comments were phrased in contrastive causal reasoning terms. SH#25 asked for multiple contrastive explanations:

- For contrastive explanation #1, SH#25 specifically stated,

> And then basically, when firefighters and police officers go to a scene, they're taught to go 45° and take up…. An AI computer can't comprehend if a fire truck is in a 45°

> angle in lane one, lane, two, lane three and then… it kind of just shuts off 'cause… It's like, oh, I've never seen this

This posits a contrastive type of CR2.3=non-common effect case of asking a differentiating factor "Why *X* vs. *Y"* type of question that focuses on the comparing the abnormality of the first responder procedure of parking at 45° angle at emergency sites when compared to normal stopping on the road. SH#25 is asking for differentiating information about whether the computer can comprehend that the obstacles are stationary first responder vehicles which follow the standard procedure of parking at a 45° angle taking up more than one lane at accident and emergency scenes. Is the AI able to detect these obstacles as opposed to other obstacles such as moving vehicles or vehicles stopped facing straight forward? This could also be categorized as CR4=Abnormality of the angle and position of first responder vehicle parking at emergency sites.

- For contrastive explanation #2, SH#25 asked, "…is the AI programmed to see lights and sirens and what that means to them, to the programming."

  This posits a contrastive type of CR2.2=the normal case of asking a "Why *X* vs. *the default value for X?"* type of question. SH#25 is asking about the AI's ability to recognize flashing lights and sirens as opposed to other normal conditions (vehicles without flashing lights, other roadway signs or lights that don't flash, etc.) on the road.

- For contrastive explanation #3, SH#25 specifically stated,

  > ...They even say the natural… natural person when they're driving, when they see lights and sirens, their eye brings them to that area, and they drive into the cars. That's why we're taught to be safe when we're doing it. But I'm wondering is… in is it programmed in AI Autopilot and what does it tell it to do if there's lights and sirens on the road?

This posits contrastive type CR2.2=the normal case of asking a "Why *X* vs. *the default value for X?"* type of question. For this contrast, SH#25 is comparing the AI's actions to those of humans who have a tendency to drive towards the things that their eyes focus upon such as flashing lights.

- Another type of causal reasoning explanation involves SH#25 asking, "… what about the camera system? Like you know the … you know how the lens or like the … how far it can project … ".

  This posits causal reasoning type CR1=abductive where SH#25 is seeking potential explanations for whether the hardware system is capable of seeing far enough ahead in order to identify obstacles and have enough time to react.

Most of the questions for the act of perception were focused on an unfolding causal chain with a proximal timeframe focus on the items that the AI algorithm did or did not perceive or comprehend. Portions of the questions could also be construed as questioning the design of the AI model, namely, the distal tasks performed by AI developers. Although I categorized the structures and levels of explanation type (Aristotle, Marr or Dennett) and the evaluation factors types (truth and coherence), these classifications proved to be too superficial for generating design requirements and I therefore omitted them from analysis in section 5.3 below.

Action Performed by AI: Motion Control (Steering and Braking)

For the actions of motion control (steering and braking), and particularly, the AV's failures in decision making and taking appropriate action with respect to motion control, SH#25 specifically stated the following,

> . . . did you see the lights, sirens or did you see the lights and did you see the apparatus blocking the roadway and if so, why didn't you stop or I would be like, yeah, what was the reasoning you continued to go forward into the vehicles, did you not see them? And if you didn't see them, what were you doing?

> . . . I would think there'd be something wrong with the algorithm of why it didn't … 'cause obviously it could. I've seen videos where it veers off quickly if someone falls onto the roadway, and the Tesla veers over, but why didn't at this moment is my question. What? What was it doing? What was it computing at that time that it didn't recognise this as a as a "brake" kind of thing or?

This is a query about the failures of multiple integrated actions including perception, motion control (braking), decision making and taking appropriate actions.

Under the "what" category of explanation, the types of explanatory information sought included:

- phenomenon type P1=material causes
- sub-type M1.2=device fails to perform appropriate action (e.g. steering away and/or braking)
- sub-type M1.3=device failure in decision making
- sub-type M1.6=device failure in information processing
- sub-type M2-failure due to design flaw
- phenomenon type P2-behavioral individual intentional reasons (pertaining to design choices made by designers)
- sub-type BIIR3=intentions (of designers)

The focus of questioning is on the decision to continue driving and not taking actions to avoid the scene ahead and whether the AV even detected or recognized the objects ahead.

Under the "how" category of explanation, the questions and their subsequent contextual comments were phrased in contrastive causal reasoning terms while focusing on an unfolding causal chain within a proximal timeframe. SH#25 wanted to understand the car's step-by-step decision-making process starting with the task of perception through to its decisions and actions of continuing to drive despite obstacles up ahead.

SH#25 asked for multiple contrastive explanations:

- For contrastive explanation #1, the question of "… why didn't you stop?", assuming it saw the obstacles, can be construed as the CR2.6=designed/expected case of asking a "Why *X* vs. *designed/expected value for X*"; CR2.1=non-occurrence of effect case of asking a "Why *X* vs.

not X"; or even CR2.2=normal case of asking "Why *X* vs. normal/default value for X" types of questions.

For this contrast, SH#25 is asking for differentiating information pertaining the AV's inaction of not stopping versus the opposite, expected, or normal action of stopping.

- For contrastive explanation #2, SH#25 stated, "… I've seen videos where it veers off quickly if someone falls onto the roadway, and the Tesla veers over, but why didn't at this moment is my question. What? What was it doing? What was it computing at that time that it didn't recognize this as a 'brake' kind of thing?".

  This posits CR2.6=designed/specified/expected case of asking a "Why *X* vs. *the designed/specified/expected value for X"* type of question or CR2.2=normal case of asking a "Why *X* vs. *default value for X*?" type of question. Here, SH#25 is asking questions about the AV's, specifically, Tesla's distal actions where it is known to regularly avoid obstacles in its path. For this contrast, SH#25 is asking for differentiating information pertaining previous incidences involving Teslas that avoided obstacles in their path that they have encountered in the past.

As mentioned above, for these tasks, although I categorized the structures and levels of explanation type and also the evaluation factor types, these classifications were deemed inadequate for generating design requirements and I therefore omitted them from analysis in section 5.3 below.

Action Performed by AI Designers: Design/Training of AI Model(s)

Overall, similar to SH#1, regardless of queries about the actions taken and/or decisions made by the AI, all explanatory information sought throughout the interview, SH#25 wanted to understand and asked questions about the design and training of the AI model especially how and which components of the AI model(s) and/or algorithm(s) failed. SH#25 is seeking a mechanistic causal reasoning process narrative that identifies the necessary and salient components in the causal chain that led to the outcome(s).

This assessment of FGCR Task #2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*, has yielded a set of explanatory information sought by SH#25. The outputs of this task include questions asked by SH#25 that represent the *explanans* sought, which were grouped by action performed by the AI and classified using the SSH-based taxonomy. The *explanans* sought by SH#25 included 2 AI actions (perception and motion control (steering/braking)) as well as the actions performed by the AI designers of design/training of the AI models. These outputs of stage 1's data analysis will feed into the corresponding assessments of FGCR Task#3 in section 5.3.3 and FGCR Task#4 in section 5.3.4 for SH#25.

5.2.5 Analysis of FGCR Task#2 – Stakeholder Consultation for Identification of Explanans for Multiple Stakeholders for AI's Act of Perception

In this section, I am interrogating and presenting an aggregated analysis of the transcripts of multiple stakeholders as it pertains to the explanatory information sought by them for the specific act of perception performed by the AV. I will present the assessment of FGCR Task#2, Stakeholder Consultation for Identification of *Explanans*, in this section, and corresponding assessments of FGCR Task#3 in section 5.3.5 and FGCR Task#4 in section 5.3.6.

To illustrate the functionality of FGCR's Task#2, I am presenting an analysis of the transcripts for the specific act of perception performed by the AV in Table 10 below. Note that to support the discussion in this section, I am presenting an analysis of causal reasoning types and descriptors of information sought by multiple stakeholders including verbatim quotes in Table 10 below.

This analysis evaluates the contextual configuration of the same phenomenon (series of car crashes) that needs an explanation, multiple *explainees*, and the same action performed by the AI. The results in this section also support the claim in Chapter 1 that explanations are contextual where the multiplicity of explanations sought is driven by the fact that different *explainees* will seek different and (sometimes multiple) explanations for the same phenomenon, and within that, the same action

performed by the AI system. Also, due to the diverse needs of each individual explainee, this holds true even when examining the data from the vantage point of the AI's actions.

As outlined in section 5.2.2's Table 6, 24 out of 26 stakeholders asked for explanatory information about the act of perception. From among all stakeholders who sought explanatory information, the main question involved the issue of why the AV did not detect the upcoming obstacles. Under the "what" category of explanation, the explanatory information sought by most stakeholders were:

- phenomenon type P1=material causes
- sub-type M1.1=device fails to adequately perform the function of perception

One exception was SH#23 who articulated her questions in anthropomorphic terms and therefore, I categorized the explanatory information sought by her as:

- P2=behavioral unintentional causes
- sub-type behavioral individual cause BIUC3= habit

The prevalence of anthropomorphic language deserves comment: Although most stakeholders phrased their queries in anthropomorphic terms (e.g. by using terms like "see", "recognize", "drive", "avoid", "tell", etc. when describing the AV's actions), when looking at the overall context and thrust of the conversation, it is evident that they knew they were talking about a non-human entity – a "thing" and not a person. Furthermore, even though they used phrases such as "see" or "did you see" amongst others, they always came back to using the pronoun "it" when speaking about the AV and its decisions and actions. The anthropomorphic language appears to serve as accessible shorthand to describe complex technical processes rather than implying that the system truly possesses human qualities.

Under the "how" category of explanation, the main question and its subsequent contextual comments were overwhelmingly phrased in contrastive causal reasoning terms. Table 9 provides a detailed causal reasoning sub-type designations per the SSH-based faceted taxonomies about the explanatory information sought by all the stakeholders. The specific contrastive comparisons varied by

stakeholder but shared a common structure: “Why did the AV fail to perceive X in these 16 cases when it successfully perceived similar instance Y in the past?”

And, when one takes a bird’s eye view of the overall thrust of the questions as they pertained to all AI actions (not only perception), but the majority of stakeholders were also asking for an explanation with respect to the failures in the design and training of the AI models themselves. This meta-level inquiry suggests that local, instance-specific explanations may be insufficient because stakeholders are seeking to understand the systemic limitations of the AI models that manifested in these particular failures.

These results of stage 1’s data analysis will serve as inputs to stage 2 of proof-of-concept testing of the FGCR’s Task#3, Generation of Detailed xAI Model Requirements, in section 5.3.5 and Task#4, Identification of Potential xAI Models that Could Provide Explanation(s), in section 5.3.6.

Table 10

Aggregate Analysis of Information Sought by Stakeholders for the AI's Act of Perception

	SH CATEGORY	SH#	Failure in Executing Act of PERCEPTION/ Detection Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
Engineer (Non-AI Expert)	Engineer (Non-AI Expert)	1	Quote: "I think the first obvious question is how come ... uh.... the AV couldn't see this stationary uh... let's say obstacle in general. **CR2.1 (Non-occurrence of Effect): [Why X vs. not X?] Contrast=could not see (X) vs. could see** Quote: "..Now my second question is why first responder vehicles? **CR2.3 (Non common case) [Why X vs. Y?] Contrast=First responder vehicles (X) vs. other types of vehicles** Quote: "Because I understood that the whole cases are first responders, responder vehicles, no? Which is the primary function and ADAS should have." **CR2.1=[Why X vs. not X?]** Quote (Engineer 7:51): "But how come a system that is mainly built to avoid obstacles, go crashing into obstacles?" **CR2.6 (Actual vs. Designed/Expected Case) [Why X vs. specified/designed/expacted value for X?"] Contrast=not avoiding obstacles (X) vs. designed/expected to avoid obstacles**
	Engineer (Non-AI Expert)	3	Quote "...do you see an object, do you see an object? -- implies did not see object. **CR2.1 (Non-occurrence of Effect) &** **CR2.6 (Actual vs. Designed/Expected Case)** **Contrast=[Why X vs. not X?] Contrast=did not see object vs. see object**
	Engineer (Non-AI Expert)	4	Quote: "...Did the vehicle sense there was a vehicle in front of it?" -- implies did not sense object. **CR2.1 (Non-occurrence of Effect) [Why X vs. not X?] && CR2.6 (Actual vs. Designed/Expected Case [WHY X vs. designed value for X] Contrast=did not sense object in front (X) vs. sense object in front**
	Engineer (Non-AI Expert)	7	N/A -- DID NOT ASK ABOUT PERCEPTION FUNCTION
	Engineer (Non-AI Expert)	9	Quote: ...when it sees lights, is it able to distinguish a typical stop light from something that is not a standard stop light? Does it distinguish colors of light?" Quote: "Yeah. Could it distinguish what is a typical traffic light from one that is an emergency light? Can a distinguished color?" Quote: "...it didn't recognize red as a signal of an issue to respond to. That's a problem. Do flashing lights become a problem.." **CR2.2=Normal Case (Actual vs. default value) &** **CR2.3=Non-common Effect (Actual vs. Something Else) &** **CR2.6 Designed/Specified/Expected Case** **[Why X vs. Y] Multiple contrastive questions: Contrast#1=emergency lights on first responder vehicles (red/yellow and flashing) (X) vs. non-flashing lights such as traffic red light Contrast#2:not recognizing red lights (X) vs. recognize typical stop red light**

	SH CATEGORY	SH#	Failure in Executing Act of PERCEPTION/ Detection
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
AI developer	AI developer	2	Quote: "...So, is the system getting the right information to process and make a decision? That's my first question. ... What exactly the input was at that point in time? ... Maybe you know the history and passing, but the model is saying something else, so that's like a failure to detect." Quote: "What was the prediction of the model? What was the detection like? You know we you see you give it a person's image and it detects something else, maybe some, you know, shadow of the person or something like that. So yeah, that's basically failure. And that's in the 1st place. Why these accidents happen, right? Failure to detect right things." Quote: "So I want to know . . . from the input, what was the input for the model and then what was the decision based on that input? You know, sometimes you have the right inputs and you get the wrong decision because of, you know failure to detect." **CR2.2=Normal Case (Actual vs. default value) &** **CR2.6 Designed/Specified/Expected Case** **[Why X vs. designed/expected/specified X?] Contrasting incorrect processing resulting in failure to detect (X) vs. correct processing which would result in accurate detection & Also Abductive Reasoning**
	AI developer	13	Quote (Stakeholder13 14:11:04): "...is there something that it saw..[or].. that maybe it didn't see?" **CR2.1 (Non-occurrence of Effect) Contrast =[Why X vs. not X] Contrast case=not seeing (X) vs. seeing** Quote (Stakeholder13 14:20:14) "I'm thinking computer vision systems. So like maybe, you know, it's like it sees a flashing light and that should register as something... it can put pieces together and maybe it knows scenes... what it's been trained on... to understand what an accident is... understand what they're seeing or just like know what pieces are there that should cause an action to occur." **CR2.6 Designed/Specified/Expected Case Contrast=[Why X vs. designed/expected/specified X?] Contrast= not registering flashing lights (X) vs. designed to register/understand flashing lights and accident & also abductive reasoning related to design activities** **Also, CR1=Abductive**
	AI developer	16	Quote (Stakeholder16 00:14:04): Multiple trucks with freaking lights going on… on top of them. So my first question is, did it detect the trucks? I assume it did detect the trucks. If it didn't detect the trucks, this would be really weird. Maybe it detected the trucks, but I thought it was like maybe it saw a row of trucks and thought it was a bridge or something. **CR2.3=Non-common Effect (Actual vs. Something Else) =[Why X vs. Y] Contrast=detecting trucks (X) vs. detecting row of trucks as bridge of lights** Quote (Stakeholder16 00:12:58):I assume they have a way to detect cars, signs, those kinds of things. **CR2.6 Designed/Specified/Expected Case=[Why X vs. designed/expected/specified X?] = Contrast=not detecting (X) vs. designed to detect** SH16 15:36-18:39: "I would look at from 2 perspectives, the perspective of object detection... And then the more reaction perspective where it's like, what did it see during its training and how did you teach it to react to during those situations?" **CR1=[Abductive reasoning related to design activities]**

	SH CATEGORY	SH#	Failure in Executing Act of PERCEPTION/ Detection
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	12	Quote: "...when it sees lights, is it able to distinguish a typical stop light from something that is not a standard stop light? Does it distinguish colors of light?" **CR2.3=Non-common Effect (Actual vs. Something Else) Contrast=[Why X vs. Y] Contrast= non-standard, non-red lights (X) vs. standard stop light which is red and not flashing** Quote: "So it was that they didn't somehow sense the object." **CR2.1 (Non-occurrence of Effect) [Why X vs. not X] Contrast= not sensing (X) vs. sensing** **Also, CR1=Abductive**
	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	6	Quote: "...the sensors are everything for these systems. So in the Tesla's case, I guess it's the cameras. So if a camera is blocked or blurred or there's snow..." N/A Focuses on hardware aspect of perception only & not on algorithm's performance of perception function.
	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	8	Quote: .. Is it because it's missing data to stop? Is it because the cameras were blinded by flashing lights? Is there some sort of camera obstruction? I know in the past it's asked me to take over when it's really sunny outside and you can, you know, the sun is blinding your eyes, so it must be blinding the front camera as well. … is it camera obstruction? It… it … was it… a really sunny day? Is it a lack of data on how to handle vehicles parked? I've also observed it not stop at a … so you've got a bus, a bus with a the… the stop extended out flights lights are flashing, … it's slowing down. I've had to take over. **CR1=Abductive Reasoning &** **CR2.3=Non-common Effect (Actual vs. Something Else)=[Why X vs. Y] Contrast= SH attempts to identify multiple differentiating factors as alternative possible causes (Y) of failures such as blocked cameras; sunny day; blinding light; etc.**
Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	18	Quote: "So AI is not somehow … is not capable of… of understanding those first responders because of the lights, you know going or because of their location." Quote: "...I would think immediately of the of the flashing lights if that's somehow screwing up the …the cameras of the Tesla." **CR2.1 (Non-occurrence of Effect)=[Why X vs. not X] Contrast=not capable of understanding first responders with flashing lights vs. capable of understanding this**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	19	Quote :I mean in terms of identity or identifying to me whether it is a.. a …a fire truck ahead or a child, or an older person, or some object that could be dangerous. Quote (Stakeholder19 00:21:00): And because the visual identification, in my mind, so long as it identifies a something that will cause a danger that would result in a …in a… in a disastrous or dangerous situation. To me, it doesn't matter whether it's a traffic light, a person, or otherwise. There there's got to be a sensor of sort. **CR1=Abductive reasoning about recognition of person, object, dangerous situation**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	21	Quote: Like, I would want to know like what system was it that failed? Was it the perception? Was it the braking? Was it the decision-making? Was it the interpretation of the inputs that it was perceiving on the road? Like what system failed? And … and once I knew what system failed, then I would want to know more about that system. **CR1=Abductive reasoning about system and its functions of perceiving and interpreting the inputs**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	22	Quote: I would like the AV system to show me how it recognised um… usually,… obstacles. But in particular this instance, why the AV system is not recognising Emergency crew or emergency vehicles or … or people? **CR2.1 (Non-occurrence of Effect)=[Why X vs. not X] Contrast= not recognizing emergency crew with people vs. recognizing emergency vehicles and people**

	SH CATEGORY	SH#	Failure in Executing Act of PERCEPTION/ Detection
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	5	Quote: ... Because those are big, big vehicles, and they have visible lights, big size, the big shape than other cars, so it's quite visible. So I just want to learn why, what he was thinking about object." Quote : "If you did see, did you categorize those big vehicles as vehicle, an object, or did you consider that something else?" **CR2.3=[Why X vs. Y] Contrast=categorize objects as big vehicles with visible lights (X) vs. categorize objects as something else** Quote: "But I'm maybe, maybe did the Autopilot didn't consider it as an object on the road. ... It can be maybe a balloon or something? There's a flashy thing. Maybe the flash, you know, maybe it impacts the hardware to see the vehicle..." **CR2.3=[Why X vs. Y] Contrast=big vehicles vs. balloon or flashing thing** **CR2.3=Non-common Effect (Actual vs. Something Else) MULTIPLE CONTRASTS: [Why X vs. Y] Contrast=categorize objects as big vehicles with visible lights (X) vs. categorize objects as something else & [Why X vs. Y] Contrast=big vehicles vs. balloon or flashing thing** **CR1=Abductive** Quote: "First question for me, why you did not realize the vehicle?" Why did it NOT recognize, when it should have? Questioning the AI's "thinking" (categorization) about the object. The stakeholder is seeking the cause of the perception failure **CR2.1 (Non-occurrence of Effect) Contrast [Why X vs. not X] Contrast=recognize vs. not recognize vehicle**
	Driver of other non-AV vehicle on the road	15	Quote: "… I mean, do you wanna how they did not interpret what it looks like ... structures up ahead as what they are, what… what mechanisms were… not performed well? " Quote: "Yes, I want to know why Autopilot is not recognizing these features coming up and they're keeps driving into a structure like where is the confusion?" Quote: So say there was an accident and the Tesla drives past it, and then the next one, this one of the 16, it drives into it. Where is the difference? Where's the dichotomy between the two. And where like the… the issue to get addressed? Explain to me where the confusion was between the one time you passed it versus this one. What do you? What are you mixing up? What are you confused about? Was it more lights? Was it structures? Was it weather? All that kind of stuff, 'cause, you know, was there fog that that threw it off? Was it daytime? Nighttime? I don't know. I mean, I don't know much about it, but I would assume fog or rain or sleet something may impact its ability to process. Are the sensors not detecting? Quote: "Why is the autonomous vehicle not interpreting it as a dramatic danger coming on." **CR2.3=Non-common Effect (Actual vs. Something Else)** **[WHY X vs. DEFAULT VALUE FOR X] AND/OR [WHY X vs. Y] & CR2.2=Normal Case (Actual vs. default value) [WHY X vs. default value for X] Contrast=drives into these 16 (X) vs. all past instances of situations like these** **CR2.1=[Why X vs. not X] Contrast=not recognizing structures up ahead (X) vs. recognizing them Contrast=not interpreting dramatic danger coming on (X) vs. interpreting dramatic danger** **CR1=Abductive**
	Driver of other non-AV vehicle on the road	17	**N/A -- DID NOT ASK ABOUT PERCEPTION FUNCTION**
	Driver of other non-AV vehicle on the road	20	Quote: "Well, I would want to know why they didn't recognize the …the obstacle that was in their way …that they hit like, why was that not seen by them?" Quote: ...Some sensor that should have taken over and been able to see that obstacle or recognized black and ... I don't know how it's set up, but it like what happened to the program that was supposed to do that? **CR2.1 (Non-occurrence of Effect) Contrast= [WHY X vs. not X] Contrast=not recognizing obstacle in the way(X) vs. recognizing obstacle** **Also, CR1=Abductive reasoning questioning design of program**

	SH CATEGORY	SH#	Failure in Executing Act of PERCEPTION/ Detection
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
OTHER	Policy maker	11	Quote: "So, like, why didn't they see the .. the vehicles that were parked in front? Why didn't they see the humans?" **CR2.1 (Non-occurrence of Effect) [Why X vs. Not X] Contrast is "why not see" vs. "seeing."[Why X vs. Not X] Contrast is "why not see" vs. "seeing."** CR2.3 Contrastive non-common effect: emphasizing the salience of the missed cues such as highly salient cues (flashing lights, warning signs and questioning size of objects to detect).. Quote: "So why didn't they see the flashing lights? Why didn't they detect that? Were there warning signs that you know, even before you got?" Quote :"...how big does an object have to be? So is it so… Did they see the when they were approaching the ...the emergency vehicles where they …that… they see a tiny little thing...is it like that …that it seems like really far away and it's like 'cause it can't see obviously. So it has to sense everything I suppose. So what was the object?" **CR2.3=Non-common Effect (Actual vs. Something Else)[Why X vs. Y] Contrast= flashing lights (X) vs. non-flashing lights; big objects vs. small objects such as humans and smaller cars** **CR1=Abductive reasoning where Policymaker is forming hypotheses (abductive) about the perception failure and focusing on the abnormality of not seeing such objects.** **Also, CR9=Mechanistic Causal Reasoning: Wants an unfolding of the perception process.**
OTHER	Pedestrian	10	Quote: "...when it sees lights, is it able to distinguish a typical stop light from something that is not a standard stop light? Does it distinguish colors of light?" Quote: "Yeah. Could it distinguish what is a typical traffic light from one that is an emergency light?" Quote: "They were. …They were red and it didn't recognize red as a signal of an issue to respond to. That's a problem. Do flashing lights become a problem..." **CR2.3=Non-common Effect (Actual vs. Something Else)=[Why X vs. Y] focuses on multiple differentiating factors: Is it able to distinguish between stop light and emergency light? Flashing lights vs. non-flashing?** **Also could be CR1=abductive reasoning by positing multiple hypotheses about sensing different types of lights**
OTHER	Insurance Industry	14	Quote: "...why did they did it not pick up these … these individuals or these vehicles as well as individuals that have high visibility in in front of them?..... I would think any vehicle, any, any obstruction in front of them it should…. recognise" Quote: "Or …or causing an influx where it's the flashes or it's the… the reflectors with the with the technology or the sensors that has caused the issue." **CR2.1 (Non-occurrence of Effect) [Why X vs. Not X] [or X vs. Y with multiple X's/Y's]: Contrast#1=not see individuals with high visibility [vests] vs. see individuals; Contrast#2=not see vehicles vs. see vehicles Contrast#3=not see any obstruction in front vs. see obstruction Contrast#4=not see flashes (i.e. flashing lights) or light reflected from high visibility vests vs. see these things** **Could also be CR2.3=Non-common Effect (Actual vs. Something Else) [Why X vs. Y]**

	SH CATEGORY	SH#	Failure in Executing Act of PERCEPTION/ Detection
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	23	Quote: what's making it attracted to maybe lights or something like why specifically emergency vehicles with lights flashing type of thing? Quote: Well, like we know that people are attracted to light, so oftentimes they do actually hit like vehicles that are at the side of the road with their lights on 'cause they you drive where you're looking, right? So that's the direction you go. **CR2.2=Normal Case (Actual vs. default value) =Anthropormorphizing by comparing to previous experience with human drivers looking in direction to where they drive and accidents up ahead attract visual attention**
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	24	Quote: "Is it capable of seeing like I mean, if… if it's specifically like emergency vehicles like they have a lot of like, you know, reflective… like symbols and stuff like that on the on the vehicle. That could be like potentially difficult for the… the sensors to pick up." Quote: What does it recognise as like tangible objects? **CR2.1 (Non-occurrence of Effect) =[Why X vs. not X] Contrasting ability to recognize/perceive high visibility vs. non-high visibility obstacles**
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	25	Quote: "...I guess the light and sirens were on or the lights were on so that the AI can interpret that the lights are on. But I guess maybe as the AI programmed to see lights and sirens and what that means to them, to the programming." Quote: And then basically, when firefighters and police officers go to a scene, they're taught to go 45° and take up…. An AI computer can't comprehend if a fire truck is in a 45° angle in lane one, lane, two, lane three and then it kind of just shuts off 'cause… It's like, oh, I've never seen this." **CR2.3=[Why X vs. Y] Contrast=not comprehending fire truck at 45 deg angle in next lane vs. regularly parked/stopped vehicle seen from back About contrasting to very specific way that first responders park at 45 degree angle at active collision scene and blocking more than one lane-- is AI programmed to recognize such objects?** Quote: "What about the camera system? Like you know the… you know how the lens or like the how far it can project?" **CR1=Questions how far / distance-wise the camera system can pick up objects** Quote: "...They even say the natural,… natural person when they're driving, when they see lights and sirens, their eye brings them to that area, and they drive into the cars. That's why we're taught to be safe when we're doing it. But I'm wondering is… in is it programmed in AI Autopilot and what does it tell it to do if there's lights and sirens on the road? Quote: " **Questions program whether trained to determine what to do with vehicles that have lights and sirens on the road.** **CR2.3=Non-common Effect (Actual vs. Something Else) [Why X vs. Y] & CR2.2=Normal Case (Actual vs. default value)=[Why X vs. Y] Contrast=ability to recognize flashing lights and sirens vs. other normal conditions on road** **Also compares to humans having tendency to drive towards lights where their eye focuses on** **CR1=Abductive Questions program whether trained to determine what to do with vehicles that have lights and sirens on the road. -- Is AI programmed to recognize such objects?**
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	26	Quote: "...how did it miss the car stopped in the side of the road if there are emergency vehicles with the lights on..." **CR2.1 (Non-occurrence of Effect) & CR2.6 Designed/Specified/Expected Case [Why X vs. not X] Contrast=not see car stopped on side of the road with emergency vehicle lights on vs. see** Quote: "...at what point did the software notice … I'm assuming this is going to be like a lane is blocked? And at what point did it notice that?" **CR1=Abductive reasoning questions when it noticed something was up ahead & CR9=Mechanistic Causal Reasoning: questions when it noticed something was up ahead**

5.2.6 Analysis of FGCR Task#2 – Stakeholder Consultation for Identification of Explanans for Multiple Stakeholders for AI's Act of System Disengagement

In this section, I am interrogating and presenting an aggregated analysis of the transcripts of multiple stakeholders as it pertains to the explanatory information sought by them for the specific act of system disengagement performed by the AV. I will present the assessment of FGCR Task#2, Stakeholder Consultation for Identification of *Explanans*, in this section, and corresponding assessments of FGCR Task#3 in section 5.3.7 and FGCR Task#4 in section 5.3.8.

To illustrate the functionality of FGCR's Task#2, I am presenting an analysis of the transcripts for the specific act of system disengagement performed by the AV in Table 11 below. Note that to support the discussion in this section, I am presenting an analysis of causal reasoning types and descriptors of information sought by multiple stakeholders including verbatim quotes in Table 11 below.

This analysis evaluates the contextual configuration of the same phenomenon (series of car crashes) that needs an explanation, multiple *explainees*, and the same action (system disengagement) performed by the AI. This contextual configuration serves as a comparator to the analysis for the separate AI action of perception. The results in this section also support the claim in Chapter 1 that explanations are contextual where the multiplicity of explanations sought is driven by the fact that different *explainees* will seek different and (sometimes multiple) explanations for the same phenomenon, and within that, the same action performed by the AI system. Similar to the act of perception as analyzed above, this holds true even when examining the data from the vantage point of the AI's actions. The results in this section also provide confidence in the ability of the FGCR to tackle the objective of identifying an appropriate xAI solution regardless of the contextual combination of factors including different AI models performing different AI actions with different and multiple *explainees* seeking explanations for the same phenomenon.

As outlined in section 5.2.2's Table 6, 17 out of 26 stakeholders asked for explanatory information about the act of system disengagement. These stakeholders posed several different queries including:

1. Why the AV disengaged from and aborted control of the task of driving – All 17 stakeholders asked for explanatory information about this.
2. Why the AV disengaged within a very short timeframe prior to the collision (<1 second on average) – 15 stakeholders asked for explanatory information about this.
3. Why the AV disengaged as opposed to other defensive actions such as motion control of braking or steering away – 4 stakeholders asked for this type of explanatory information.
4. Why the AV disengaged as opposed to communicate with the human driver with enough time for them to react – 3 stakeholders asked for this type of explanatory information.

Under the "what" category of explanation, the explanatory information sought by all 17 stakeholders were:

- phenomenon type P1=material causes
- sub-type M1.4=device fails to engage/disengage

Also, some stakeholders asked about the following sub-types in addition to M1.4:

- M1.3=device failure in decision making
- M1.5=device communication failure
- M1.6=device failure in information processing
- M1.7=device fails to perform/fulfill primary function
- M2=failure due to design flaw

Also, some stakeholders articulated their questions in anthropomorphic terms and therefore, I categorized the explanatory information sought by them as:

- P2=behavioral unintentional causes
- sub-type behavioral individual cause BIIR3= intentions

Even though they used anthropomorphic phrases such as "why didn't you. . .", when looking at the overall context and thrust of the conversation, it is evident that they knew they were talking about a non-human entity – a "thing" and not a person. The anthropomorphic language appears to serve as accessible shorthand to describe complex technical processes rather than implying that the system truly possesses human qualities.

Under the "how" category of explanation, the questions above and their subsequent contextual comments involved 11 stakeholders asking for mechanistic causal reasoning information while 12 asked for information in contrastive causal reasoning terms. Note, there were overlaps between these two where some stakeholders asked for both. Table 10 provides detailed causal reasoning sub-type designations per the SSH-based faceted taxonomies about the explanatory information sought by the stakeholders with verbatim quotes. The specific mechanistic causal reasoning questions that stakeholders asked were varied, but they were seeking answers about the inner workings of the AI's ". . . decision making pathway. . ." or ". . . logic. . .". They were all seeking answers along the lines of how the system made its decision to disengage. The specific contrastive comparisons also varied by stakeholder but centered around the themes of the following two comparative questions: "Why the AV disengaged so late versus sufficient time for the human driver to react?" or "Why the AV disengaged versus brake or steer away?"

And, when one takes a bird's eye view of the overall thrust of the questions as they pertained to all AI actions, the majority of stakeholders were asking for an explanation with respect to the failures in the design and training of the AI models themselves. This meta-level inquiry suggests that local, instance-specific explanations may be insufficient because stakeholders are seeking to understand the systemic limitations of the AI models that manifested in these particular failures.

These results of stage 1's data analysis served as inputs to stage 2 of proof-of-concept testing of the FGCR's Task #3, Generation of Detailed xAI Model Requirements), in section 5.3.7, and Task#4, Identification of Potential xAI Models that Could Provide Explanation(s), in section 5.3.8.

Table 11

Aggregate Analysis of Information Sought by Stakeholders for the AI's Act of System Disengagement

	SH CATEGORY	SH#	Failure in Executing Act of SYSTEM DISENGAGEMENT Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
Engineer (Non-AI Expert)	Engineer (Non-AI Expert)	1	Quote (Engineer 1443):... I don't understand how a second before it realises that there is a problem and try to give control to the driver by aborting the the control." Quote (Engineer 5404): "And why the conclusion of this interpretation was just to.. to give up?" (Referring to the AI disengaging). **CR4=Abnormality – late (<1 second before crash) disengagement is abnormal and decision to abort when faced with a problem is also abnormal** **Also CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
	Engineer (Non-AI Expert)	3	Quote (SH3_Engineer 6:16): "Why are you releasing to humans? Aborting to humans just before a crash...Is there a back door for liability...?" **CR4=Abnormality (Late disengagement is abnormal)** **Could also be CR1=Abductive- questioning whether program is designed to abort for liability reasons or [Why X vs. Expected X] Contrastive CR2.6 Designed/Specified/Expected Case =Why disengage before crash (X) vs. not disengaging (Designed/Expected)** **Distal when questioning designers' intentions and goals.**
	Engineer (Non-AI Expert)	4	Quote (Engineer 7:01): I'd like to know when the vehicle sensed there was another vehicle. Quote (Engineer 7:37):Did it sense that there was something there before it dropped? Quote (Engineer 8:36): "...at is it the response the dropping the autopilot is that a response. . . an immediate response or was the response you know a second or two before that, and then it took the system a second or two to to to say "oh OK, it's time to drop and give it to the driver to take over?" **CR1=Abductive = seeking multiple hypotheses regarding;** **-timeframe of decision to disengage - immediately before crash or a second or two before crash? This could also be contrastive [Why X vs. Y?] CR2.3 Non-Common Contrast=Why disengage <1 second before crash (X) vs. disengage ~1-2 seconds before crash (Y)** **- did it sense (or not) that there was an obstacle before it disengaged? This could also be contrastive [Why X vs. not X?] CR2.1 Non-Occurence Contrast=Why sense obstacle (X) vs. not sense obstacle before it disengaged** **Could also be CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
	Engineer (Non-AI Expert)	7	Quote (Stakeholder7 5:43): "...I would want to know: Why the Autopilot didn't notice this potential risk earlier?" **[Why X vs. not X] CR1.2 Non-occurence Contrast=Why timing of recognizing potential risk earlier (X) vs. too late (<1 second before crash)** Quote (Stakeholder7 7:44): "Why did it decide to simply disengage rather than engaging in emergency braking, for instance?" **Contrastive [Why X vs. Y] CR2.3 Non-Common Contrast=Why disengagement (X) vs. braking (Y)** **or Contrastive [Why X vs. default value for X?] CR2.2 Normal Contrast=Why disengagement (X) vs. normal case of emergency braking** Quote (Stakeholder7 8:47): "I guess emergency steering would also be an option." **Contrastive [Why X vs. Y] CR2.3 Non-Common Contrast=Why disengagement (X) vs. steering away (Y)or Contrastive [Why X vs. default value for X?] CR2.2 Normal Contrast=Why disengagement (X) vs. normal case of emergency steering away** Quote (Stakeholder7 8:56): "And why that wasn't considered as an option?" Quote (Stakeholder7 15:44): "...I want to know what were the different options that the AI considered before disengaging immediately before the ...the impact? What other options were considered? When were they considered? And why? Why did it choose one option over another?" **CR1=Abductive – seeking hypotheses about options the AV considered before making decision to disengage** **Could also be CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
	Engineer (Non-AI Expert)	9	Quote (Stakeholder9 00:03:27): "Why did you not take the actions quickly with more lead time?" **Contrastive CR2.1 [Why X vs. not X] Contrast=timing of taking action earlier (X) vs. too late (<1 second before crash)** Quote (Stakeholder9 00:33:57): "Not brake. Not even slow down. You're so slowing down at all?" **Contrastive CR2.3 [Why X vs. Y] Contrast=disengagement (X) vs. braking (Y) or** **Contrastive CR2.2 [Why X vs. default value for X?] contrast=disengagement(X) vs. normal case of emergency braking** Quote (Stakeholder9 00:34:18): "Because it's not. .,, It's not at the problem yet." (Implying a flawed definition or perception of "the problem"). Quote (Stakeholder9 00:34:46): "Yeah. So it's got, it's got more confidence in its ability to kick over to a driver and a driver's response time that is realistic." (Questioning this confidence in driver's ability to take over quickly). **CR1=Abductive – Attempting to identify multiple possible causes of failure**

	SH CATEGORY	SH#	Failure in Executing Act of SYSTEM DISENGAGEMENT
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
AI developer	AI developer	2	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
	AI developer	13	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
	AI developer	16	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	12	Quote (Stakeholder12 00:06:52): "...why did it not disengage and gives some alarms to say there's an object and apply the brakes? Like the sensors should have caught that and it should have interpreted it to do the right actions." Quote (Stakeholder12 00:07:49): "...was stuck in some sort of loop before it could get to that decision? Was there too many factors that it had to analyze, say before it could come to the right reaction?" Quote (Stakeholder12 00:08:18): "...what it is… is the logic or this… or the software decision making when it decides OK, if this then do this." Quote (Stakeholder12 00:15:04): "So what is its decision pathway? OK. Because obviously, if it took too long, it's getting caught up in some sort of loop or, umm, branching off too far. And can you simplify its process so it can come to the right decision quicker?" **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain** **Could also be CR1=abductive -- Seeking multiple hypothesese regarding decision to keep driving –- was it failure to process information? Was it cognitive load of algorithm with too many factors to analyze? Was the logic correct? Seeking information about the decision pathway.** **Could also be Contrastive [Why X vs. default value for X] CR2.2 = Normal case contrasting disengagement 9X)vs. how a competent driver would behave or** **Contrastive [Why X vs. designed/Expected Value for X] CR2.6 contrasting disengagement (x) vs. Autopilot's "normal" or expected operation** **Focus on the algorithm's internal logic, its process for deciding on an action (or inaction) after perception, and its reaction capability. Failure to disengage properly by communicatuing decision to disengage.** **The core inquiry is about how the Autopilot system failed in its decision-making process, leading to the crashes.**
	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	6	Quote (Stakeholder6 7:40): "I guess the first thing that comes to my mind is how the system is making its decision to disengage the Autopilot." Quote (Interviewer summarizing Stakeholder6 49:15): "So initially your first question was… the seconds before it disengages, "Why did it disengage?" was your first question." **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
	Driver of Automated Vehicle (AV) (e.g. Tesla, Waymo, Cruise, etc.	8	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy	23	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy	24	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy	25	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
	First Responder (e.g. Police, Ambulance, Fire, search and rescue, heavy equipment operators, etc.)	26	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**

	SH CATEGORY	SH#	Failure in Executing Act of SYSTEM DISENGAGEMENT
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
Driver of other non-AV vehicle on the road	Driver of other non-AV vehicle on the road	5	Quote (Stakeholder5 10:49): "Why one second?" (Questioning the timing of the abort). Quote (Stakeholder5 11:11): "Why one second? Yeah, I mean exactly if you because . . . if you think categorically it is not a vehicle on the road, you will need to keep driving until you crash. We don't have the abort the vehicle about the control." Quote (Stakeholder5 12:59): "So why did you stop driving a second before?" (Questioning the decision to stop only a second before). **[Why X vs. default value for X] CR2.2 Normal Contrast: Why disengage vs. normally, AVs are trained not to hit objects, implying earlier/avoidance action** **Could also be CR3=Counterfactual (implicitly, "if it were like a human/properly trained, it would have acted differently/earlier")** **Could also be CR4=abnormality: Comparing AI behavior to human behavior and its training. The stakeholder is trying to understand the "reasoning" or "cause" for the specific timing of its final action (Malle, 1999; Miller, 2019). The explanation sought is for the abnormality of the late disengagement timing (Hilton, 1990).** **Could also be CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
	Driver of other non-AV vehicle on the road	15	Quote (Stakeholder15 00:06:48): why isn't there a safeguard in that .. it alarm… it alerts the human? So obviously it didn't, you say one second before. There's obviously something there before and it just stops right before impact? One second? No, no, it stops, meaning like the autopilot turns off. **[Why X vs. Expected X] CR2.6 Contrastive Designed/Specified/Expected Case: Why X =it disengages <1 second before impact without warning vs. expected warning**
	Driver of other non-AV vehicle on the road	17	Quote (Stakeholder17 00:09:22): "I'm thinking that you have one second, did you say? To alert." (Focusing on the minimal time given). Quote (Stakeholder17 00:09:28): "I don't think you can prevent that accident in one second. And that's the case then you really can't trust the AV." Quote (Stakeholder17 00:11:54): "Why did it stop functioning? A second before the accident." Quote (Stakeholder17_NonAV Driver 00:11:59): A second before the accident. And, how it has to notify the person in the driver's seat to give them more time to take over if, … if the… the AV system is stopping. Quote (Stakeholder17 00:12:28): "why it didn't do it sooner." **[Why X vs. Designed/Specified/Expected X] CR2.6 Contrastive= Why disengage <1 second before accident (X) vs. longer lead time to both notify driver and disengage** or **[Why X vs. ideal value for X] CR2.5 Contrastive=Why disengage <1 second before accident (X) vs. longer lead time to both notify driver and disengage**
	Driver of other non-AV vehicle on the road	20	Quote (Stakeholder20 00:09:38): "..… or give it back to the driver like a millisecond before, like it wasn't enough time, whatever that time is obviously off. Like it was going to give …Autopilot realized there was a problem and needed to give it back to the driver, one second before isn't going to cut it. So that has to be looked at in the program. **Questioning actual time as opposed to sufficient time to release control and give back to driver [Why X vs. ideal value for X] CR2.5 Ideal Contrast=Why disengage<1 second before crash vs. more lead time** **Also CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
OTHER	Policy maker	11	Quote (PolicyMaker 9:04): "...The one second that they need or is it 16 seconds they need in order to engage the OR disengage the autopilot?" Quote (PolicyMaker 13:11): "...the car made the decision to release control. So why did they? Why did it do that? What was what was in the… its,… its AV brain...why did it release?" Quote (PolicyMaker 13:31): "Because you said it released like less than a second before the impact. So why did it release? What? What triggered it to release?" Quote (PolicyMaker 15:06): "...So what? What made it released? So it made some kind of a decision. Based on all this information that they had. So what was that information?" **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain**
OTHER	Pedestrian	10	Quote (Stakeholder10 00:03:27): "'Why did you not take the actions quickly with more lead time?'" Quote (Stakeholder10 00:03:42): "...why was there only one …? Is it because… it, Autopilot didn't detect it until a second before? Or it was going so fast that it didn't detect the issue until a second before?" Quote (Stakeholder10 00:05:13 onwards, implicitly): The discussion about human driver distraction leading to late response is analogously applied to the AI's late response/disengagement. **[Why X vs. Y] CR2.3=non-common effect contrast=focuses on multiple differentiating factors: Why release with only 1 second? (X) vs. longer lead time (Y) OR Because it didn't detect object until a second before (X) vs. detecting object before < 1 second (Y) or was it going too fast to react (X) vs. slow enough to react (Y)** **Also could be [Why X vs. default value for X] CR2.2-Normal case = comparing AV not taking any action (x) vs. what a human that is distracted would end up doing** **Also could be CR1=ABDUCTIVE by positing multiple hypotheses**

	SH CATEGORY	SH#	Failure in Executing Act of SYSTEM DISENGAGEMENT
			Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought
OTHER	Insurance Industry	14	**NO QUESTIONS ABOUT SYSTEM DISENGAGEMENT**
Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	18	Quote (Stakeholder18_Lawyer 7:09): So it's not it. I… t seems to me that it's not something that was not tuned enough if it was a breakdown in the system, so it's not it. So it's we're not looking at just a little tweak. We're talking about something that was working from 100% to 0%. Quote (Stakeholder18 00:14:17): "OK. Yeah. No, it's …it's… it's…. it's just absolutely amazing that …That it did not stop, for one thing, when clearly the road was blocked." Quote (Stakeholder18 00:18:17): "So this is going from the what… that's what I mean going from 100% so-called intelligence to 0 intelligence." (Referring to its failure to act appropriately until the last second). Quote (Stakeholder18 00:18:29): ...And the fact that it that it only released a second prior is also puzzling, because you would expect either to not release at all because clearly was not no longer functioning properly, but how could it actually waited… waited a second when it was obvious that the crash was going to happen? Quote (Stakeholder18 00:19:00): "So again, it's puzzling that even released one second prior." Quote (Stakeholder18 00:05:30): "So immediately my … my thought is OK, something triggered that malfunction, and I… I would like to know the nature of it." Quote (Stakeholder18_Lawyer 19:40): And why would it release if it if it didn't recognise, right? If it didn't recognise, then why release? If it recognised then it released because it recognised and didn't know it, couldn't it? And it knew it couldn't do anything to avoid it. So let it go. **CR1=Abdictove & could also be CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain. Seeking information about the breakdown as well as the timing of disengagement being so late and what triggered the breakdown. Also questioning logic of decision to disengage – if it didn't recognize obstacles ahead, then why did it disengage --- is it because it didn't know what the object was or couldn't do anything to avoid it?** Quote (Stakeholder18_Lawyer 21:33):What I would want to know is what was this car thinking 5 seconds or ten seconds prior? **[Why X vs. ideal value for X] CR2.5 contrast= ideal case of releasing 5-10 seconds prior as opposed to <1 second (X)**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	19	Quote (Stakeholder19 00:15:46): "...if there is an inherent design flaw. Or if there is a weakness in the system that manifests itself by disengaging one … one second before reaching a first responder vehicle..." **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain.** Quote (Stakeholder19 00:22:15):I don't understand how it is that the whole autopilot disengaged? So is there, I'll call it a manufacturing defect, because of lack of better knowledge, but is there a defect in the motion control where a confused putting on the brake to disengaging the whole autopilot? **CR1=Abductive -- potential hypotheses for failure** (confusing brakes with disengagement mechanism?)
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	21	Quote (Stakeholder21 :28:26): And I would also be very interested in knowing more like your scenario adverts to the fact that… that the Autopilot can disengage and hand over controls to the human driver. Quote (Stakeholder21 29:45): It can disengage and ask the human to take over, so I would want to know how that works, how it communicates to the human that it's going to do so. How the… like how its internal mechanisms work or how its levels are set to … to trip the threshold of when it's supposed to do that because you know, if the idea is to hand control back to the human once it gets into trouble, it did that too late here, right? Like, I know in driving situations, one second can be a long time, but, … Quote (Stakeholder21 29:45): I... if the idea is to hand control back to the human once it gets into trouble, it did that too late here, right? Like, I know in driving situations, one second can be a long time, but, … Quote (Stakeholder21 30:32):Probably one second prior to the first impact wasn't enough time for the human to react and avert the accident, and so I would want to know more about. **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain** **Could also be CR1=abductive -- Looking at multiple causes in decision to disengage**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	22	Quote (Lawyer 14:13): "...Like major major issues in regards to …to the release of control. How come the car did not stop then?" **[Why X vs. what ought to be X] CR2.4=prescribed case = Evaluates actions against normative/prescriptive standard of stopping vs. continuing driving and disengaging (X)** **Could also be CR8=Dennett's Process Narrative**

	SH CATEGORY	SH#	Failure in Executing Act of SYSTEM DISENGAGEMENT	Failure in Executing Act of SYSTEM DISENGAGEMENT
			Causal Reasoning Type (CR) & Descriptors of Information Sought	**Quotes of Explanatory Information Sought by Explainee(s) with highlighted Causal Reasoning Types & Descriptors of Information Sought**
Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	18	**CR1=Abdictove & could also be CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain. Seeking information about the breakdown as well as the timing of disengagement being so late and what triggered the breakdown. Also questioning logic of decision to disengage -- if it didn't recognize obstacles ahead, then why did it disengage --- is it because it didn't know what the object was or couldn't do anything to avoid it?** **[Why X vs. ideal value for X] CR2.5 contrast= ideal case of releasing 5-10 seconds prior as opposed to <1 second (X)**	Quote (Stakeholder18_Lawyer 7:09): So it's not it. I… t seems to me that it's not something that was not tuned enough if it was a breakdown in the system, so it's not it. So it's we're not looking at just a little tweak. We're talking about something that was working from 100% to 0%. Quote (Stakeholder18 00:14:17): "OK. Yeah. No, it's …it's… it's…. it's just absolutely amazing that …That it did not stop, for one thing, when clearly the road was blocked." Quote (Stakeholder18 00:18:17): "So this is going from the what… that's what I mean going from 100% so-called intelligence to 0 intelligence." (Referring to its failure to act appropriately until the last second). Quote (Stakeholder18 00:18:29): ...And the fact that it that it only released a second prior is also puzzling, because you would expect either to not release at all because clearly was not no longer functioning properly, but how could it actually waited… waited a second when it was obvious that the crash was going to happen? Quote (Stakeholder18 00:19:00): "So again, it's puzzling that even released one second prior." Quote (Stakeholder18 00:05:30): "So immediately my … my thought is OK, something triggered that malfunction, and I… I would like to know the nature of it." Quote (Stakeholder18_Lawyer 19:40): And why would it release if it if it didn't recognise, right? If it didn't recognise, then why release? If it recognised then it released because it recognised and didn't know it, couldn't it? And it knew it couldn't do anything to avoid it. So let it go. **CR1=Abdictove & could also be CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain. Seeking information about the breakdown as well as the timing of disengagement being so late and what triggered the breakdown. Also questioning logic of decision to disengage -- if it didn't recognize obstacles ahead, then why did it disengage --- is it because it didn't know what the object was or couldn't do anything to avoid it?** Quote (Stakeholder18_Lawyer 21:33):What I would want to know is what was this car thinking 5 seconds or ten seconds prior? **[Why X vs. ideal value for X] CR2.5 contrast= ideal case of releasing 5-10 seconds prior as opposed to <1 second (X)**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	19	**CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain.** **CR1=Abductive -- potential hypotheses for failure (confusing brakes with disengagement mechanism?)**	Quote (Stakeholder19 00:15:46): "...if there is an inherent design flaw. Or if there is a weakness in the system that manifests itself by disengaging one … one second before reaching a first responder vehicle..." **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain.** Quote (Stakeholder19 00:22:15):I don't understand how it is that the whole autopilot disengaged? So is there, I'll call it a manufacturing defect, because of lack of better knowledge, but is there a defect in the motion control where a confused putting on the brake to disengaging the whole autopilot? **CR1=Abductive -- potential hypotheses for failure** (confusing brakes with disengagement mechanism?)
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	21	**CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain** **Could also be CR1=abductive -- Looking at multiple causes in decision to disengage**	Quote (Stakeholder21 :28:26): And I would also be very interested in knowing more like your scenario adverts to the fact that… that the Autopilot can disengage and hand over controls to the human driver. Quote (Stakeholder21 29:45): It can disengage and ask the human to take over, so I would want to know how that works, how it communicates to the human that it's going to do so. How the… like how its internal mechanisms work or how its levels are set to … to trip the threshold of when it's supposed to do that because you know, if the idea is to hand control back to the human once it gets into trouble, it did that too late here, right? Like, I know in driving situations, one second can be a long time, but, … Quote (Stakeholder21 29:45): I... if the idea is to hand control back to the human once it gets into trouble, it did that too late here, right? Like, I know in driving situations, one second can be a long time, but, … Quote (Stakeholder21 30:32):Probably one second prior to the first impact wasn't enough time for the human to react and avert the accident, and so I would want to know more about. **CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain** **Could also be CR1=abductive -- Looking at multiple causes in decision to disengage**
	Legal Expert (e.g. lawyer, judge, lawmaker, etc.)	22	**[Why X vs. what ought to be X] CR2.4=prescribed case = Evaluates actions against normative/prescriptive standard of stopping vs. continuing driving and disengaging (X)** **Could also be CR8=Dennett's Process Narrative**	Quote (Lawyer 14:13): "...Like major major issues in regards to …to the release of control. How come the car did not stop then?" **[Why X vs. what ought to be X] CR2.4=prescribed case = Evaluates actions against normative/prescriptive standard of stopping vs. continuing driving and disengaging (X)** **Could also be CR8=Dennett's Process Narrative**

5.3 Stage 2 of FGCR Testing

In this section, I will summarize the results of proof-of-concept testing of FGCR Tasks #3 (Generation of Detailed xAI Model Requirements) and #4 (Identification of potential xAI Models that Could Provide Explanation(s)). This stage of testing uses the outputs of FGCR Task#2 from stage 1 of the testing and applies them to these subsequent FGCR tasks.

The sections below that describe the results of the FGCR's proof-of-concept testing include:

- An analysis of one individual stakeholder's (Stakeholder #1) interview for multiple AI actions (sections 5.3.1 - 5.3.2)
- An analysis of another individual stakeholder's (Stakeholder#25) interview for multiple AI actions (sections 5.3.3 - 5.3.4)
- An analysis of multiple stakeholder interviews of one AI action – perception (sections 5.3.5 - 5.3.6)
  An analysis of multiple stakeholder interviews of another AI action – system disengagement (sections 5.3.7 - 5.3.8) – note, these latter two analyses are disaggregated by AI action but aggregated by multiple stakeholders. i.e. explanatory information sought by multiple stakeholders (aggregated) for two individual and separate AI actions (disaggregated).

The outputs of stage 2 testing will include detailed xAI model requirements as well as identification of potential xAI models that could satisfy the explanatory needs of stakeholders. The outputs for Task #3 will include:

- a description of questions asked by stakeholders with direct quotes if applicable
- classification of explanations sought per the SS-based taxonomy if applicable
- identification of AI application, AI model, or AI actions taken or decisions made by the AI system that requires an explanation

And for Task #4, the outputs will include:

- determination of the type of explanation problem: global, local or model inspection
- potential xAI models that could provide explanations

5.3.1 Analysis of FGCR Task#3 – Generation of Detailed xAI Model Requirements for Stakeholder#1

To illustrate the functionality of FGCR Task#3, Generation of Detailed xAI Model Requirements, I am presenting the analysis for Stakeholder #1 (SH#1). I will demonstrate how to generate detailed xAI model requirements for the *explanans* identified by FGCR Task#2 and presented

in section 5.2.3 above. I will summarize the detailed classifications of the explanatory information sought by *explainees* per the SSH-based taxonomies as well as groupings by action performed by the AI into a set of xAI model requirements. Note that these can be articulated in multiple formats (textual; tables; databases; etc.) allowing xAI developers the flexibility to choose the format that best suits their needs. For the purposes of this research project, I am presenting the set of requirements in textual format as follows:

(1) The xAI model shall provide explanation(s) for the AI's act of perception by answering the main question of why the AI failed to perceive stationary objects.

Specifically, the xAI model shall provide answers in the form of contrastive (i.e. comparative or differentiating factors) explanations for the following queries posed by SH#1:

a. Why did the AI not recognize or perceive first responder vehicles as opposed to. other types of vehicles?
b. Why did a system that avoided (including the task of perception) many obstacles prior to these accidents as opposed to not doing so in these 16 circumstances?

An important point to note here is that the detailed requirements as stated above do not use SSH-based contrastive case type terminology. For the purposes of requirements validation and inclusion of requirements into the specification for xAI model identification and/or development, a high level of precise and detailed information is a necessary. Therefore, the focus in this task within the FGCR is stating highly specific differentiating information sought by the stakeholder and not SSH-based terminology. Note that for facilitating the FGCR task of stakeholder consultation especially during the iterative task(s) providing explanatory information, the SSH-based terminology within the how category of explanation will be useful for quickly identifying causal reasoning types and communicating explanations in that form.

(2) The xAI model shall provide explanation(s) for the AI's actions of steering and braking: The xAI model shall provide answers with respect to the AV's failures with respect to decision making and taking the appropriate action of avoiding obstacles ahead.

Here again, the xAI model shall provide answers in the form of contrastive (i.e. comparative or differentiating factors) explanations for why the AV did not change direction nor brake to avoid the

obstacles. And more specifically, answer the question of why the AV avoided obstacles in the past vs. not avoiding obstacles in these 16 cases.

(3) The xAI model shall provide explanation(s) for the AI's actions of communication/warning to humans and system disengagement decision:

a. The xAI model shall provide answers with respect to the AV's decision to disengage a second before the crash.
b. The xAI model shall provide answers with respect to the AV's realization of the problem at that specific moment and decision to disengage rather than notifying the driver of a problem.

Specifically, why the AV aborted control a second before the crash with an implication that the AV "realized" there was a problem at that point and rather than communicate this to the driver, and instead, chose to abort control.

(4) The xAI developers shall provide global model explanatory information as it pertains to each AI action queried above.

This is due to the fact that underlying all of the above requirements, SH#1 is seeking an explanation for the design and training of the AI model and an explanation for the mismatch between training and reality.

As outlined above, the heterogeneity of complex AI systems with multiple AI models performing multiple AI tasks combined with the contextuality of explanations is driving the need for multiple explanations. This analysis for SH#1 shows that even for the same *explainee,* multiple explanations are needed for the same phenomenon (series of car crashes) which depend upon the task performed by the AV's multiple AI systems. These results support the claim in Chapter 1 that there will not be one explanation that will meet the explanatory needs of individual stakeholders. These outputs of FGCR Task#3 analysis will now serve as a set of xAI model requirements that will feed into the assessment of FGCR Task#4 in section 5.3.2 below.

5.3.2 Analysis of FGCR Task#4 – Identification of Potential xAI Models for Stakeholder#1

To demonstrate the FGCR's Task#4, Identification of Potential xAI Models that Could Meet the Explanatory Needs of SH#1, I will attempt to match the requirements in section 5.3.1 above with the taxonomy of the xAI landscape. Note that, rather than analysing all four actions performed by AI, this section demonstrates the feasibility of the proposed FGCR Task#4 by conducting a full analysis for the act of perception only, with the assumption that the AI model is a black box artificial neural network (ANN).

For the act of perception, one could initially point out that SH#1 is seeking a local explanation for the moments leading up to the crash. Here, as described in section 3.3 above, there are various existing xAI models available for use by the AI developers: LIME, SM, CAM, Grad-CAM, and VisualBackProp models may be able to provide an explanation by identifying objects of the scene ahead on a second-by-second snapshot of the video. However, because these xAI models provide visual representations that highlight regions of an image that are particularly influential in the final classification of an image, they do not provide information in contrastive terms. In this case, because they do not explicitly provide contrastive information, the *explainee* would very likely have to form their own conclusions about the factors that led to the failure of perception. Therefore, they may not sufficiently provide adequate explanations that answer most of the contrastive questions asked by SH#1. SH#1's specific contrastive queries pertained to why the AV did not perceive or recognize first responder vehicles versus other types of vehicles and obstacles and why it did not recognize and therefore avoid crashing into many other obstacles in its history of driving versus in these 16 circumstances.

To address the requirements of meeting these specific contrastive queries, two other explanators, a prototype selection (PS) explanator or a pointing and justification-based explanation (PJ-X) explanator might suffice. A PS provides examples of similar scenarios to the predicted outcome that make explicit

which criteria informed the classification decision and a PJ-X, a multi-modal explanator for visual decisions provides explanations for an image by simultaneously highlighting the relevant parts and generating a textual justification. Both could complement the visual representation xAI models listed above and potentially provide information such as "following are past scenarios where the system successfully identified emergency vehicles—note the distinguishing features that differed between previous scenarios and these failure cases".

If these xAI models do not suffice, perhaps other model inspection or global model explanators might be appropriate. A saliency mask (SM) explanator might point out the saliency of individual features that cause a certain outcome. A feature importance (FI) explanator might quantify the weight and magnitude of inputs features that are used to produce the results of the AI model. Or a partial dependence plot (PDP) explanator might provide insights into the relationship between one or more input variables and the output of the AI model. All of these xAI models might complement and augment the explanations provided by the others while addressing SH#1's queries about the design and training of the AI system.

These latter technical xAI models (SM, FI, PDP) produce outputs that require translation by human AI experts into understandable layperson terms for non-AI expert stakeholders. The FGCR process takes this limitation of technical xAI models into account by including Task#5, Usage of Existing xAI Models to Answer Stakeholders/*Explainees'* Questions via Consultation. Therefore, it is crucial that xAI developers perform the iterative tasks of consulting with the stakeholder(s) by providing them with explanations and interactively getting to agreement that they meet the stakeholder's explanatory needs. A key point to note here is that regardless of the explanation produced by the FGCR process, ultimately, it is up to the *explainee* to determine whether they have received an adequate explanation from both the xAI model and their developers.

The outputs of FGCR Task#4 analysis include a list of potential xAI models that could provide explanations to SH#1 for only one AI action (perception). This analysis shows that even when

narrowing down the focus to one AI action (perception) as it pertains to the same phenomenon that needs an explanation (series of car crashes), the same *explainee* will not only seek multiple explanations, but will require multiple xAI solutions.

The need for multiple xAI solutions is driven by the need to consider the complex combination of contextuality of explanations and the heterogeneity of AI systems. Contextuality requires providing different explanations for the same phenomenon, in the same situation to the same *explainee*. And within the same AI domain application, we had multiple AI models performing multiple tasks (e.g. perception, steering, braking, system disengagement, etc.) that required different explanations for the same *explainee*. Furthermore, it isn't only that different AI models require different explanations, but also, for the outputs of the same AI domain application, and within that, the same AI model performing the same task (perception), multiple xAI models will be required for the same phenomenon and/or task to be explained in the same situation.

5.3.3 Analysis of FGCR Task#3 – Generation of Detailed xAI Model Requirements for Stakeholder#25

To illustrate the functionality of FGCR Task#3, Generation of Detailed xAI Model Requirements, I am presenting the analysis for Stakeholder#25 (SH#25). I will demonstrate how to generate detailed xAI model requirements for the *explanans* identified by FGCR Task#2 and presented in section 5.2.4 above. I will summarize the detailed classifications of the explanatory information sought by *explainees* per the SSH-based taxonomies as well as groupings by action performed by the AI into a set of xAI model requirements. Based on the outputs of Task#2, following are the set of requirements in textual format:

(1) The xAI model shall provide explanation(s) for the AI's act of perception by answering the questions pertaining to why the AI failed to perceive stationary objects at the crash scene(s).

Specifically, the xAI model shall provide answers in the form of contrastive (i.e. comparative or differentiating factors) explanations for the following queries posed by SH#25 (quoted verbatim):

a. … is the AI programmed to see lights and sirens and what that means to them, to the programming?

For this query, the stakeholder was asking about whether the AV is capable of perceiving and recognizing vehicles with flashing lights and sirens as opposed to vehicles or other road objects with regular lights.

b. . . . And then basically, when firefighters and police officers go to a scene, they're taught to go 45° and take up…. An AI computer can't comprehend if a fire truck is in a 45° angle in lane one, lane, two, lane three and then… it kind of just shuts off 'cause… It's like, oh, I've never seen this. . .

For this query, the stakeholder was asking about whether the AV is capable of perceiving and recognizing first responder vehicles parked at 45° angles and occupying several lanes at emergency scenes as opposed to typical front/back facing vehicles parked on sides of roads.

(2) The xAI model shall provide explanation(s) for the AI's actions of motion control, in particular, braking: The xAI model shall provide answers with respect to the AV's failures with respect to decision making and taking the appropriate action of braking and stopping prior to crashing into the obstacles ahead.

Here again, the xAI model shall provide answers in the form of contrastive (i.e. comparative or differentiating factors) explanations for why the AV did not brake to avoid crashing into the obstacles ahead. The xAI model shall provide explanations for the following questions posed by SH#25 (quoted verbatim):

a. . . . did you see the lights, sirens or did you see the lights and did you see the apparatus blocking the roadway and if so, why didn't you stop?

For this query, the stakeholder is asking about the decision-making process and logic of the AV, first by probing whether the AV perceived the flashing lights and sirens of the obstacles blocking the road ahead and if it did, why it didn't stop.

b. . . . I would think there'd be something wrong with the algorithm of why it didn't 'cause obviously it could. I've seen videos where it veers off quickly if someone falls onto the roadway, and the Tesla veers over, but why didn't at this moment is my question. What? What was it doing? What was it computing at that time that it didn't recognise this as a as a "brake" kind of thing or . . .?

For this query, the stakeholder is asking about the comparison to other incidents of Teslas avoiding obstacles ahead as opposed to these cases where it did not even brake.

(3) The xAI developers shall provide global model explanatory information as it pertains to each AI action queried above.

This is due to the fact that underlying all of the above requirements, is the fact that SH#25 is seeking an explanation for the design and training of the AI model. This could also require providing a process narrative identifying necessary and salient components in the causal chain that led to the outcome(s).

As outlined above, the heterogeneity of complex AI systems with multiple AI models performing multiple AI tasks combined with the contextuality of explanations is driving the need for multiple explanations. This analysis for SH#25 also shows that the same individual *explainee* is seeking multiple explanations depending upon the task performed by the AV's multiple AI systems. And when compared to SH#1's analysis, different *explainees* will seek different (and sometimes multiple) explanations for the same phenomenon. These results support the claim in Chapter 1 that there will not be one xAI solution that will meet the explanatory needs of multiple individual stakeholders. These outputs of FGCR Task#3 analysis will now serve as a set of xAI model requirements that will feed into the assessment of FGCR Task#4 in section 5.3.4 below.

5.3.4 Analysis of FGCR Task#4 – Identification of Potential xAI Models for Stakeholder#25

To demonstrate the FGCR's Task#4, Identification of Potential xAI Models that Could Meet the Explanatory Needs of SH#25, I will attempt to match the requirements in section 5.3.3 above with the taxonomy of the xAI landscape. In this section, in addition to providing an analysis for the act of perception, which will serve as a comparator to section 5.3.2's SH#1 results, I am choosing to also analyze the AI action of motion control. Note that for the purposes of this exercise, I will assume that for the actions performed by the AI above, the AI models are black box artificial neural nets (ANNs).

Action Performed by AI: Perception

For the act of perception, one could initially point out that SH#25 is seeking a local explanation for the moments leading up to the crash. Here, similar to the analysis for SH#1 in section 5.3.2 above, there are various existing xAI models available for use by the AI developers: LIME, SM, CAM, Grad-CAM, and VisualBackProp xAI models may be able to provide an explanation by identifying objects of the scene ahead on a second-by-second snapshot of the video. However, because these xAI models provide visual representations that highlight regions of an image that are particularly influential in the final classification of an image, they do not provide information in contrastive terms. And because they do not explicitly provide contrastive information, they may not provide be sufficient to provide adequate explanations that answer SH#25's specific contrastive queries: why the AV did not perceive or recognize first responder vehicles with flashing lights versus other types of lights on other objects on the road and also whether it is capable of recognizing first responder vehicles parked at a 45º angle occupying more than one lane versus other types of vehicles that would be front facing.

To address the requirements of meeting these specific contrastive queries, again, two other explanators: a prototype selection (PS) explanator or a pointing and justification-based explanation (PJ-X) explanator. Note that SH#25's contrastive cases are different from SH#1's and therefore, the same xAI model may not provide adequate explanations to both stakeholders.

Again, if these xAI models do not suffice, perhaps other model inspection or global model explanators might be appropriate. A saliency mask (SM) explanator, a feature importance (FI) explanator, or a partial dependence plot (PDP) explanator might provide insights that could complement and augment the explanations provided by the others while addressing SH#25's queries about the design deficiencies of the AI system.

Action Performed by AI: Motion Control (Steering and Braking)

For the act of motion control, in particular, the braking function, SH#25 is seeking insights about the "logic" or thinking process of the AI. Specifically, assuming that it perceived the obstacle ahead, why the AV didn't brake and stop. This particular explanation sought by SH#25, could be designated as a model explanation or model inspection problem. Here, a feature importance (FI) explanator that quantifies the weight and magnitude of the input features used to make the predictions and which features are important to the output might provide answers to which input features contributed to decision to not brake and stop. A partial dependence plot (PDP) explanator that provides a graphical representation of the relationship between one or more input variables, and the outcome might also provide answers. A more in-depth inspection of the NN could be conducted by using an activation maximization (AM) explanator which could provide information about the input patterns that maximize the activation of a certain neuron in a certain layer with respect to particular inputs. And if none of these suffice, then either decision tree (DT) or decision rule (DR) explanators might be able to provide information about the set of rules that describe the logic behind the black box AI model. Even though there are overlaps in the types of xAI models (FI, PDP) recommended for both the AI tasks of motion control and perception, these xAI models will be different models. This is due to the fact that the underlying AI black box models will be different.

Again, because the xAI models that address model inspection and global model explanation problems (SM, FI, PDP, AM) produce technical outputs that require translation into layperson terms for non-AI expert stakeholders the FGCR process takes this limitation of technical xAI models into account

by including Task#5, Usage of Existing xAI Models to Answer Stakeholders/*Explainees'* Questions via Consultation. Therefore, xAI developers will have to perform the iterative tasks of consulting with the stakeholder(s) by providing them with explanations and interactively getting to agreement that they meet the stakeholder's explanatory needs. And regardless of the explanation produced by the FGCR process, ultimately, it is up to the *explainee* to determine whether they have received an adequate explanation from both the xAI model and their developers.

The outputs of FGCR Task#4 analysis include a list of potential xAI models that could provide explanations to SH#25 for the AI actions of perception and motion control (steering/braking). This analysis shows that not only will the same *explainee* (SH#25) seek multiple explanations, but also, when compared to another stakeholder's (SH#1) analysis, different *explainees* will seek different (and sometimes multiple) explanations for the same phenomenon (series of car crashes). These will require multiple xAI solutions.

The need for multiple xAI solutions is driven by the need to consider the complex combination of contextuality of explanations and the heterogeneity of AI systems. Contextuality requires providing different explanations for the same phenomenon, in the same situation to the same *explainee*. And when compared to another stakeholder, SH#1, different and sometimes multiple) explanations will be required for different *explainees.* For this case study, we had the added complexity of having multiple AI models within the same AI domain application performing (e.g. perception, steering, braking, etc.) that required different explanations for the same *explainee*. Furthermore, it isn't only that different AI models require different explanations, but also, for the outputs of the same AI domain application, and within that, the same AI model performing the same task (perception), multiple xAI models will be required for the same phenomenon and/or task to be explained in the same situation. Therefore, especially for complex AI applications, multiple xAI models and/or solutions will be required because individual *explainees* will seek not only different, but multiple explanations.

5.3.5 Analysis of FGCR Task#3 – Generation of Detailed xAI Model Requirements for AI's Act of Perception for Multiple Stakeholders

To illustrate the functionality of FGCR Task#3, Generation of Detailed xAI Model Requirements, I am presenting the analysis for multiple stakeholders for the AI's act of perception. The above results about the *explanans* sought by multiple stakeholders for the AI's act of perception from section 5.2.5 can be summarized as a set of requirements:

(1) The xAI model(s) shall provide explanation(s) for the act of perception by answering the question of why the AI failed to perceive stationary objects.

More specifically, the xAI model shall provide answers in the form of contrastive explanations for the following specific differentiating factors.

Why did the AV not perceive (recognize/categorize/see/sense) or differentiate between:

- First responder vs. other types of vehicles
- Emergency lights on first responder vehicles that are red/yellow and flashing vs. red non-flashing traffic lights
- First responder large vehicles (such as fire trucks, tow trucks, ambulances) vs. other objects in the distance such perhaps a balloon or a row of lights on a bridge in a distance
- Highly visible large first responder vehicles and people with highly visible vests vs. all other objects
- These particular 16 highly dangerous accident scenes vs. all other similar scenes the AV encountered in the past
- Large fire trucks and ambulances parked at 45-degree angles covering more than one lane vs. regularly positioned vehicles on the road

Two important things to note: These detailed requirements stated above do not use SSH-based contrastive case type terminology as articulated in the analysis of FGCR Task#2 in section 5.2.5. For the purposes of requirements validation and inclusion into an xAI model development specification, as well as identifying existing xAI models, the focus in this task within the FGCR is stating highly specific differentiating information sought by multiple stakeholders and not SSH-based terminology. In contrast, for facilitating the FGCR task of stakeholder consultation especially during the iterative task(s) providing explanatory information, the SSH-based terminology within the how category of explanation could prove to be useful for quickly identifying contrastive reasoning types and communicating explanations in specific differentiating terms.

Also, as stated in previous sections, underlying all of the above requirements, is the fact that most stakeholders are seeking explanations with respect to the design and training of the AI model and explanations for the potential mismatch between training and testing environment and real-world deployment of the AI system. Therefore,

(2) xAI developers shall provide global model explanatory information as it pertains to the task of perception.

These outputs of FGCR Task#3 analysis will now serve as a set of xAI model requirements that will feed into the assessment of FGCR Task#4 in section 5.3.6 below.

5.3.6 Analysis of FGCR Task#4 – Identification of Potential xAI Models for AI's Act of Perception for Multiple Stakeholders

To demonstrate the FGCR's Task#4, Identification of Potential xAI Models that Could Meet the Explanatory Needs for multiple stakeholders, I will attempt to map the requirements in section 5.3.5 above with the taxonomy of the xAI landscape. Again, for the purposes of this exercise, I will assume that the AI model is a black box neural net in all AI actions discussed.

As outlined above in section 5.3.2's analysis of Stakeholder #1, for the task of perception, one could initially point out that the majority of stakeholders are seeking a local explanation for the moments leading up to the crash. And perhaps, if we can apply an analysis by visual representation xAI models (such as LIME, SM, CAM, Grad-CAM, or VisualBackProp) to a second-by-second snapshot of the video leading up to the crash, we may be able to provide an explanation to the stakeholders. However, these xAI models only highlight regions of an image that are particularly influential in the final classification of objects of an image and do not explicitly provide contrastive information. For example, they may highlight areas of the image of the accident with labels such as "lights" without providing further information about the types of lights, colours, whether the lights belonged to objects such as vehicles or people as opposed to streetlights, etc. Because they do not explicitly provide contrastive information, *explainees* will have to form their own conclusions about the factors that led to the failure

of perception. Therefore, it may not provide adequate explanations for the many differentiating factors that are listed above.

In addition to these visual representation xAI models, for explanatory information pertaining to the specific contrastive queries, perhaps a prototype selection (PS) explanator or a pointing and justification-based explanation (PJ-X) could also be used. These could potentially highlight and provide distinguishing features that were different between previous outputs and these particular outputs.

And if the above-mentioned xAI models do not suffice, other model inspection or global xAI models might be appropriate such as saliency mask (SM), feature importance (FI), or partial dependence plot (PDP) explanators. These additional xAI models might augment the explanation provided by the combination of visual representation and contrastive explanator type xAI models.

Also, when considering that the overarching thrust of the queries of most stakeholders was with respect to the design and training of the AI models, xAI developers will need to delve into the training data itself in order to answer some of their questions. Furthermore, the data provided by the stakeholders' list of differentiating factors in section 5.3.5 may actually serve as feedback for AI developers. These differentiating factors may be included as types of features and additional inputs in future training datasets for evaluating and improving the AI models.

Again, because the xAI models that address model inspection and global model explanation problems (SM, FI, PDP) produce technical outputs that require translation into layperson terms for non-AI expert stakeholders the FGCR process takes this limitation of technical xAI models into account by including Task#5, Usage of Existing xAI Models to Answer Stakeholders/*Explainees'* Questions via Consultation. Therefore, xAI developers will have to perform the iterative tasks of consulting with the stakeholder(s) by providing them with explanations and interactively getting to agreement that they meet the stakeholder's explanatory needs. And regardless of the explanation produced by the FGCR process, ultimately, it is up to *explainees* to determine whether they have received an adequate explanation from both the xAI model and their developers.

The outputs of FGCR Task#4 analysis include a list of potential xAI models that could provide explanations to multiple *explainees* for the AI action of perception. This analysis shows even for the same AI action, due to the diverse needs of *explainees,* different and multiple explanations are needed. These will in turn, drive the need for multiple xAI solutions.

This analysis of multiple stakeholders' explanatory needs also yielded the expected results that xAI solutions need to consider contextuality. Contextuality requires providing different and multiple explanations for the same phenomenon, to different *explainees*. And even when examining the situation from the vantage point of one AI action, *explainee*-driven contextuality is paramount. As anticipated and demonstrated above, within the same AI domain application, and within that, the same AI task and AI model (perception), multiple xAI models will be required for the same phenomenon and/or task to be explained in the same situation because different *explainees* will seek different explanations. Essentially, multiple xAI models and/or solutions will be required because each individual *explainee* will seek different and in some cases, multiple explanations.

5.3.7 Analysis of FGCR Task#3 – Generation of Detailed xAI Model Requirements for AI's Action of System Disengagement for Multiple Stakeholders

To illustrate the functionality of FGCR Task#3, Generation of Detailed xAI Model Requirements, I am presenting the analysis for multiple stakeholders for the AI's action of system disengagement. The above results about the *explanans* sought by multiple stakeholders for the AI's action of system disengagement from section 5.2.6 can be summarized as a set of requirements:

(1) The xAI model(s) shall provide explanation(s) for the act of system disengagement by identifying the necessary and salient components in the causal chain.

More specifically, the xAI model shall provide information about the following:

- When (specific time before the crash) did the AV sense there was an obstacle up ahead?
- Did the AV even perceive there was an obstacle up ahead before it disengaged?
- Why did the AV simply disengage rather than engage in emergency braking or steering away?
- What were the different options that the AV considered before it made the decision to disengage?
- What is the software's decision-making logic?

- Why disengage <1 second before the crash rather than several seconds or minutes sooner?
- Was the AV stuck in a loop of considering too many factors to analyze during the moments leading up to the crash?
- If the AV sensed the obstacle ahead, then why did it disengage? vs. If the AV did not sense the obstacle ahead, then why did it disengage? Note that in either scenario the decision to disengage is unacceptable.

(2) The xAI model(s) shall provide explanation(s) for the act of system disengagement by answering the following comparative questions about the decision to disengage.

More specifically, the xAI model shall provide answers in the form of contrastive explanations for the following specific differentiating factors. Why, specifically, did the AV:

- Disengage vs. not disengage
- Disengage <1 second before the crash vs. disengage 1-2 seconds before the crash
- Sense the obstacle vs. not sense the obstacles
- Disengage vs. brake
- Disengage vs. steer away from the obstacles
- Disengage without warning/notification with enough time for human to react vs. disengage with warning/notification with ample time for human to react
- Disengage <1 second before the crash if it sensed the object ahead vs. disengage <1 second before the crash if it did not sense the obstacle ahead – note that in either scenario the reaction and/or decision to disengage is unacceptable.

Also, as stated in previous sections, underlying all of the above requirements, is the fact that most stakeholders are seeking explanations with respect to the design and training of the AI model and explanations for the potential mismatch between training and testing environment and real-world deployment of the AI system. Therefore,

(3) xAI developers shall provide global model explanatory information as it pertains to the task of perception.

These outputs of FGCR Task#3 analysis will now serve as a set of xAI model requirements that will feed into the assessment of FGCR Task#4 in section 5.3.8 below.

5.3.8 Analysis of FGCR Task#4 – Identification of Potential xAI Models for AI's Action of System Disengagement for Multiple Stakeholders

To demonstrate the FGCR's Task#4, Identification of Potential xAI Models that Could Meet the Explanatory Needs for multiple stakeholders, I will attempt to map the requirements in section 5.3.7

above with the taxonomy of the xAI landscape. Again, for the purposes of this exercise, I will assume that the AI model is a black box neural net.

Since the overall thrust of the explanatory information sought by stakeholders is focused on the inner workings of the AV and the decision-making pathway or logic that led to the final decision to disengage the system, the explanation problem here is a model inspection or global model explanation problem. For neural nets, global model explanation solutions include explanation via single-tree (DT) and explanation via rule extraction (DR) (Guidotti et al., 2018). And model inspection solutions include the use of Feature Importance (FI), Partial Dependence Plot (PDP), or Activation Maximization (AM) xAI models (Adadi & Berrada, 2018; Guidotti et al., 2018). An FI explanator might quantify the weight and magnitude of inputs features that contributed to the decision to disengage the system within <1 second of the car crash. A PDP explanator might provide insights into the relationship between one or more input variables and the output of the AI model. And an AM explanator could provide information about the input patterns that maximize the activation of a certain neuron in a certain layer with respect to particular inputs. In addition to these, counterfactual explanators which describe the minimum conditions that would have led to an alternative decision could also be useful (Adadi & Berrada, 2018).

Furthermore, it is essential to recognize that certain stakeholder concerns necessitate a system-wide perspective. The query "Did the AV even perceive there was an obstacle up ahead before it disengaged?" cannot be answered by examining system disengagement in isolation. While the preceding analysis examines system disengagement in isolation, a robust explanation must acknowledge that Autonomous Vehicles (AVs) are goal-based systems comprised of highly interconnected subsystems. Because a decision to disengage is often the result of a causal chain—originating in perception and concluding in motion control—a holistic xAI strategy is required. Consequently, developers must account for multi-model interactions between perception, motion planning, and control functions. Addressing these dependencies, such as timing and data handovers, requires an integrated

xAI approach that synthesizes responses from multiple underlying models to generate a unified, coherent explanation for stakeholders.

Taken together, the analysis above reveals that the explanatory information sought by stakeholders for the AI's action of system disengagement will require a dual approach: xAI developers will need to provide insights into the inner workings of not only individual AI black box models that perform discrete functions, but also the overall black box nature of AI systems characterized by complex, inter-related multi-model interactions. Consequently, it is very likely that AI experts will have to leverage model inspection and global model explanators that provide technical outputs about individual AI models and then integrate them into cohesive, actionable outputs about the intricate system interactions. And these technical outputs will very likely require translation into human-centric layperson terms for non-AI expert stakeholders. The FGCR process takes this limitation of technical xAI models into account by including Task#5, Usage of Existing xAI Models to Answer Stakeholders/*Explainees'* Questions via Consultation. Therefore, xAI developers will have to perform the iterative tasks of consulting with the stakeholder(s) by providing them with explanations and interactively getting to agreement that they meet the stakeholder's explanatory needs. And regardless of the explanation produced by the FGCR process, ultimately, it is up to *explainees* to determine whether they have received an adequate explanation from both the xAI model and their developers.

The outputs of FGCR Task#4 analysis include a list of potential xAI models that could provide explanations to multiple *explainees* for the AI action of system disengagement. This analysis shows that even for the same AI action, due to the diverse needs of *explainees,* different and multiple explanations are needed. These will in turn, drive the need for multiple xAI solutions.

This analysis of multiple stakeholders' explanatory needs also yielded the expected results that xAI solutions need to consider contextuality. Contextuality requires providing different and multiple explanations for the same phenomenon, to different *explainees*. And even when examining the situation from the vantage point of one AI action, *explainee*-driven contextuality is paramount. As anticipated

and demonstrated above, within the same AI domain application, and within that, the same AI action and AI model (system disengagement), with the added complexity of comparing to another AI action and AI model (perception), multiple xAI models will be required for the same phenomenon and/or task to be explained in the same situation because different *explainees* will seek different explanations. This analysis of a second AI action evaluated the ability of the FGCR to tackle the objective of identifying a suitable xAI solution regardless of the contextual combination of factors (different AI models performing different AI actions, different and multiple *explainees*, etc.) Essentially, multiple xAI models and/or solutions will be required because each individual *explainee* will seek different and in some cases, multiple explanations.

5.4 Discussion

The findings from the proof-of-concept testing of the FGCR and the analysis presented above have revealed achievements as well as challenges and opportunities for future research. The primary accomplishment of this project was the successful demonstration of the feasibility and functionality of the FGCR's core components, Tasks#1 through #4, using a proof-of-concept assessment. The challenges included lack of access to real-world proprietary AI models involved in the driving tasks performed by AVs which necessitated a narrowed scope and prevented a full assessment of the FGCR. To complete the development and rollout of the FGCR, full acceptance testing remains for future research efforts.

The main objective of the FGCR's proof-of-concept testing was to assess whether its process and components can function as intended. The testing enabled me to assess the process workflow, evaluate the functionality of components and the utility of its tools and aids, especially the taxonomies. Essentially, the testing assessed whether the FGCR can assist xAI developers with generating detailed xAI model requirements for meeting the needs of various and diverse sets of stakeholders thereby facilitating the determination of the best suited xAI solution. In addition to assessing the functionality of

the FGCR's components, the data from the testing enabled me to finetune and optimize the process as well as the taxonomies.

During proof-of-concept testing, I assessed and evaluated the feasibility and versatility of the FGCR in two sequential stages. In stage 1, I employed case studies research and used interviews to gather real-world data about the explanatory needs of stakeholders pertaining to the outputs of an AI system. Stage 1 enabled me to execute and assess the feasibility of FGCR Task#1, Stakeholder Identification, and Task#2, Stakeholder Consultation. The discussion of the results of stage 1 testing is presented in section 5.4.1. And in stage 2, to assess the feasibility of the FGCR process and components, I used the analyzed data from stage 1 to evaluate the feasibility of FGCR Task#3, Generation of xAI Requirements, and Task#4, Identification of Potentially suitable xAI Solutions. The discussion of the results of stage 2 testing is presented in section 5.4.2. Furthermore, a discussion of the findings from Phase 2 of this research project and their implications for the field of xAI, particularly, requirements engineering for xAI, is presented in section 5.4.4 and 5.4.5.

5.4.1 Discussion of Results of Stage 1 Testing

5.4.1.1 FGCR Process

In stage 1, I employed case studies research and used interviews to gather real-world data about the explanatory needs of stakeholders pertaining to the outputs of an AI system. Specifically, I focused on the AI domain of AVs and chose a case study that explored a real-world scenario of a series of car crashes involving AVs. Stage 1's data gathering efforts inherent to case study research were functionally equivalent to the first two tasks of the FGCR – Task#1, Identification of Stakeholders/*Explainees* and Task#2, Stakeholder/*Explainee* Consultation for Identification of *Explanans*. Executing these two tasks enabled me to assess their functionality.

While the original framework (Figure 12 in Appendix B) design suggested a linear workflow, the testing revealed a more fluid reality. During execution, I found that stakeholder identification, grouping and consultation tasks often occurred simultaneously rather than sequentially. Consequently, the FGCR

process has been optimized and streamlined to reflect this operational concurrency, ensuring it remains practical for real-world applications. A comprehensive account of the changes made to the original FGCR is provided in Appendix B below.

In addition to evaluating the first two tasks of the FGCR's process I was also able to assess two of the tools outlined in Chapter 4's FGCR framework—the Stakeholder Requirements Definition and Analysis Tools as well as the SSH-based Taxonomies of Explanation.

5.4.1.2 Stakeholder Requirements Management Tools

As outlined in Chapter 3, diverse stakeholder consultations are a necessary part of robust product development activities especially the validation tasks of identifying whether the right thing is being built for the relevant stakeholder. A cornerstone of stakeholder consultation is the execution of the requirements definition and analysis process and usage of the tools as outlined in the requirements management knowledge areas of INCOSE (2023), PMBOK (2021) and SWEBOK (2014). The proof-of-concept testing conducted in this project corroborated this principle directly. By executing FGCR Tasks #1 and #2, interview techniques were used to elicit, identify, and prioritize stakeholder needs, ultimately translating them into meaningful requirements for inclusion in a formal xAI model specification. The testing showed the value of requirements management processes and tools and xAI developers can use not only interview techniques, but also some of the other tools such as questionnaires, surveys, focus groups, etc., as appropriate in the future because they will undoubtedly prove useful for identifying the needs of stakeholders.

The testing supports the claim that the requirements definition and analysis tools are indispensable to the product validation process. Critically, due to the abundance of technical xAI solutions, the results showed that the current tendency to prioritize computational functionality over human-centric considerations has produced a gap: the needs of non-expert stakeholders are not being met. This was illustrated by the results of FGCR Task#4 which yielded a great many available xAI solutions that are highly technical and require translation in human-centric terms for non-AI experts.

Recent work such as Nazat et al. (2024) and Bojarski et al. (2017) on xAI applied specifically to the autonomous vehicle domain further illustrates the computational-centric approach of current xAI initiatives. As outlined in section 3.3, Bojarski et al.'s (2017) VisualBackProp explanator was developed as a debugging tool for CNN-based self-driving car systems and therefore serves as a technical tool to be used by AI experts for addressing the model inspection problem. Similarly, Nazat et al.'s (2024) evaluation framework for anomaly detection in autonomous driving networks focuses on computational properties of xAI techniques such as SHAP and LIME without consulting any end-user stakeholders about whether the resulting explanations are adequate or comprehensible. Without deliberate human-centric requirements elicitation techniques — such as interviews, questionnaires, surveys, and focus groups — the nuanced and varied explanatory needs of non-expert stakeholders will continue to go unmet, leaving them excluded from xAI development activities entirely.

The results showed not only that stakeholder consultation is the cornerstone of effective xAI development but also underscored the importance of expanding the product validation process to encompass a holistic stakeholder-centered approach. A robust validation process must engage diverse populations, which include end-users and the specific *explainees* who are the intended audiences for explanation, throughout the development life cycle. Because what constitutes an adequate explanation Is ultimately determined by stakeholders/*explainees,* sustained and structured stakeholder consultation is necessary for meeting their needs by producing genuinely human-centric xAI solutions.

5.4.1.3 SSH-Based Taxonomies

As detailed in Chapter 3, there are many different ways to classify and categorize the SSH-based concepts of explanation. The taxonomies (the "What" and the "How" of Explanation) I have produced represent a snapshot of my interpretation of the current offerings in the SSH fields. They are not exhaustive in either breadth or depth and are intended as a starting point. Therefore, they can be modified and added to in the future by xAI developers as required. During the course of conducting

proof-of-concept testing of the FGCR, I iteratively refined and expanded upon these taxonomies, thereby exemplifying the ongoing and evolving nature of this research.

In particular, during the analysis of interview data, I encountered instances that did not align with and challenged the existing categorization in the original taxonomies, necessitating the expansion of the taxonomies to include additional categories. A comprehensive account of the changes made to the original FGCR process and taxonomies is provided in Appendix B below.

A few points with respect to the changes to both SSH-based Taxonomies of explanation that were driven by the testing merit mention here:

Under the "What" category, because a lot of the participants' queries were focused on specific device failures, I expanded the "M1-Device Failures" category by adding sub-categories (M1.1-M1.8) as well as adding a "M2=Failure due to Design Flaw" sub-category.

Under the "How" category, based on the participants' queries, I made modifications to the Causal Reasoning List. First, an overarching observation was that a majority of queries made by *explainees* during the interviews were contrastive in nature. People typically asked causal questions in contrastive or differentiating terms where they were comparing facts against foils especially when they were hypothesizing about causes.

And as a result, I amended the causal reasoning list of the original SSH-based faceted taxonomy of explanation by providing additional contrastive sub-types (CR2.1-CR2.5). Furthermore, when it comes to the phrasing of contrastive questions, contrary to the 'rather than' structure emphasized in SSH literature, people favored colloquial connectors like 'in lieu of', 'compared to', 'on the flip side', 'versus', 'instead', 'as opposed to', etc. This observation is critical for xAI developers in that the activities involved in identifying the requirements for human-centric xAI solutions must accommodate a broad linguistic spectrum.

Notably, despite this change to the contrastive reasoning list, the execution of FGCR Tasks#3 and #4 testing yielded a key insight: during the categorization of detailed quotes, I observed that

multiple sub-categories were often assigned to the same query, revealing that queries could be interpreted in multiple ways. This ambiguity could lead to unclear and ineffective requirements. Consequently, it became evident that despite the expansion of the contrastive reasoning list, xAI developers should prioritize identifying the specific differentiating factor highlighted by the *explainee,* rather than fixating on the precise categorizations of contrast cases within the provided SSH-based taxonomy. Therefore, the articulation of the differentiating factors must involve clear definitions and descriptions of both the fact and the foil. By focusing on articulating the underlying differentiating factors in the explanatory information sought by the *explainee(s)*, xAI developers can effectively define and clearly articulate that as a concrete requirement in the specification.

These findings emphasize the need for xAI developers to be flexible when it comes to generating xAI specification requirements that accurately reflect the explanatory needs of diverse *explainees*. xAI developers need to accommodate and interpret not only varied linguistic structures but also provide a clear description of both the fact and the foil in non-academic theoretical SSH-based linguistic terms. Because the FGCR aims to bridge the gap between academic SSH-based knowledge with everyday explanatory queries, in order to accurately identify the explanatory needs of *explainees,* these taxonomies should be seen as flexible instruments rather than absolute constraints requiring xAI developers to apply contextual interpretation to ensure the resulting explanations remain human-centric.

In the future, xAI developers will have to engage in iterative practices of modifying the taxonomies because as demonstrated by results of the data analysis activities of this thesis, and the dynamic nature of research in the SSH-based knowledge area of explanation. Consequently, due to the contextual demands of the validation activities of xAI solutions, and as SSH research on explanation advances, more categories will be identified, and useful ones can be integrated into the existing taxonomies as appropriate for solving the problem of providing adequate explanations to stakeholders.

5.4.2 Discussion of Results of Stage 2 Testing

And in stage 2, to assess the feasibility of the FGCR process and components, I used the analyzed data from stage 1 to assess the functionality of Task#3, Generation of xAI Requirements, and Task#4, Identification of Potential xAI Models/solutions.

5.4.2.1 FGCR Process

The evaluation of the process in this stage centered on the framework's robustness across varied contextual configuration. As stated above, while the original framework (Figure 12 in Appendix B) design suggested a linear workflow, the testing of the tasks in stage 2 also revealed a more fluid reality. During execution, I found that the majority of tasks from the original framework often occurred simultaneously rather than sequentially. Consequently, the FGCR process has been optimized and streamlined to reflect this operational concurrency, ensuring it remains practical for real-world applications. A comprehensive account of the changes made to the original FGCR provided in Appendix B below.

To assess the FGCR's capacity to handle a variety of contextual configurations, I chose to interrogate and present the findings of the testing through multiple contextual perspectives. To assess the feasibility and versatility of the FGCR while illustrating the pivotal role that contextuality plays in explanations, I subjected it to contextual conditions that are composed of multiple combinations of factors including multiple AI models performing multiple AI functions and multiple *explainees* seeking explanations for the same phenomenon (a series of car crashes) in the case study presented to interviewees.

Four different contextual configurations assessed the feasibility of the FGCR, namely, whether it can serve as a guide for identifying the best suited xAI solution regardless of AI domain, AI application, AI model, action performed by the AI, or *explainee(s)* seeking an explanation. They illustrated the pivotal role that contextuality not only plays in explanations but also drives the need for a multiplicity of explanations even for the same phenomenon which in turn drives the need for tailoring xAI solutions.

The contextual configurations that evaluated the functionality and versatility of the FGCR included two analyses that approached the data from the vantage point of *explainee(s)* and probed the need for tailoring xAI solutions to the stakeholders impacted by AI systems; and two analyses that approached the data from the vantage point of the AI's actions and probed whether the need for tailoring xAI solutions still centers on the stakeholders impacted by AI systems. Employing these four analyses exercised the FGCR's capacity to handle a variety of contextual configurations and demonstrated its versatility. They also illustrated that even when approaching the data from the vantage point of the AI's function, the need for tailoring xAI centers on the stakeholders and ultimately *explainee(s)* impacted by the AI system. The objective of tackling the complex and contextually variable nature of explanations was met, as I was able to use the FGCR to classify explanatory information sought into concrete and detailed xAI model requirements, which in turn supported the identification of suitable xAI model(s) capable of providing adequate explanation(s) to explainees.

In addition to demonstrating the functionality of the FGCR, the proof-of-concept testing and analysis of the results supported the claim in Chapter 1 that xAI solutions need to be tailored to the explanatory needs of diverse stakeholders. As shown above, xAI developers must navigate the tensions that arise from the combination of three issues: (i) the contextuality of explanations and the needs of diverse *explainee(s)*; (ii) the heterogeneity of AI applications where even within the same AI application there could be multiple AI models performing multiple tasks; and (iii) the articulation of explanations in human-centric layperson terms for most stakeholders.

The element of the pivotal role that contextuality plays in explanations and how these drive the need for tailored xAI solutions was supported by evidence during testing. By involving multiple and diverse *explainees* in the testing process, the FGCR was evaluated for its capacity to generate requirements tailored to varied stakeholder needs. The proof-of-concept testing showed that using the FGCR yielded different requirements for xAI solutions, and this was driven by the needs of diverse *explainees*.

The element of the diverse nature of AI applications and how that drives the need for tailored xAI solutions was manifested during testing. The AI application of AVs is composed of multiple AI models operating simultaneously to perform multiple tasks involved in the motion control operations of a vehicle. This enabled the testing of the FGCR to see if it can produce requirements for more than one type of AI model. The proof-of-concept testing showed that depending upon the actions taken and/or decisions made by the AI system, *explainees* sought explanatory information for, the FGCR yielded different requirements for xAI solutions.

With respect to the element of human-centric layperson termed explanations, the need to articulate explanations in human-centric layperson terms for most stakeholders was underscored by the cross-section of society that the participants represented. And the analysis of the results showed that the currently available suite of highly technical and computational xAI solutions may not provide adequate explanations for most stakeholders, and they may require xAI developers to translate their outputs into layperson terms.

Even though the results of proof-of-concept testing of the FGCR made evident that tailoring xAI solutions requires balancing the tensions between all three elements above, they also showed that it is the needs of *explainees* that take primacy in determining a contextually appropriate xAI solution. Therefore, it is crucial and necessary for xAI developers place diverse stakeholder consultation at the center of xAI development activities.

5.4.2.2 Faceted Taxonomy of the xAI Landscape

As detailed in Chapter 3, there are many different ways to classify and categorize the landscape of AI and xAI models. The Taxonomy of the xAI Landscape that I have produced represents a snapshot of my interpretation of the current offerings in both fields with a focus on the ML technique of neural nets (NN). This focus on NNs was intended to serve as a starting point and an example for proof-of-concept testing of the FGCR. Therefore, it is not exhaustive in either breadth or depth and it is anticipated that it will be modified and added to in the future by xAI developers as required. During the

course of conducting proof-of-concept testing of the FGCR, much like the SSH-based taxonomies, I iteratively refined and expanded upon this taxonomy, thereby exemplifying the ongoing and evolving nature of this research.

In particular, during the analysis of interview data, I encountered instances that did not align with and challenged the existing categorization in the original taxonomy, necessitating the expansion of the taxonomies to include additional categories. A comprehensive account of the changes made to the original taxonomy is provided in Appendix B below.

Furthermore, because there is a significant quantity and extensive range of current AI models, xAI developers will have to add to the proposed taxonomy of the FGCR depending upon the AI model that is the target of xAI development. Also, because newer models of both AI and xAI continue to be developed, xAI developers will have to continue with the activities of searching for the most current and latest available xAI models and adding to the proposed taxonomy during execution of requirements management activities as part of the validation process of xAI model development.

5.4.3 Discussion of Future Research Opportunities

5.4.3.1 Pertaining to FGCR Development and Rollout

For this thesis, the scope of testing the FGCR was limited to proof-of-concept testing and did not encompass comprehensive acceptance testing involving xAI developers. The objective was to assess its soundness, feasibility and functionality. Chiefly, whether the FGCR and its components enable xAI developers to generate detailed xAI model requirements that attempt to meet the needs of various and diverse sets of stakeholders thereby facilitating the determination of a contextually suitable xAI solution.

Due to the lack of access to proprietary AI models involved in the driving tasks performed by AVs, the iterative tasks of consulting with stakeholders with actual explanations from xAI solutions to determine if the explanations are adequate for the *explainee(s)* could not be conducted. Therefore, FGCR Task#5, Usage of Existing xAI Model(s) to Answer Stakeholder Questions, could not be evaluated. In addition, due to the limited scope of this research project, namely, proof-of-concept

testing, the outputs of Task#3 were deemed adequate for this exercise and FGCR Task#6, Generation of Detailed xAI Model Requirements for Inclusion in a Specification to Modify Existing or Create New xAI Model(s), was also not evaluated. Therefore, the testing of these latter tasks of the FGCR remains the purview of future research efforts within the scope of real-world acceptance testing.

Moreover, because the scope of this research project's testing was limited to proof-of-concept, it did not encompass acceptance testing involving xAI developers to assess whether the framework and all its components is truly useful and practical in real-world development contexts. The next step of comprehensive acceptance testing in a full field trial within industrial environment of AV manufacturers and with the participation of xAI developers to assess the FGCR's practical applicability for developing usable human-centric xAI models remains the purview of future research efforts.

5.4.3.2 Pertaining to Taxonomies (one of the FGCR's Tools)

As discussed in sections 5.4.1.3 and 5.4.2.2, the taxonomies I have produced for this thesis represent a snapshot of my interpretation of the current offerings in both the SSH and AI/xAI fields. They are not exhaustive in either breadth or depth and are intended as a starting point designed for modification and expansion by xAI developers as required. The iterative refinements made during proof-of-concept testing exemplify this ongoing and evolving nature of the research.

This flexibility is both warranted and supported by the contemporary literature in this field. Recent studies such as Mangold et al. (2025), Kim et al. (2024), and Omeiza et al. (2022), among others, have introduced taxonomies that categorize xAI methods, user groups, and stakeholder types. These categories overlap in part, with the taxonomies addressed in this thesis and rather than rendering this work redundant, these studies reinforce the case for treating the taxonomies as living instruments: their categorical structures offer xAI developers a ready source of reference material for future refinements and expansion. Omeiza et al. (2022) in particular, provides a classification of stakeholder types that could prove particularly valuable when carrying out FGCR Task#1, Identification and Recruitment of Stakeholders/*Explainees*. Furthermore, note that these works remain functionally distinct from the

present research as they focus primarily on post-hoc evaluations of xAI approaches. In contrast, the FGCR proposed in this research is uniquely designed as a forward-looking roadmap that guides xAI developers during architectural identification and implementation phases. The FGCR addresses the "what-to-build" challenge rather than "how-to-measure what is being/has been built" challenge.

Following the adaptive approach undertaken in this thesis, in the future, xAI developers will undoubtably have to engage in similar iterative practices of modifying the taxonomies. This is necessitated by the dynamic nature of the research in fields of SSH as well as AI and xAI—they are in a state of constant flux. Consequently, as research advances, and new SSH-based categories of explanation are identified, and AI and xAI models are developed, the taxonomies will require ongoing updates to remain relevant and effective. Crucially, xAI developers must treat these taxonomies as flexible instruments rather than rigid constraints in order to solve the problem of providing adequate explanations to stakeholders. Because the FGCR aims to bridge the gap between academic SSH-based knowledge with everyday explanatory queries, in order to accurately identify the explanatory needs of *explainees,* xAI developers must apply contextual interpretation to these taxonomies to ensure the resulting xAI solutions deliver human-centric explanations.

A comprehensive account of the changes made to the original FGCR process and taxonomies is provided in Appendix B below.

5.4.4 Shortcomings in Current Requirements Engineering for xAI: Findings and Implications for the Field

While emerging research in current requirements engineering practices for xAI treats explainability as an explicit software engineering requirement, shortcomings persist within current xAI literature. A rigorous analysis of the empirical data generated across Phase 3 directly illuminates why current xAI research remains misaligned with human-centric expectations. While Chapter 1 identified these theoretical imbalances as problems, the proof-of-concept testing of the FGCR framework provides concrete, empirical evidence of these operational shortcomings in action. Four core areas where current

xAI research remains incomplete are synthesized below. By demonstrating how the FGCR operationalizes these challenges, this section outlines where current approaches fall short and highlights the specific contributions of this empirical work.

5.4.4.1. The Absence of a Stakeholder-Centered, Context-Driven Requirements Process

The Shortcoming: Current literature lacks a structured, repeatable process for generating xAI model requirements that is both stakeholder-centered and responsive to the contextual nature of human explanatory needs. Köhl et al. (2019) recognized that ". . . a systematic and overarching approach to ensure explainability by design is still missing" and proposed treating explainability as a non-functional requirement (NFR) to be elicited, specified, and verified. Habiba et al. (2022) similarly identified the need for a mediating role between ML engineers and stakeholders, proposing a five-step requirements engineering framework that explicitly asks, "explainable for whom?" and "how much explainability is enough for which stakeholder?"

Both works represent meaningful contributions: they shift explainability from an isolated algorithmic concern into the domain of engineering process, and both gesture toward stakeholder differentiation as a design consideration. However, neither provides a fully *operationalized* process — one equipped with practical tools, taxonomies, and structured tasks — for translating the contextual needs of real-world stakeholders into concrete model requirements. Köhl et al.'s framework, while theoretically rich, remains a conceptual scaffold and its verification mechanisms rely on hypothetical stakeholder representatives rather than empirically grounded consultation. Habiba et al.'s framework, while explicitly user-oriented, stops at classification of requirements without providing the mechanisms by which those requirements feed dynamically into a development cycle that accounts for changing user contexts. As hypothesized in Chapter 1's treatment of Contextuality in Real-World Explanations (Problem 1), the adequacy of an explanation cannot be judged statically — it is entirely dependent on the situational realities of the *explainees*. Most critically, xAI demands a shift from static requirements specification to an adaptive requirements management process that tracks and responds to volatile,

situation-dependent human cognitive needs. Neither framework provides a structured method to elicit, trace, and negotiate evolving requirements across diverse and shifting stakeholder contexts. Without this, even technically sophisticated xAI systems remain misaligned with the non-technical stakeholders they serve.

The FGCR Resolution: The FGCR addresses this operational missing piece directly. It goes beyond identifying explainability as an NFR or proposing an abstract conceptual framework by providing a tested, multi-step process that xAI developers can follow from initial stakeholder identification through to requirement specification. The FGCR provides a proof-of-concept tested, multi-step process with aids such as taxonomies that guide developers from initial stakeholder identification through to requirement specification. The Phase 3 data demonstrated the feasibility of this capability. When confronted with the exact same phenomenon—a series of AV disengagement and subsequent crashes—individual stakeholders generated radically distinct queries depending on their context. By providing structured tasks (Tasks #1 through #4), the FGCR successfully captured the disaggregated nuances across different contextual configurations and stakeholders and generated precise, context-dependent requirements, showing that even for the same phenomenon, multiple explanations are needed depending on the tasks performed by the AV's multiple AI systems. Phase 3's proof-of-concept testing demonstrated that while the complexity of context plays a pivotal role, the need for tailoring xAI ultimately centers on the stakeholders (*explainee(s)*) impacted by the AI system. Because current xAI practices lack execution mechanisms to elicit these dynamic human variations, the FGCR's requirement of central stakeholder consultation offers a structural remedy.

5.4.4.2. Inadequate Integration of the Social Sciences and Humanities

The Shortcoming: Current xAI research remains heavily dominated by purely computational approaches that treat explanation as a computational problem to be computed rather than a communicative relational act. This highlights the shortcoming resulting from the Paucity of Human-Centric xAI Models (Problem 3) outlined in Chapter 1. Advanced post-hoc methods like LIME (Ribeiro

et al., 2016) and SHAP (Lundberg & Lee, 2017) rely entirely on mathematical models of feature importance rather than how humans actually seek, construct, and evaluate explanations. This results in technical tools built *by* developers *for* developers, without systematically leveraging philosophy, psychology, and cognitive science insights regarding how humans participate in giving and receiving explanations (Miller et al., 2017; Miller, 2019). While pioneering work called for the systematic integration of social science insights into xAI (Miller, 2019), and subsequent works adopted user-centered approaches involving contrastive explanations (Madumal et al., 2019), the broader field has yet to consistently translate this call into practical engineering methodologies.

The FGCR Resolution: This research directly operationalizes that interdisciplinary call. The FGCR's SSH-based taxonomies provide developers with a structured, empirically grounded resource to identify the qualitative "what" and "how" of human-centric explanations. The empirical data from Phase 3 showed that stakeholders do not think or query in terms of feature-importance matrices or mathematical weight vectors. Instead, their natural cognitive schema involves multiple dynamic and contextual approaches such as causal reasoning (e.g. contrastive, counterfactual, etc.); BDI; identification of material causes; different types of conversational models; etc. By embedding the SSH-Based Taxonomy of Explanation directly into the requirements loop, the FGCR bridges academic SSH knowledge with everyday explanatory queries, transforming abstract communication needs into functional criteria.

5.4.4.3. The Static Checkbox Vulnerability and the Verification-Validation Dichotomy

The Shortcoming: Current software engineering trends carry an inherent risk of treating explainability as a static, one-time checkbox rather than an adaptive, iterative process. In INCOSE (2023) systems engineering terms, existing approaches focus heavily on *verification* (confirming the system matches a fixed technical specification) rather than *validation* (confirming the right system was built for the people using it). Certification-focused frameworks (Köhl et al., 2019) and sequential elicitation processes (Habiba et al., 2022) provide structured starting points but lack the iterative,

stakeholder-engaged mechanisms that true validation demands. This ignores the operational realities tied to the Heterogeneous Technology Challenge (Problem 2) and the Contextuality Challenge of real-world explanations (Problem 1) introduced in Chapter 1. AI applications, the *explananda*, are not monolithic; they are highly fragmented, fluid ensembles of distinct models performing different acts (e.g., perception vs. motion control vs. system disengagement). And the complexities involved in the contextual nature of explanations also contribute to the fluid nature of explainability. Therefore, treating explainability as a uniform checkbox fails because a static framework cannot adapt to these distinct challenges, creating a deep dichotomy between algorithmic verification and human-centric validation.

The FGCR Resolution: The FGCR offers a roadmap to facilitate xAI developers with the task of navigating an adaptive, iterative validation process that keeps the human *explainee* at the center. The two user friendly resources, namely, the SSH-based taxonomy of explanation and the Taxonomy of the xAI Landscape, make the vast knowledge base of distinct disciplines accessible in order to facilitate the efficient identification of crucial xAI requirements. The architectural alignment of the FGCR explicitly operationalizes engineering development lifecycle validation model. Rather than treating validation as a late-stage backend checkpoint, the FGCR maps Tasks 1 through 5 as anchors to the front-end phase of the engineering development lifecycle. It aims to systematically answer early validation questions regarding stakeholder needs and map them into actionable xAI requirements.

5.4.4.4 The Environmental De-contextualization of Explanation Design

The Shortcoming: Current xAI research tends to isolate explanation design from the environmental and operational ecosystems where the AI is deployed. Traditional xAI tools have been developed within abstract laboratory settings utilizing simulation environments to evaluate explanations. Consequently, they often overlook the complex and messy, real-world environmental variables that human *explainees* focus on when a system fails, especially embodied systems that act as agents.

The FGCR Resolution: The proof-of-concept data demonstrated that a human-centric explanation cannot be divorced from environmental variables present when embodied AI systems are

deployed in real-world high-stakes situations. During Phase 3 consultations, stakeholders' queries were heavily dependent on the specific environmental context of the case study. The FGCR process explicitly integrates these environmental factors into Task #2 (Stakeholder Consultation for Identification of Explanans) and Task #3 (Generation of Detailed xAI Model Requirements), ensuring that the resulting requirements reflect the real-world operational context of the deployment environment.

5.4.5 Strategic Roadmap for Future xAI Research

As stated above in section 5.4.3, to build upon the contributions of the FGCR and address broader systemic challenges in human-centric AI requirements management, future research should pursue full acceptance testing of the FGCR involving xAI developers to assess whether the framework and all its components (including the untested Tasks #5 and #6), is truly useful and practical in real-world development contexts. And future research should prioritize the following strategic avenues:

Cross-Domain Generalization and Scalability Testing: Although the FGCR's proof-of-concept utilized a safety-critical autonomous vehicle use case, its underlying socio-technical framework is AI domain-agnostic and model-agnostic. A critical next step, upon completion of full acceptance testing of the FGCR and all of its tasks within the AV domain, is testing the scalability and boundaries of the FGCR within other high-stakes socio-technical ecosystems within healthcare, legal, or financial sectors. This will validate whether the framework's process and baseline taxonomies natively map to diverse and variable AI applications and stakeholders.

Standardized Human-Centric Validation Metrics: While this thesis demonstrates the necessity of shifting focus from computational verification to human validation, the wider xAI field still lacks standardized metrics to measure human-centric explanatory needs. Future research may establish repeatable, empirical validation test suites incorporating a range of human-centric measures such as structured cognitive workload scales, situational awareness metrics, and knowledge-transfer tasks. The goal is to measure whether explanations generated via the FGCR meet the range of human explanatory

needs such as understandability, trust, fairness, transparency, ethicality, etc. Meeting these needs has the potential to drive wider acceptance and adoption of these new technologies.

Yet this recommendation carries an important caveat. Given the inherently fluid, dynamic, and context-dependent nature of explanation established throughout this thesis, any such standardized suite must be treated with caution: the contextuality of human explanatory needs may ultimately resist reduction to a fixed battery of metrics. This is precisely why, for the purposes of this research, the FGCR narrows the task of validating an explanation during stakeholder consultation to the simple question embedded within its feedback loops: *"Is this explanation adequate?"* Indeed, the empirical work of this research suggests that the most defensible and epistemically honest validation instrument may be this deceptively simple question — posed directly to the stakeholder, in context. More elaborate validation schemes may build upon it, but none can leave it out. Ultimately, when confronted with the realities of human-centric explainability, this question is the one thing that cannot be stripped out without losing what matters most.

6 Conclusion

6.1 Summary of Research and Contributions

This research project was undertaken to address two challenges faced by the field of xAI: the paucity of human-centric explainable AI (xAI) models (Gap 1 from Chapter 1); and the complexity involved in identifying and tailoring explanations to the contextual needs of diverse *explainee(s)* (Gap 2 from Chapter 1). To address these challenges, I have developed and conducted a proof-of-concept assessment of the **Framework for Generating Context-Dependent Requirements (FGCR).** The FGCR serves as a practical guide and toolkit to aid xAI developers with the generation of requirements that feed into the engineering development cycle in the creation of human-centric xAI models.

The **main contribution** of this work to **the field of xAI development** is a framework that is a multi-disciplinary bridge between Social Sciences and Humanities (SSH) and Engineering facilitating a shift in AI development from a purely computational focus toward genuine intelligibility for the humans it affects. The usage of the FGCR will enhance the development of human-centric xAI solutions for a broad range of stakeholders by facilitating xAI developers with the generation of context-driven requirements that feed into the software requirements specification.

The key contributions of this framework to xAI development include:

- **Enhancing Current Systems and Software Engineering Knowledge**: This framework is designed to aid xAI development efforts by enabling the application of current systems and software engineering requirements knowledge (INCOSE, 2023; Bourque & Fairley, 2014) and has the potential to augment these standards and practices by expanding their scope and applicability to the burgeoning field of AI development.
- **Multi-disciplinary Taxonomies of Explanation**: By synthesizing the vast research from the Social Sciences and Humanities (SSHs) on the topic of explanation, I developed SSH-based

taxonomies of explanation that provide xAI designers with a structured way to efficiently identify and classify the "what" and "how" of human-centric explanations.

- **A Taxonomy of the xAI Landscape:** I developed a proposed structure in the form of a taxonomy that provides an efficient way to approach the identification of potential xAI models by assisting with identification of AI models that need an explanation and potential xAI problems (model inspection, global model explanation and/or local explanations).
- **The FGCR Process:** I provided a structured, iterative process for xAI developers to move from stakeholder consultation to the generation of specific and actionable xAI model requirements. The FGCR adapts established systems engineering and requirements management practices (such as those from INCOSE (2023) and Bourque & Fairley's (2014) SWEBOK) to the specific needs of xAI development.
- **Proof-of-Concept Assessment of the FGCR:** Using case study research involving 26 stakeholders in the AI domain of automated vehicles (AVs), I assessed whether the FGCR could serve as an effective mechanism to capture the explanatory needs of a range of stakeholders, from AI experts to layperson pedestrians, drivers, lawyers, and first responders. This provided empirical support for central claim of this thesis that "context matters," by showing how explanatory needs shift depending on the specific stakeholder's/*explainee's* explanatory needs and the AI action being queried. Furthermore, the data from the case study enabled me to test the functionality of the FGCR's process and components. In particular, using the FGCR to produce a set of requirements from the gathered real-world data and identify potential xAI solutions provided empirical evidence that the framework could function as intended. Essentially, the proof-of-concept testing provided empirical evidence for the feasibility of the FGCR to not only capture the explanatory needs of diverse stakeholders but also to effectively translate them into a set of requirements for the identification of a contextually suitable human-centric xAI solution.

Secondary contributions to society of this research project:

- **A Roadmap for Human-Centric AI Development Across Industries—Empowering the *Explainee***: By placing the ***explainee*** at the center of the development cycle, the FGCR offers a design for human-centric AI development – a roadmap for creating AI that is not just technically "transparent" but genuinely "intelligible". This research empirically demonstrates the value of placing the needs of the diverse human stakeholders impacted by transformative technologies such as AI at the center of the development cycle. The FGCR offers a scalable methodological blueprint that could achieve wider societal acceptance and adoption of AI, particularly in high-stakes environments such as autonomous vehicles, healthcare, food production, and law, amongst others. The FGCR offers a robust methodology and toolkit that enables AI researchers and developers to transcend the traditional "developer-centric" model. While this study utilized AVs as a primary case study, the framework is designed to be domain-agnostic. For example, in the healthcare sector, the implementation of FGCR requirements management would mandate the inclusion of not only medical practitioners but also patients, who are both the source of data and the ultimate recipients of AI-driven diagnostic outputs. By facilitating the consultation of a diverse spectrum of stakeholders, the FGCR ensures that building trust is no longer an afterthought, but a standard engineering practice. Ultimately, this work provides the necessary tools to move AI development beyond purely computational goals, ensuring that transformative technologies are built for, and understood by, the society they serve.
- **Automated Vehicle (AV) Industry:** The findings from the interviews will be of particular interest for AV manufacturers – not only Tesla but all AV manufacturers. They provide insights into the particular aspects of the AI systems that operate these vehicles and in order to increase trust and adoption of AV technology, it behooves AV manufacturers to take into account the concerns of consumers. The diverse set of stakeholders interviewed wanted explanatory information about the decisions made and actions taken by AVs especially those

pertaining of their safe operation. The data gathered during proof-of-concept testing has the potential to inform not only xAI development but also the transparency of the underlying AI systems being explained.

- **Contribution to SSH Disciplines**: While the bounded focus of this research project has been to use insights from the social sciences and humanities and apply them to xAI initiatives, it is worth noting that this research can offer a reciprocal contribution back to the very fields it drew from. Chapter 1's section 1.1.1.1 offers a unique explication of the multiple layered factors at play involving the complex nature of contextuality within explanations. And Chapter 3.1 offers something new to SSH scholars: a reorganisation of the existing body of knowledge on explanation the two top-level categories — the *what* and the *how* — along with a set of novel sub-classifications, particularly for the *how* of explanation. These taxonomies were built for xAI engineers, but they also represent a structured, operationalised treatment of explanation that may be of independent value to SSH researchers. The act of making SSH concepts legible to engineers required precision that can, in turn, clarify concepts within SSH itself.

6.2 Answering Research Questions

6.2.1 RQ1: Applying Multidisciplinary Insights to xAI

In order to apply multidisciplinary insights to the field of xAI, this research project addressed **RQ1) "What multi-disciplinary insights and knowledge from the fields of social sciences, humanities, engineering, project management, and xAI can improve xAI development efforts by assisting with the identification of suitable context-dependent, human-centric xAI solutions?"**

To comprehensively answer this main question, the research was driven by the following three sub questions:

RQ1a: "What constitutes an adequate explanation for humans?" This sub question focused on how humans explain thoughts and actions to each other, drawing from social sciences and humanities (SSH) to ensure xAI models provide explanations in layperson terms.

RQ1b: "What are the current state-of-the-art xAI approaches?" This investigation required a deep dive into the current AI and xAI landscape to identify where existing initiatives fail to meet contextual human-centric needs.

RQ1c: "What insights and practices from the field of engineering and project requirements management can inform future xAI development?" This sub question explored how established requirements management practices can be used to capture and validate context-dependent needs for xAI models.

This research identified that insights from the SSHs (regarding the concept of everyday explanations and the role that contextuality plays in explanations) and systems engineering (regarding stakeholder requirements definition) are essential for building xAI that laypersons can understand. These insights will assist with and enhance xAI development by moving beyond purely computational solutions for a narrow band of stakeholders, namely, AI experts, toward human-centric layperson termed explanations for a broad range of stakeholders impacted by AI systems.

6.2.2 RQ2: The Framework for Generating Context-Driven Requirements

In order to improve human-centric xAI development efforts, this research project addressed **RQ2) "What is a framework that can be employed for identifying and classifying explanatory information sought by diverse stakeholders and generating context-dependent requirements for AI systems' explainability?"**

In response to RQ2, the FGCR was developed as a viable framework for identifying and classifying the varied explanatory needs of stakeholders. This project developed a framework composed of a process and a suite of tools and aids such as taxonomies that effectively translates complex human explanatory needs into technical xAI specifications. The step-by-step process consists of stakeholder

identification, consultation, requirements generation, and xAI model selection. The proof-of-concept testing, utilizing a case study of Autonomous Vehicle crashes, demonstrated the feasibility of the FGCR as a tool for identifying and classifying explanatory information enabling the translation of complex human explanatory needs into technical requirements.

6.3 Challenges and Limitations

A primary challenge encountered during this research was the lack of access to proprietary real-world AI models used in current AV systems, such as Tesla's Autopilot. Without such access, it was not possible to generate authentic xAI explanations of AV decisions, which constrained the research to a proof-of-concept approach built on assumptions. This included treating varied AI driving actions as performed by a single machine learning technique (neural networks). This limitation of lack of access also had a direct consequence for the iterative stages of the FGCR process where Task#5, Usage of Existing xAI Models to Answer Stakeholders/*Explainees'* Questions via Consultation and Task#6, Generation of Detailed xAI Model Requirements for Inclusion in a Specification to Modify Existing or Create New xAI Models could not be evaluated. Testing was therefore limited to FGCR Tasks#1 through #4. As a result, the research focused on testing the generation of requirements that xAI solutions would need to meet and the identification of potential xAI that could satisfy those requirements rather than the validation of actual xAI solutions and outputs with the participation of xAI developers and iterative stakeholder consultation.

A further limitation concerns the scope of the domain in which the FGCR was tested. This research was conducted entirely within the autonomous vehicle domain. The field of AVs is characterized by considerable technological complexity which requires the sophisticated integration of multiple AI models performing multiple tasks such as motion planning, vehicle localisation, perception, and pedestrian and traffic signal detection which employ multiple ML techniques including neural networks, support vector machines, and reinforcement learning. While this complexity offered an

opportunity to examine the model-agnostic design of the FGCR, its domain-agnostic claims remain unexamined. While the FGCR is designed to be both model-agnostic and domain-agnostic, its applicability to other high-stakes AI domains — such as healthcare, law, and finance — was not examined. Establishing the framework's generalisability across such domains represents an important direction for future research.

The proof-of-concept testing also involved a number of methodological constraints that warrant acknowledgement. The xAI model identification exercise (Task#4), in addition to assuming that all AI actions were performed by a single machine learning technique (neural networks), was conducted as a researcher-led analysis of interview data. No xAI developer independently applied the framework, and no *explainee* evaluated the adequacy of an explanation produced through the FGCR process. Similarly, the classification and requirements-generation activities — including the mapping of stakeholder questions to SSH-based categories in the taxonomy and then to xAI requirements — were performed by the author of this thesis. This interpretive assessment was neither independently replicated by another analyst nor carried out by an xAI developer, which represents a limitation in terms of inter-rater reliability and practitioner validation.

These limitations collectively point to the most critical directions for future research. What remains necessary is comprehensive real-world acceptance testing involving xAI developers working with deployed AI models, alongside the iterative consultation of stakeholders with authentic explanations generated by actual xAI solutions in order to determine whether those explanations are adequate for the *explainee(s)*. Such testing would provide the validation that the proof-of-concept approach of this research could not and would assess the FGCR's practical applicability for developing usable, human-centric xAI models across varied deployment contexts.

6.4 Future Work

Building on the foundation laid by this research project, and insights discussed in section 5.4.2 above, future research should focus on:

- **Full Acceptance Testing of the FGCR:** A vital next step that should be undertaken is the full acceptance testing of the FGCR within an industrial setting involving the AI domain of AVs. Future research should involve partnerships with AV manufacturers and xAI developers to apply the FGCR to live, proprietary AI models. This would allow for validation of the adequacy of the generated explanations for stakeholders by undertaking iterative consultation with the stakeholders interviewed for this research project. Furthermore, since the FGCR is designed to be both model-agnostic and domain-agnostic, future research should undertake full acceptance testing across industries, AI applications, and AI models, especially those involving high-stakes AI domains (such as healthcare, law, finance, etc.) in order to establish its generalisability.
- **Refining FGCR Tools and Taxonomies:** There is also a significant opportunity to refine the FGCR tools, especially the taxonomies as AI technology evolves. This will involve the adaptation of the existing taxonomies to address the unique explanatory requirements from a domain-specific perspective. The domain-specific taxonomy expansion could move beyond autonomous transportation sectors to other high-stakes sectors beyond autonomous transportation such as healthcare, law, etc.
- **Societal Safety Concerns and Regulatory Integration:** Although not part of the primary scope of this research project, one observation deserves particular mention: Interview data revealed a pattern of societal safety concerns that aligns with the findings of Li et al. (2016). Specifically, the participants in this study echoed Li et al.'s (2016) findings that "Participants confronted with a self-driving car at fault allocated greater responsibility to the manufacturer and the government than participants who were confronted with a human driver at fault did".

This pattern suggests a "disconnect" between current legislative and academic focus and public expectations. Historically, global policy frameworks have prioritized data privacy, exemplified by the European Union's *Regulation (EU) 2016/679 (General Data Protection Regulation)*. And academic discourse has frequently been preoccupied with theoretical moral dilemmas like the "Trolley Problem". In contrast, the stakeholder consultations in this research project reveal that the public's primary concern lies in operational safety, institutional accountability, and robust regulatory oversight. To bridge this gap, this study advocates for global adoption and rigorous enforcement of frameworks and laws similar to the more recent European Union's *Artificial Intelligence Act (Regulation (EU) 2024/1689)*. Furthermore, the implementation of AI technologies should be guided by the consensus-based approach established by Fjeld et al.'s (2020) *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles*. Their mapping of ethical and rights-based approaches provides a vital blueprint for drafting comprehensive legislation that transcends the focus on the much discussed data privacy issues to address the multidimensional requirements of societal safety, institutional accountability and human rights, ultimately providing the empirical evidence needed to advance AI licensing and public-interest accountability. The global adoption and rigorous enforcement of legislation that prioritizes public interest, coupled with an increased transparency is essential to addressing the actual needs of citizens. The findings from the interview data gathered for this study can serve as empirical evidence for societal safety concerns and advocacy for AI licensing and accountability as part of future research efforts. Leading from this, to enhance safety of and trust in AI applications that have real-world high-stakes outcomes, components of the FGCR could be leveraged as a basis for formal transparency guidelines. Future research could focus leveraging the requirements management practices of the FGCR as a

basis for formal transparency guidelines and standardized regulatory protocols especially in high-stakes AI domains such as healthcare, law, and transportation.

6.5 Closing Remarks

This research reinforces this thesis' central proposition: In xAI development, context matters, and *explainee*-driven contextuality is paramount for human-centric xAI development. The FGCR offers a roadmap for providing human-centric xAI solutions by shifting the focus from "how the machine works" to "what the human needs". Building the trust and transparency necessary for AI's continued adoption in society depends on it. Trust in AI will not be won through computational sophistication alone, it will be earned through explanations that are intelligible, relevant, and tailored to the humans they serve. For human-centric xAI, context does not just matter, it is the starting point for every requirement, every explanation, and every solution.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352. https://doi.org/10.1016/0010-0277(94)00640-7

Alpaydin, E. (2014). *Introduction to machine learning* (Third edition.). The MIT Press.

Aristotle, & Charlton, W. (1984). *Physics* (1st edition). Clarendon Press.

Aristotle, Hardie, R. P., & Gaye, R. K. (1984). Physics. In J. Barnes (Ed.), *Aristotle: The Complete Works. Electronic Edition. : The Complete Works of Aristotle. BOLLINGEN SERIES LXXI - 2 Volume I :: The Complete Works of Aristotle (Volume 1)*. http://pm.nlx.com.proxy.bib.uottawa.ca/xtf/view?docId=aristotle/aristotle.01.xml;chunk.id=div.aristotle.v1.28;toc.depth=2;toc.id=div.aristotle.v1.28;brand=default

Bachute, M. R., & Subhedar, J. M. (2021). Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms. *Machine Learning with Applications*, *6*, 100164. https://doi.org/10.1016/j.mlwa.2021.100164

Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., & Zieba, K. (2017). *VisualBackProp: Efficient visualization of CNNs* (arXiv:1611.05418). arXiv. https://doi.org/10.48550/arXiv.1611.05418

Boston Dynamics. (n.d.). *Stretch™—Autonomous Case Handling Robot*. Boston Dynamics. Retrieved 9 April 2023, from https://www.bostondynamics.com/products/stretch

Bourque, P., & Fairley, R. E. (Eds.). (2014). *Guide to Software Engineering Body of Knowledge (SWEBOK), Version 3*. IEEE Computer Society. https://www.computer.org/education/bodies-of-knowledge/software-engineering/

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Centers for Disease Control and Prevention (US), National Center for Chronic Disease Prevention and Health Promotion (US), & Office on Smoking and Health (US). (2010). *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General.* Centers for Disease Control and Prevention (US). https://www.ncbi.nlm.nih.gov/books/NBK53010/

Ceva. (2016, April 14). AI leaps forward, mastering the ancient game of Go. *Ceva*. https://www.ceva-ip.com/blog/artificial-intelligence-leaps-forward-mastering-the-ancient-game-of-go/

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://doi.org/10.48550/arXiv.2307.09009

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dennett, D. C. (1989). *The intentional stance*. MIT press. https://books.google.ca/books?hl=en&lr=&id=J7GzEAAAQBAJ&oi=fnd&pg=PR9&ots=4A3Uv1Dq-N&sig=-vGumVqGn3tpOi3rBzC2HA_Lr_Q

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds* (First Edition.). W.W. Norton & Company.

Dongre, V., Rossi, R. A., Lai, V. D., Yoon, D. S., Hakkani-Tur, D., & Bui, T. (2025). *Drift No More? Context Equilibria in Multi-Turn LLM Interactions* (arXiv:2510.07777v1 [cs.CL]). arXiv. https://arxiv.org/html/2510.07777v1

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Escape Velocity Labs. (2023, August 7). *How AI beat the Go world champion | by Escape Velocity Labs | Medium*. https://medium.com/@evlabs/how-ai-beat-the-world-champion-at-the-game-of-go-47d92f7197b4

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

European Parliament. (2020, April 9). *What is artificial intelligence and how is it used?* Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

European Union. (2018, May 25). *Regulation (EU) 2016/679 (General Data Protection Regulation)*. GDPR.Eu. https://gdpr.eu/tag/gdpr/

European Union. (2024, June 13). *The Act Texts | EU Artificial Intelligence Act (Regulation (EU) 2024/1689)*. https://artificialintelligenceact.eu/the-act/

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (SSRN Scholarly Paper No. 3518482). https://doi.org/10.2139/ssrn.3518482

Ghavanati, S., Motger, Q., & Frattini, J. (2025). AIRE 2025: 12th International Workshop on Artificial Intelligence and Requirements Engineering. *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)*, 1–3. https://doi.org/10.1109/REW66121.2025.00091

Gitlin, J. M. (2022, June 15). *NHTSA data shows Teslas using Autopilot crashed 273 times in less than a year*. Ars Technica. https://arstechnica.com/cars/2022/06/teslas-using-autopilot-crashed-273-times-in-less-than-a-year/

Gitlin, J. M. (2023, July 18). *NHTSA investigating Tesla Autopilot after yet another fatal crash*. Ars Technica. https://arstechnica.com/cars/2023/07/feds-open-yet-another-safety-investigation-into-tesla-autopilot/

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 41–58. Scopus.

Google DeepMind, & Kohs, G. (Directors). (2020, March 13). *AlphaGo—The Movie | Full award-winning documentary*. https://www.alphagomovie.com/. https://www.youtube.com/watch?v=WXuK6gekU1Y

Google DeepMind. (2024, May 14). *Technology AlphaGo*. Google DeepMind. https://deepmind.google/technologies/alphago/

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44–58.

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), e61. https://doi.org/10.1002/ail2.61

Habiba, U., Bogner, J., & Wagner, S. (2022). Can Requirements Engineering Support Explainable Artificial Intelligence? Towards a User-Centric Approach for Explainability Requirements. *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, 162–165. https://doi.org/10.1109/REW56159.2022.00038

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, *74*(1), 88–95. Scopus.

Hesslow, G. (1988). The problem of causal selection. *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, 11–32. Scopus.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81. Scopus. https://doi.org/10.1037/0033-2909.107.1.65

Hilton, D. J., McClure, J. L., & Slugoski, B. R. (2005). The course of events: Counterfactuals, causal sequences, and explanation. In *The Psychology of Counterfactual Thinking* (pp. 44–60). Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-33747890590&doi=10.4324%2f9780203963784&partnerID=40&md5=b3260f097ad2d2991634a6e1cf7c3b1c

Hirshorn, S. R., Voss, L. D., & Bromley, L. K. (2017). *NASA Systems Engineering Handbook*. https://ntrs.nasa.gov/citations/20170001761

INCOSE. (2023). *INCOSE systems engineering handbook*. John Wiley & Sons. https://books.google.ca/books?hl=en&lr=&id=HhjBEAAAQBAJ&oi=fnd&pg=PR9&dq=INCOSE&ots=hPADS7uXqY&sig=K-K9RLY-u9ZnQIYMXn82rQc2o2I

International Organization for Standardization. (2015). *Quality management systems — Fundamentals and vocabulary* (ISO Standard No. 9000:2015). https://www.iso.org/standard/45481.html

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. *Judgment under Uncertainty: Heuristics and Biases*, 201–208. Scopus.

Karma, R. (2025, September 7). Just How Bad Would an AI Bubble Be? *The Atlantic*. https://www.theatlantic.com/economy/archive/2025/09/ai-bubble-us-economy/684128/

Kass, A., & Leake, D. (1987). Types of Explanations. *Types of Explanations*, 1–75. Scopus.

Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, *57*, 227–254.

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1456486

Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368. https://doi.org/10.1109/RE.2019.00046

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., … Chan, L. (2025). *Measuring AI Ability to Complete Long Tasks* (No. arXiv:2503.14499). arXiv. https://doi.org/10.48550/arXiv.2503.14499

Lewis, D. (1986). Causal explanation. *Philosophical Papers*, *2*, 214–240. Scopus.

Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). *From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars*. SAE International. SAE 2016 World Congress and Exhibition. https://doi.org/10.4271/2016-01-0164

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, *27*, 247–266. Scopus.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48. Scopus. https://doi.org/10.1207/s15327957pspr0301_2

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). *Explainable Reinforcement Learning Through a Causal Lens* (arXiv:1905.10958). arXiv. https://doi.org/10.48550/arXiv.1905.10958

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Mangold, A., Zietz, J., Weinhold, S., & Pannasch, S. (2025). *On the Design and Evaluation of Human-centered Explainable AI Systems: A Systematic Review and Taxonomy* (arXiv:2510.12201; Version 1). arXiv. https://doi.org/10.48550/arXiv.2510.12201

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040297511&partnerID=40&md5=388589d723800f6eee76df36692c0587

Mayes, G. R. (n.d.). *Theories of Explanation.* Internet Encyclopedia of Philosophy. Retrieved 22 August 2023, from https://iep.utm.edu/explanat/

Maxwell, J. (2022). Interactive approaches to qualitative research design. In U. Flick (Ed.) Interactive approaches to qualitative research design (Vol. 2, pp. 41-54). SAGE Publications Ltd, https://doi.org/10.4135/9781529770278.n4

McLeod, C. (2023). *Trust*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2023/entriesrust/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum. *IJCAI-17 Workshop on Explainable AI (XAI)*, 36–42. Scopus.

Mohammed, M., Bashier, E. B. M., & Khan, M. B. (2017). *Machine learning: Algorithms and applications*. CRC Press. https://doi.org/10.1201/9781315371658

Moreschi, A., Nejman, A., & Deal, L. (2023, July 24). *Mystery Accidents: Teslas in 'Autopilot' crashing into emergency vehicles*. CBS Austin Spotlight on Ameria. https://cbsaustin.com/news/spotlight-on-america/responders-at-risk-nhtsa-probes-driver-assistance-systems-after-a-series-of-crashes-involving-teslas-and-emergency-vehicles

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y., & Ryu, K. H. (2019). An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. *Sustainability*, *11*(3), Article 3. https://doi.org/10.3390/su11030699

Nazat, S., Arreche, O., & Abdallah, M. (2024). On Evaluating Black-Box Explainable AI Methods for Enhancing Anomaly Detection in Autonomous Driving Systems. *Sensors*, *24*(11), 3515. https://doi.org/10.3390/s24113515

Ng, A., Katanforoosh, K., & Mourri, Y. B. (n.d.). *Neural Networks and Deep Learning—Introduction to Deep Learning—Week 1*. Coursera. Retrieved 28 December 2023, from https://www.coursera.org/learn/neural-networks-deep-learning/home/welcome

NHTSA. (n.d.). *Search Safety Issues | NHTSA* [Text]. Retrieved 3 August 2023, from https://www.nhtsa.gov/search-safety-issues

NHTSA. (2022) *NHTSA Action Number EA22002 Open Investigation Autopilot & First Responder Scenes.* DOT NHTSA ODI Document. https://static.nhtsa.gov/odi/inv/2022/INOA-EA22002-3184.PDF

NOVA PBS Official (Director). (2018, May 16). *Meet the Robots at Amazon*. https://www.youtube.com/watch?v=HSA5Bq-1fU4

Papaioannou, D., Sutton, A., Carroll, C., Booth, A., & Wong, R. (2010). Literature searching for social science systematic reviews: Consideration of a range of search techniques. *Health Information & Libraries Journal*, *27*(2), 114–122. https://doi.org/10.1111/j.1471-1842.2009.00863.x

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2022). Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865

Pearl, J. (with Mackenzie, D.) (2018). *The book of why: The new science of cause and effect* (First edition.). Basic Books.

Peirce, C. S. (1903). Harvard lectures on pragmatism. *Collected Papers*, *5*, 188–189.

Project Management Institute. (2021). *The standard for project management and a guide to the project management body of knowledge (PMBOK guide).* (Seventh Edition.). Project Management Institute, Inc.

Reevely, D. (2025). *We tried the AI tool Evan Solomon used to summarize an AI bill. It didn't go well—The Logic*. Thelogic.Co. Retrieved September 7, 2025, from https://thelogic.co/news/evan-solomon-ai-podcast/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why should i trust you?' Explaining the predictions of any classifier*. *13-17-August-2016*, 1135–1144. Scopus. https://doi.org/10.1145/2939672.2939778

Schlosser, M. (2019). *Agency*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2019/entries/agency/

SEBoK Editorial Board, & Hutchison, N. (2024). *The Guide to the Systems Engineering Body of Knowledge (SEBoK), v. 2.10*. https://sebokwiki.org/wiki/Guide_to_the_Systems_Engineering_Body_of_Knowledge_(SEBoK)

Stoklosa, A. (2022, June 15). *NHTSA Finds Teslas Deactivated Autopilot Seconds Before Crashes*. MotorTrend. https://www.motortrend.com/news/nhtsa-tesla-autopilot-investigation-shutoff-crash/

Swallow, R. (2023, July 28). AlphaGo and the Future of Games. *Rupert Swallow*. https://rupertswallow.com/2023/07/28/alphago-and-the-future-of-games/

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. Scopus. https://doi.org/10.1017/S0140525X00057046

Theobald, O. (2017). *Machine learning for absolute beginners: A plain English introduction* (Vol. 157). Scatterplot press London, UK. https://ammroc-files.edgegroup.ae/s3fs-public/jobs/pdf-machine-1learning-for-absolute-beginners-a-plain-english-introduc-oliver-theobald-pdf-download-free-book-63cb5bb.pdf

Usman, M., Britto, R., Börstler, J., & Mendes, E. (2017). Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method. *Information and Software Technology*, *85*, 43–59. https://doi.org/10.1016/j.infsof.2017.01.006

Walton, D. (2011). A dialogue system specification for explanation. *Synthese*, *182*(3), 349–374. Scopus. https://doi.org/10.1007/s11229-010-9745-z

Watson, D. S., & Floridi, L. (2021). The explanation game: A formal framework for interpretable machine learning. *Synthese*, *198*(10), 9211–9242. Scopus. https://doi.org/10.1007/s11229-020-02629-9

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10.

Woodward, J., & Ross, L. (2021). *Scientific Explanation*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/

Yousafzai, M., & Christina, L. (2013). *I Am Malala: The Girl Who Stood Up for Education and Was Shot by the Taliban* (First Edition). New York, NY: Little Brown and Company. https://web-p-ebscohost-com.proxy.bib.uottawa.ca/ehost/ebookviewer/ebook?sid=3eb39158-c638-455f-b054-0661e941f354%40redis&vid=0&format=EK

Appendix A: Details of Components of Data Collection Process

Research Ethics Board Approval Letter

03/02/2025

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

CERTIFICAT D'APPROBATION ÉTHIQUE | CERTIFICATE OF ETHICS APPROVAL

Numéro du dossier / Ethics File Number	H-01-25-10973
Titre du projet / Project Title	CONTEXT MATTERS: A Framework for Generating Context-dependent Requirements for Developing Human-Centric Explainable AI (xAI) Models
Type de projet / Project Type	Thèse de doctorat / Doctoral thesis
Statut du projet / Project Status	Approuvé / Approved
Date d'approbation (jj/mm/aaaa) / Approval Date (dd/mm/yyyy)	03/02/2025
Date d'expiration (jj/mm/aaaa) / Expiry Date (dd/mm/yyyy)	02/02/2026

Équipe de recherche / Research Team

Chercheur / Researcher	Affiliation	Role
Unnati PATEL	École de conception et d'innovation pédagogique en génie / School of Engineering Design and Teaching Innovation	Chercheur Principal / Principal Investigator
André VELLINO	École des sciences de l'information / School of Information Studies	Superviseur / Supervisor

Conditions spéciales ou commentaires / Special conditions or comments

550, rue Cumberland, pièce 154 Ottawa (Ontario) K1N 6N5 Canada
550 Cumberland Street, Room 154 Ottawa, Ontario K1N 6N5 Canada

613-562-5387 • 613-562-5338 • ethique@uOttawa.ca / ethics@uOttawa.ca
www.recherche.uottawa.ca/deontologie | www.recherche.uottawa.ca/ethics

Recruitment Flyer & Recruitment Letters

Recruitment Flyer

uOttawa
University of Ottawa

Digital Transformation and Innovation Program

PARTICIPANTS NEEDED FOR RESEARCH IN THE FIELD OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (xAI)

What is this study about?

This study aims to advance explainable AI (xAI), which tackles AI's "black box problem" by enabling users to understand the outputs of opaque AI systems. By applying insights from social sciences and humanities on how humans engage in acts of explanation, this project seeks to make xAI more accessible to a broader range of stakeholders beyond xAI developers. The findings will help develop more human-centric xAI models.

Who can participate?

English speaking adults with ability to speak, understand, and write fluently in English.

We are seeking participants from various roles in society (e.g., police officers, lawyers, judges, first responders, drivers, AI experts, pedestrians, policymakers, passengers, etc.) affected by AI systems like automated vehicles (AVs). Your professional and lived experiences offer valuable insights into the kind of explanations you would seek regarding AV decisions and actions.

A screening questionnaire will be used to select participants. Participants will be selected on a first come, first serve basis for each category of stakeholder.

What's involved?

Your participation will involve an either an in-person (face-to-face) or remote (telephone, video-call, or web-based live human chat interface) interview depending upon your preference. The interview will last approximately 45 to 90 minutes.

This study will use a descriptive case study approach to ask about the specific explanations you would seek from AV AI systems regarding their decisions and actions. The focus is on gathering your views and insights based on your professional and lived experiences, and not on any of your formal professional duties or responsibilities.

For more information or to participate in this study:

Please fill out a screening questionnaire: [redacted]
Or call the dedicated phone number for this study: [redacted]
Contact Information for Principal Investigator: [redacted]

Recruitment Letter – Invitation to Participate

uOttawa
University of Ottawa

Digital Transformation and Innovation Program

Date: February 6, 2025

[Name of interviewee]
[Designation, Organization]

Dear [insert name of interviewee],

Subject: Request for Participation in PhD Dissertation Study

Please allow me to introduce myself, my name is Unnati Patel, and I am a constituent of yours in the riding of Oakville. I am a PhD candidate in the Tri-Faculty Digital Transformation and Innovation Program at the University of Ottawa. I am writing to request your participation in my research project in the field of explainable AI (xAI) which addresses AI's current "black box problem". My dissertation project, A Framework for Generating Context-Dependent Requirements for Developing Human-Centric Explainable AI (xAI) Models, is being conducted under the supervision of Dr. André Vellino.

This project is situated within the field of explainable AI (xAI), a set of processes and methods that would allow human users to understand the outputs of opaque AI systems. This project focuses on the human aspects of explainability by applying ideas from the social sciences on how humans engage in acts of explanation to the field of xAI. This project aims to aid the development of xAI that meet the human-centric layperson explanatory needs of a broad range of stakeholders beyond AI experts.

I am seeking participants who represent a diverse range of stakeholders across different sectors of society who are affected by the deployment of AI systems such as automated vehicles (AVs) in our communities. And I believe that your professional lived experiences in your role as a legislator, policy maker, lawyer, and academic offers valuable insights into the kinds of explanations you would seek regarding the decisions made and actions taken by AVs. Participants must be English-speaking adults with ability to speak, understand, and write fluently in English. A screening questionnaire will be used to identify the category of stakeholder role in society that you hold, and participants will be selected on a first come, first serve basis for each of the identified category of stakeholder.

Our conversation will last approximately 45 to 90 minutes. This project will use a descriptive case study approach to ask you about the specific explanations you would seek from AV AI systems regarding their decisions and actions. Please note that the focus will be on gathering your views, opinions, and insights based on your professional and lived experiences and not on any of your formal professional duties, responsibilities, or actions.

Participation in this study is, of course, completely voluntary. Your personal information will remain confidential, with your anonymity protected throughout all stages of the project. Only my supervisor and I will have access to the information provided by you. I would like to schedule each conversation at a date and time of your convenience during the period spanning from the present to August 31, 2025.

If you wish to participate in this study or have any additional questions, please feel free to reach out via the following contact information for this study: Dedicated phone number for this study – [redacted] or Screening questionnaire – [redacted]

Thank you for considering my request. Your assistance would be invaluable in helping me complete this research, and I truly appreciate your time and support.

Best regards,

[redacted]

Unnati Patel, PhD candidate
Tri-Faculty Digital Transformation & Innovation Program
University of Ottawa
Email: [redacted]

Recruitment Letter – Request for Recruitment Assistance

uOttawa
University of Ottawa

Digital Transformation and Innovation Program

Date: February 6, 2025

[Name of interviewee]
[Designation, Organization]

Dear [insert name of contact], ,

Subject: Request for Assistance with Participant Recruitment for PhD Dissertation Study

Please allow me to introduce myself, my name is Unnati Patel, and I am a PhD candidate in the Tri-Faculty Digital Transformation and Innovation Program at the University of Ottawa. I am reaching out to you in the hopes that you can assist me with recruiting individuals who may be interested in contributing to the important research in the field of explainable AI (xAI) which addresses AI's "black box problem". My dissertation project, A Framework for Generating Context-Dependent Requirements for Developing Human-Centric Explainable AI (xAI) Models, is being conducted under the supervision of Dr. André Vellino.

This project is situated within the field of explainable AI (xAI), a set of processes and methods that would allow human users to understand the outputs of opaque AI systems. This project focuses on the human aspects of explainability by applying ideas from the social sciences on how humans engage in acts of explanation to the field of xAI. This project aims to aid the development of xAI that meet the human-centric layperson explanatory needs of a broad range of stakeholders beyond AI experts.

I am seeking participants who represent a diverse range of stakeholders across different sectors of society who are affected by the deployment of AI systems such as automated vehicles (AVs) in our communities. It is my hope to connect with various contacts you may have within professions such as police offers, lawyers, judges, AI experts, transportation policy experts, etc., to invite them to participate in this research project. I believe that their professional lived experiences in their role as police officer, lawyer, judge, policy maker, transportation expert, etc., offers valuable insights into the kinds of explanations they would seek regarding the decisions made and actions taken by AVs. Participants must be English-speaking adults with ability to speak, understand, and write fluently in English. A screening questionnaire will be used to identify the category of stakeholder role in society that the participant holds, and participants will be selected on a first come, first serve basis for each of the identified category of stakeholder.

Each participant conversation will last approximately 45 to 90 minutes. This project will use a descriptive case study approach to ask participants about the specific explanations they would seek from AV AI systems regarding their decisions and actions. Please note that the focus will be on gathering participants' views, opinions, and insights based on their professional and lived experiences and not on any of their formal professional duties, responsibilities, or actions.

Participation in this study is, of course, completely voluntary. Each participant's personal information will remain confidential, with their anonymity protected throughout all stages of the project. Only my supervisor and I will have access to the information provided by each participant. I would like to schedule each conversation at a date and time of the participant's convenience during the period spanning from the present to August 31, 2025.

If you could assist with distributing the attached recruitment flyer and contact information for this study (Screening questionnaire: [redacted] or dedicated phone number for this study: [redacted]) with people within your organization who have relevant professional and lived experiences in your field and be interested in participating in this project, I would appreciate it.

If you have any additional questions, please feel free to contact me (Email: [redacted]) or my supervisor (Email: [redacted]) directly. Thank you for considering my request. Your assistance would be invaluable in helping me complete this research, and I truly appreciate your time and support.

Best regards,

[redacted]

Unnati Patel, PhD candidate
Tri-Faculty Digital Transformation & Innovation Program
University of Ottawa
Email: [redacted]

Consent Form

Title of the study: A Framework for Generating Context-Dependent Requirements for Developing Human-Centric Explainable AI (xAI) Models

Principal Investigator: Unnati Patel, PhD Candidate
Digital Transformation and Innovation,
Faculty of Engineering, University of Ottawa
Email address: [redacted]

Supervisor: Dr. André Vellino
Information Studies
Faculty of Arts, University of Ottawa
[redacted]

Invitation to Participate: I am invited to participate in the abovementioned doctoral thesis research study conducted by Unnati Patel under the supervision of Dr. André Vellino.

Purpose of the Study: The purpose of the study is to advance current explainable AI (xAI) initiatives which tackle AI's "black box problem" by enabling users to understand the outputs of opaque AI systems. The findings will aid the development of more human-centric xAI models.

Participation: My participation will consist of participation in one interview lasting 45-90 minutes which will be either in-person (face-to-face) or remote (telephone call, video call, or web-based live human chat interface). This study will use a descriptive case study approach to ask about the specific explanations I would seek from AV AI systems regarding their decisions and actions. The focus will be on gathering my views and insights based on my professional and lived experiences, and not on any of my formal professional duties or responsibilities. By signing this form, I agree to have my interview recorded. After the meeting, a transcript of the conversation will be provided within 48 hours. I will be asked to confirm its veracity and provide any additional comments or feedback within 7 days of receipt. If no feedback is provided within the 7-day period, it will be assumed that I have reviewed the transcript and confirmed its accuracy. Beyond this period, I forfeit the right to request changes to the transcript, and the content will be considered final and authorized for its intended use.

Risks: There is no risk that is known or associated with my participation in this study. I do not have to answer the questions that I do not want to answer. I can decide to end or withdraw from the interview at any time I wish. My participation in this study is completely voluntary.

Benefits: My participation in this study will benefit society by contributing to the advancement of knowledge about how opaque AI systems arrived at their outputs. By improving transparency of otherwise AI "black boxes", this study will increase trust in and broader acceptance of AI systems throughout society.

Confidentiality and Privacy I have received assurance from the researchers that my personal information will remain strictly confidential with my anonymity protected throughout all stages of the project. To support the findings of this study, quotations and excerpts from the interviews will be labelled with pseudonyms representing the category of stakeholder that I represent, (e.g. police, policy maker, lawmaker, first responder, vehicle drivers, pedestrian, etc.) to protect the identity of all participants. My name will not appear in the thesis or reports resulting from this study.

Conservation of Data: The data collected will include either audio and/or video recordings for in-person or remote interviews or text from the online web-based live human chat interface interview; the resulting transcripts of the interview; researcher's notes and this consent form. It will be kept in a secure manner in locked cabinets and password protected computers at the home of Unnati Patel, the principal investigator (PI), which will be accessible only to her. Electronic copies of audio/video recordings and transcripts will be accessible to her supervisor, Dr. André Vellino, as required for the purposes of this study only. All data will be destroyed 5 years after the completion of the PI's PhD thesis defense.

Compensation: There is no compensation for my participation in this study.

Voluntary Participation: I am under no obligation to participate and if I choose to participate, I can withdraw from the study at any time and/or refuse to answer any questions, without suffering any negative consequences. If I choose to withdraw, all data gathered until the time of withdrawal will be removed from the dataset and not used in the study.

If I have any questions about the study, I may contact the researcher or their supervisor. If I have any questions regarding the ethical conduct of this study, I may contact the Office of Research Ethics and Integrity via email (ethics@uottawa.ca) or telephone 613-562-5800 ext. 5387.

It is recommended that I *(keep/print/save)* a copy of this consent form for my records.

Choose one of the options below, as it applies for the interview being conducted:

For in-person interview, signed consent:

Acceptance: By signing my name below, I agree to participate in this research study.

Participant's name: _____ Date: _____

Participant's signature: _____ Date: _____

Researcher's signature: _____ Date: _____

For remote (telephone or video call) interview, verbal consent:

Acceptance: By selecting the consent statement below, I agree to participate in this research study.

- ☐ Yes, I want to participate.
- ☐ No, I do not want to participate.

For online web-based live human chat interface interview, online consent:

Acceptance: By selecting the consent statement below, I agree to participate in this research study.

- ☐ Yes, I want to participate.
  - *(Name):* _____
  - *Email/Telephone number:* _____
- ☐ No, I do not want to participate.

Interview Protocol & Questionnaire

Procedure to be Followed by Interviewer

1. Recruit and categorize interview subjects (**Note, final list of representative sample of interviewees may be a subset of this list. This list is merely a starting point for recruitment and categorization of interviewees. ***)

   Category of stakeholder:

   a) Pedestrian: _____

   b) Driver of AV vehicle: _____

   c) Driver of other non-AV vehicle on the road _____

   d) First responder (e.g. Police, fire, ambulance, search and rescue, heavy equipment operators, etc.) – specify your role _____

   e) Government Agency (e.g., Transport Canada, Government of Ontario Ministry of Transportation, etc…) – specify your role and organization _____

   f) Crash investigator – specify your role and organization _____

   g) Policy Maker – specify your role and organization _____

   h) Legal Expert (e.g. Lawyer, Judge, lawmaker etc.) – Specify your role, area of expertise, organization _____

   i) Insurance Industry – specify your role _____

   j) Automotive Industry – specify your role and organization _____

   k) Engineer -- specify your role and organization _____

   l) AI developer – specify your role and organization _____

   m) Researcher -- specify your role and organization _____

   n) Passenger in Vehicle _____

   o) Other – specify: _____

2. Review interview introduction and background information with interviewee and provide interviewee instructions.
3. Provide a description of the case study scenario to interviewee.
4. Conduct face-to-face, telephone, video call or web-based live human chat interface discussion using questions from questionnaire and nudging prompts to keep the interview on track.
5. Provide a transcript of the interview within 48 hours of interview and request that a response be provided by the interviewee within 7 days of receipt of transcript.

Interview Preparation Documents for Interviewees

Currently, AI systems are seen as suffering from the "black box problem" because of their opacity regarding the rationale behind their decisions and actions. This results in a lack of suitable explanations to human users about the outputs of AI systems. The field of explainable AI (xAI) aims to provide explanations that would allow human users to understand the outputs of AI systems.

This project is conducting research in the field of xAI by developing a framework that can guide xAI developers to produce xAI models that provide explanations to diverse sets of stakeholders in human-centric layperson terms. Part of the research is to study the domain of automated vehicles (AVs). One potential way to investigate xAI is to utilize a case study of a real-life scenario involving AVs to query different stakeholders. Stakeholders that would be impacted by AV technology will be queried about specific explanatory information they would seek from AV AI systems regarding certain algorithmic decisions and actions. The focus will be on gathering the views and insights of the interviewee (you) based on your professional and lived experiences, and not on any of your formal professional duties or responsibilities.

This interview will be conducted either either in-person (face-to-face) or remote (telephone call, video call, such as Zoom or MS Teams, or web-based live human chat interface that will serve as a dialogue-window between the human interviewer (me) and human interviewee (you). With your consent, our discussion will be recorded and after the meeting, a transcript of the conversation will be provided within 48 hours of the interviewee. You will be asked to confirm its veracity and provide any additional comments or feedback within 7 days of receipt. If no feedback is provided within the 7-day period, it will be assumed that you have reviewed the transcript and confirmed its accuracy.

The purpose of this interview is to gather in-depth information on the subject of explanation from the perspective of humans. Therefore, please remember that this isn't a test for you. This interview is only gathering information about **your views, perspectives, thoughts, questions, and opinions** about a real-life scenario. There are no right or wrong answers.

You will be provided with the description of a real-life scenario followed by interview questions to elicit your views, perspectives, thoughts, questions, and opinions.

Description of AV Case Study Scenario

This case study scenario involves a real-life case within the AI application of automated vehicles (AVs). It involves the occurrences of actual car crashes involving one particular AV brand, Tesla and its advanced driver assistance system (ADAS), "Autopilot". Tesla's Autopilot system controls the steering, braking and acceleration functions of the AV without any assistance from the human driver. Furthermore, note that Autopilot could, at any time, disengage and hand over controls to the human driver.

According to USA NHTSA's Office of Defects Investigation (ODI), EA22-002, between January 2018 and January 2022, Tesla's AVs, with Autopilot engaged, were involved in 16 crashes where they struck highly visible stationary in-road or roadside first responder vehicles (e.g., police, ambulance, fire trucks, road maintenance vehicles) that were attending to pre-existing collision scenes. Furthermore, ". . . on average, in these crashes, Autopilot aborted vehicle control less than one second prior to the first impact".

Interview Questionnaire

Remember, this isn't a test for you, it is gathering information about your views, perspectives, thoughts, questions, and opinions about the case study scenario below. There are no right or wrong answers. The purpose of this interview is to gather in-depth information on the subject of explanation from the perspective of humans. Let's proceed with the questionnaire:

In order to identify the varied explanatory information that multiple stakeholders would seek regarding these crashes, following are specific questions for you:

(1) Based on the scenario described above, you are seeking explanatory information about these car crashes from "Autopilot", the AI system that controls the steering, braking and acceleration functions of the AV (i.e., motion control functions).

What specific questions would you ask of and/or what types of information would you seek from the AI system about the decisions made and actions taken by Autopilot (an agential entity) that is performing the human driving tasks of motion control of the AVs? i.e., controlling the steering, braking and acceleration functions of the AV.

Instruction Note to Interviewer: for question 1 specifically, researcher should only use Nudging Prompts provided at the end of this questionnaire to obtain more detailed information in a manner that minimizes and limits the amount of researcher bias and increase objectivity. *

(2a) How much do you know about AI systems?

On a scale of 1 to 5 where:
1=no knowledge of AI systems technology
2=knowledge of different types of AI systems and have used some applications
3=top level abstract understanding of AI algorithms – types of applications/types of algorithms for different applications/functions (e.g., NN's, CNN's, RL's)
4=software engineering knowledge of various AI systems and applications
5=expert software engineering knowledge of different types of AI systems including being part of design team that has developed these systems.

Answer: _____

(2b) If you are 3 or above, do you know what types of AI algorithms are present in AVs and what functions they perform in driving the vehicle?

Answer:

_____

_____

_____

_____

_____

(3) When you hear "Autopilot" what do you assume as the level of driving automation?

a) Fully automated _____

b) Partially automated _____

(4) Why would you want this information?

a) Trust? _____

b) Transparency? _____

c) Other, please specify _____

(5) If you wish to clarify any of the information you have provided so far, please do so here. Do you have any additional questions or comments about the research project and/or researcher?

NUDGING PROMPTS FOR INTERVIEWER

For both the in-person and the web-based live human chat interface Q&A interviews, the following are a set of prompts that can be used during the conversations in order to ensure that the interviewee is focused on the topic. These are nudging prompts for the purpose of limiting and minimizing the amount of researcher bias and increasing objectivity.

Nudging Prompt #1: If it helps, you can pretend that you are conversing with the AI named Autopilot.

Nudging Prompt #2: In order to answer the question above, you may finish the sentences such as: "I would ask x, y, z" or "I would want to know x, y, z" or "I want information about x, y, z"

Nudging Prompt #3: Those are all excellent questions. However, for the purposes of this experiment, please assume that all other systems of the car are functioning correctly (e.g., all hardware including engine, steering wheel, pedals, cameras, LIDAR, sensors, etc.… and all software such as engine optimization, fuel efficiency, etc.…). Also assume that all environmental conditions are optimal (e.g., visibility, weather, road conditions) with the exception that there are stopped vehicles ahead of you including first responder vehicles and their personnel that are attending to a crash site ahead. What explanatory information would you seek about the decisions made and actions taken by the AI agential entity named Autopilot that is performing the human driving tasks of motion control of the AVs?

Nudging Prompt #4: How does this relate to a specific question for Autopilot that is performing the human driving tasks of steering, braking, and accelerating? Can you elaborate by providing more specifics?

Nudging Prompt #5: Those are excellent questions/suggestions/responses. Would you please elaborate on the type of information you would seek with respect to the AI algorithm that makes the decisions and takes action to perform the human driving tasks of motion control of the AVs?

Nudging Prompt #6: If your objective is to ensure crashes such as these don't occur in the future, what specific things (e.g., features, functions, methodology, etc.) would you be looking for in the algorithm?

Nudging Prompt #7: What questions do you have about the algorithm itself?

Nudging Prompt #8: How would you go about improving the algorithm? e.g., Provide information about what features, functions, methodology, etc. that you would seek information about, and you would want to improve.

Appendix B: Original FGCR Components

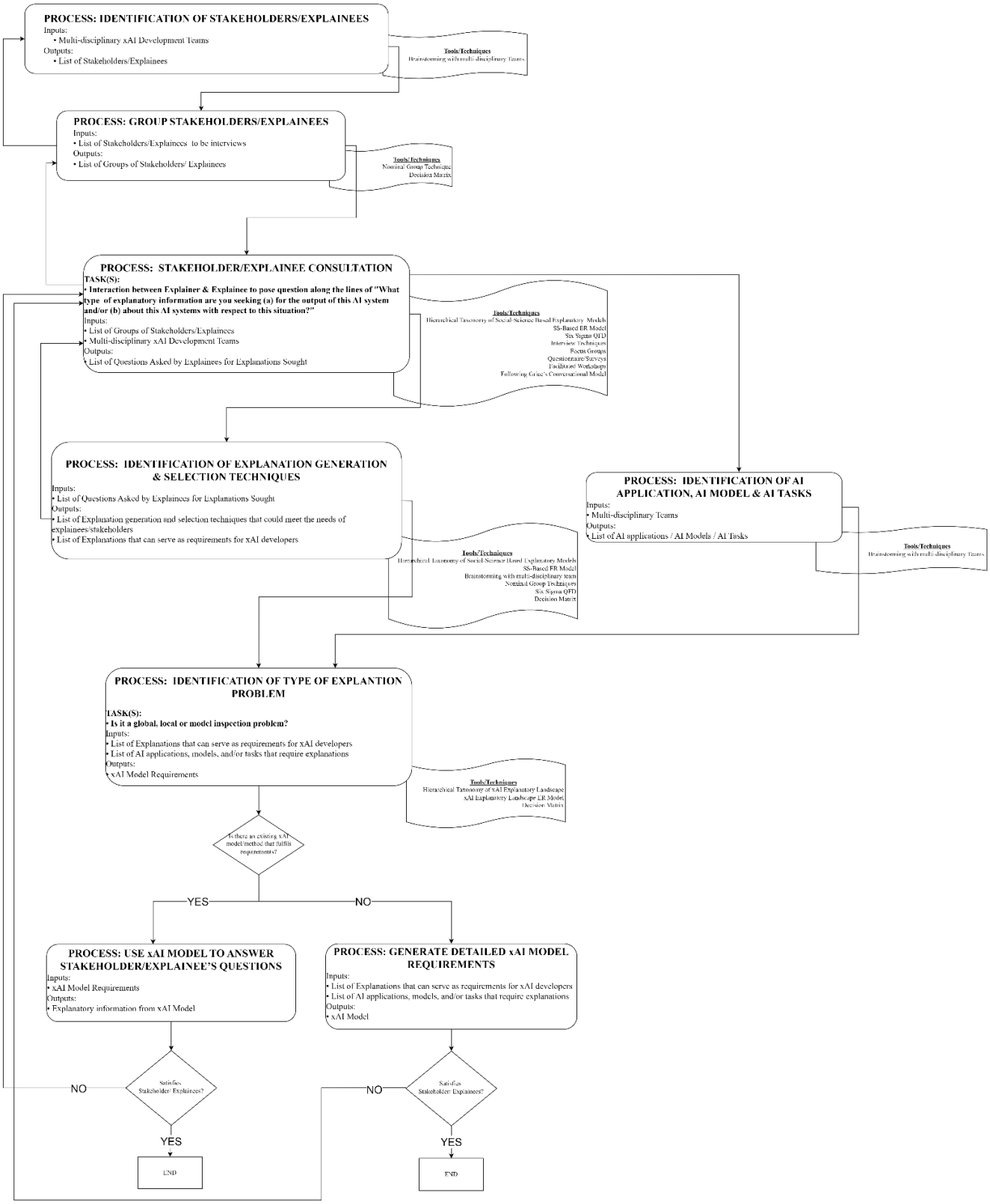
Appendix B includes the original versions of the FGCR process and the taxonomies which were amended after Phase 3's proof-of-concept testing of the FGCR process (Chapter 5). Contextual note: An extended abstract and evolutionary analysis from a research design perspective of the baseline materials contained in this Appendix is integrated into the main body of the thesis in section 4.3 (Summary of Structural Evolution of the FGCR). This material is preserved here to maintain a complete engineering development audit trail.

Original FGCR Process

The original FGCR process is illustrated below in Figure 12. Based on the results of the testing of the FGCR, I was able to finetune and optimize the process. During proof-of-concept testing of the FGCR, the original process was followed to obtain data regarding the stakeholders' needs and generating detailed xAI model requirements. However, during execution of the process, I realized that I was conducting a majority of the tasks simultaneously rather than sequentially. Therefore, I modified the original process by streamlining the presentation. The main changes involved the grouping of the original tasks of "Stakeholder/Explainee Consultation"; "Identification of Explanation Generation & Selection Techniques"; "Identification of AI Application, AI Model & AI Tasks"; and "Identification of Type of Explanation Problem" into the main task of "Task#2: Stakeholder/Explainee Consultation for Identification of *Explanans*". I also clarified and modified the inputs and outputs list of this task to include the details from the original tasks. Other modifications to the process involved minor changes of renaming the task boxes from the usage of the word "process" to explicit usage of the word "Task" and renaming of other tasks.

Figure 12:

Original Framework for Generating Context-Dependent Requirements (FGCR) Process



Original Taxonomies

Based on the results of FGCR testing including the data analysis activities, I was also able to finetune and optimize the taxonomies. The patterns in the data that emerged from the interview data about the explanatory information sought by stakeholders enabled me to modify the SSH-Based Faceted taxonomies of explanation as follows:

Under the "What" category, I modified the Material Causes List when it became evident that a lot of the participants' queries were focused on specific device failures – they viewed the AVs as engineered devices and were questioning the failures based on the functions that devices performed. Therefore, I modified the "M1=Device Failures" category by providing sub-categories (M1.1-M1.8) as well as adding a "M2=Failure due to Design Flaw" sub-category.

Under the "How" category, I added Keil's (2006) causal relations sub-types (CCT6-CCT9) to the causal chain type list (CT) which already had the types per Hilton et al.'s (2005) causal chain list that focused on the relationships between distal versus proximal events (CCT1-CCT5). Keil's (2006) causal relations list focused on the ways that causes result in effect such as converging chains of events or serial chain of events, etc.

Under the "How" category, I also modified the causal reasoning list (CR) by adding detailed sub-types of contrastive reasoning (CR2.1 – CR2.5) from Hilton (1990). In addition to this list based on Hilton (1990), I introduced a contrast case within the "design fault or bug" type of cause, "CR2.6=designed/Specified/Expected Case" with the associated type of implied question phrased as "Why X rather than *the designed/specified/expected value* for X?"

A major change under the "How" category arose from the analysis of the fourth contextual configuration discussed in Chapter 5, sections 5.2.6, 5.3.7-5.3.8 – the AI action of System Disengagement. When analysing this particular contextual configuration, it became clear that the majority of questions asked by most stakeholders were "How" questions. As discussed in Chapter 3,

section 3.1.3.2, they were asking for explanatory information about process that underlies the relationship between cause and effect and seeking salient components of the causal chain that had causal force or necessity leading to the end result. I realized that this was a major gap in my proposed SSH-based taxonomy when it came to categorizing many of the "How" questions asked by explainees about the phenomenon for which they were seeking explanations. Therefore, I added two new classes of causal reasoning based on Miller's (2019) model of classes of explanatory questions and reasoning required to answer them – "CR5=Associative" and "CR6=Interventionist". In addition, recognizing that neither of these accurately represented the "How" questions that most stakeholders were asking, I added Dennett's (2017) classifications of two sub-types of "Why" questions which I termed as "CR7=Dennett's Process Narrative" and "CR8=Dennett's Intentional Reasoning". And upon reflection, finding that Dennett's (2017) classification also did not adequately represent the type of explanatory information sought by the stakeholders interviewed, I proposed a new class of reasoning that I term "CR9=Mechanistic Causal Reasoning". As a result of these new additions to the taxonomies, I updated the categorizations in all 26 individual stakeholder analyses especially pertaining to the explanatory information sought about the actions of the AI designers. The stakeholders that asked about the failures in design/training of the AI model made queries about identifying what was overlooked in the design process and training as well as testing of the system prior to deployment in the field and posed those queries in simple "How" termed questions. I designated these as CR9=Mechanistic Causal Reasoning.

Based on the data analysis activities of FGCR Task#4 testing, I was also able to finetune and optimize the Faceted Taxonomy of the xAI Landscape. During the analysis of the explanatory information sought about the AI's act of perception, I found that LIME, SMs, and PDPs weren't the only xAI models available, therefore, I searched for more xAI models that could be useful and added CAM, AM, etc. to the list of potential models. Also note that due to the limitation of not having access to real-world AI systems operating AVs, I assumed that all AI actions performed by the AV were executed by only one type of ML technique – Neural Nets. This enabled me to manage the scope of the project while

conducting proof-of-concept testing of the FGCR. However, this limitation means that the proposed Taxonomy of the xAI Landscape (Figure 7) does not currently encompass the vast range of AI models that would typically be the target of xAI development. As a result, the taxonomy will require future expansion to accommodate emerging AI and xAI models, as well as evolving requirements management activities.

In fact, as research in SSH-based definitions of explanation, AI, and xAI continues to advance, the proposed taxonomies will necessarily undergo revisions, expansions, and updates to reflect the latest developments in these fields.

Figure 13:

Original SSH-Based Hierarchical Taxonomy

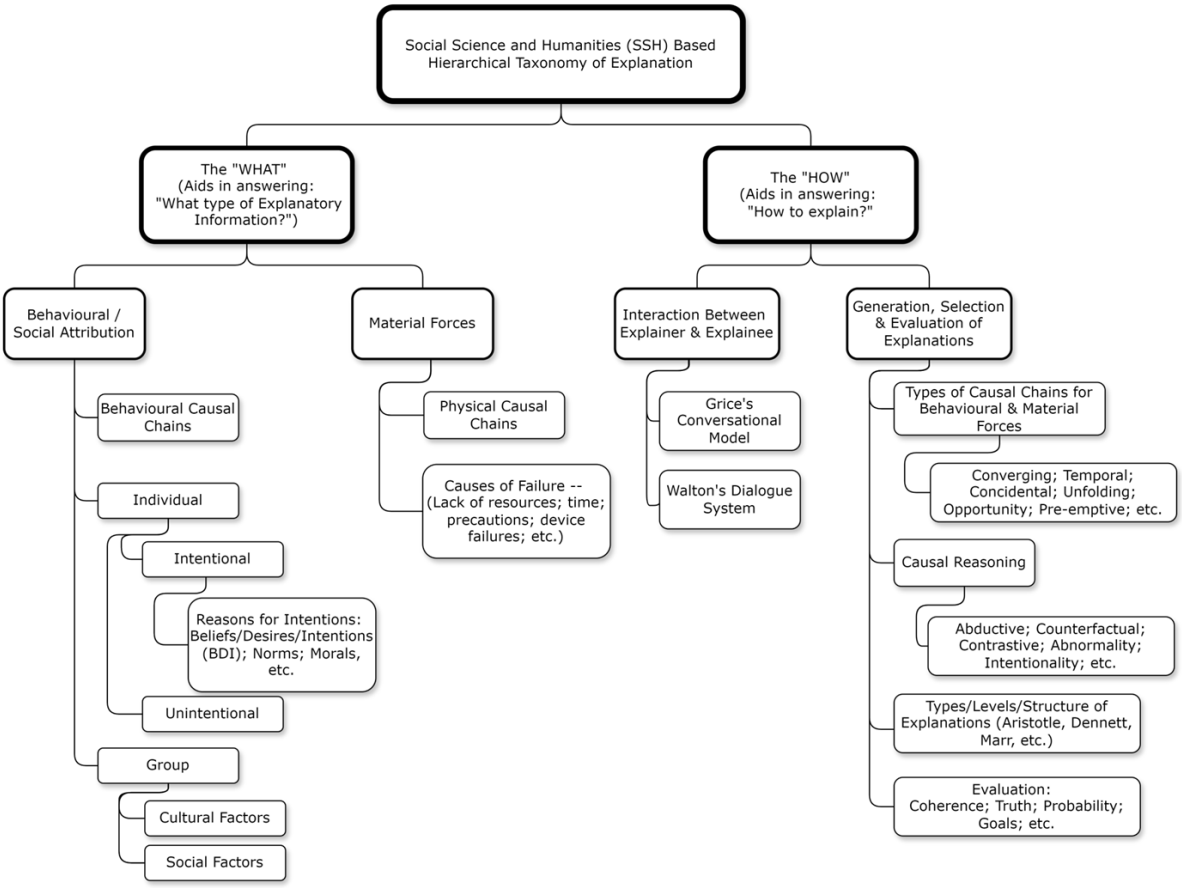


Figure 14:

Original SSH-Based Faceted Taxonomy of Explanation—The "WHAT"

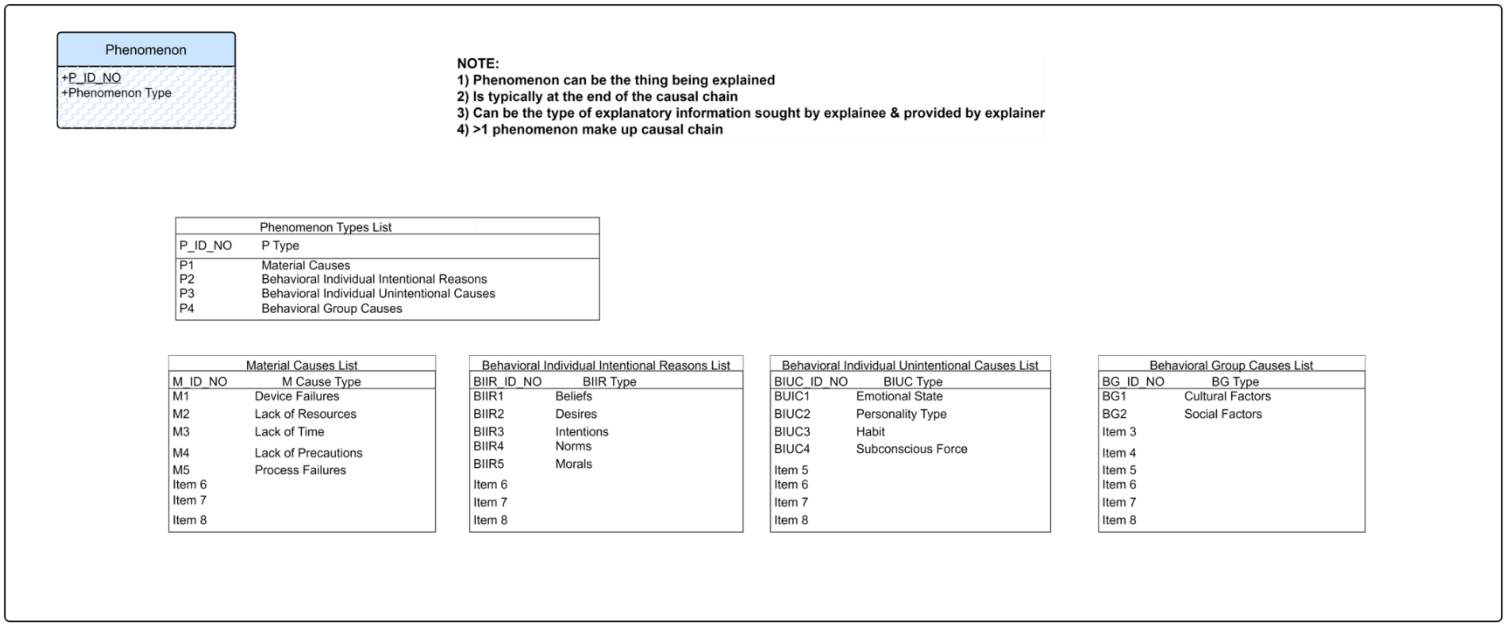


Figure 15:

Original SSH-Based Faceted Taxonomy of Explanation—The "HOW"

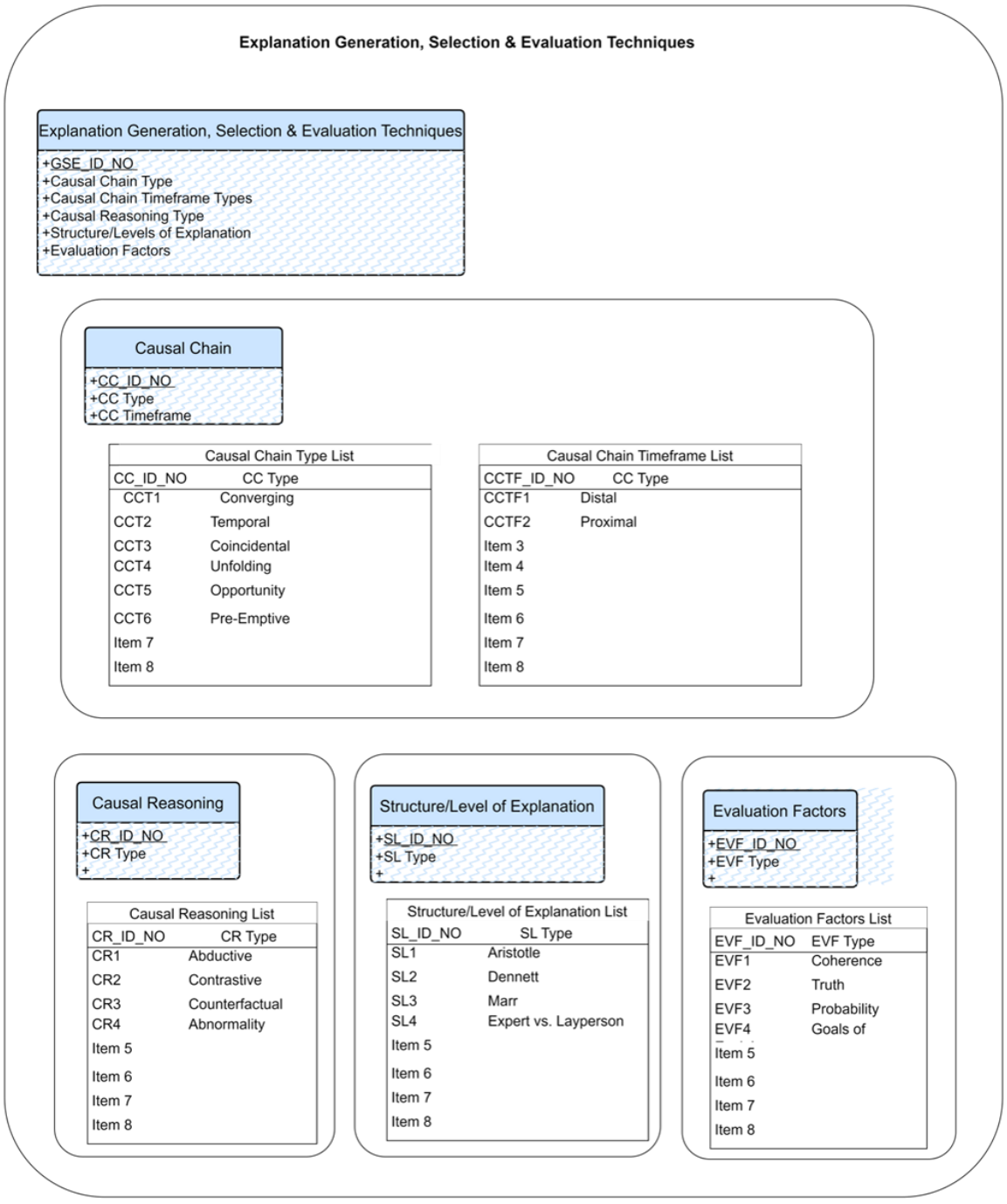


Figure 16:

Original Faceted Taxonomy of the xAI Landscape

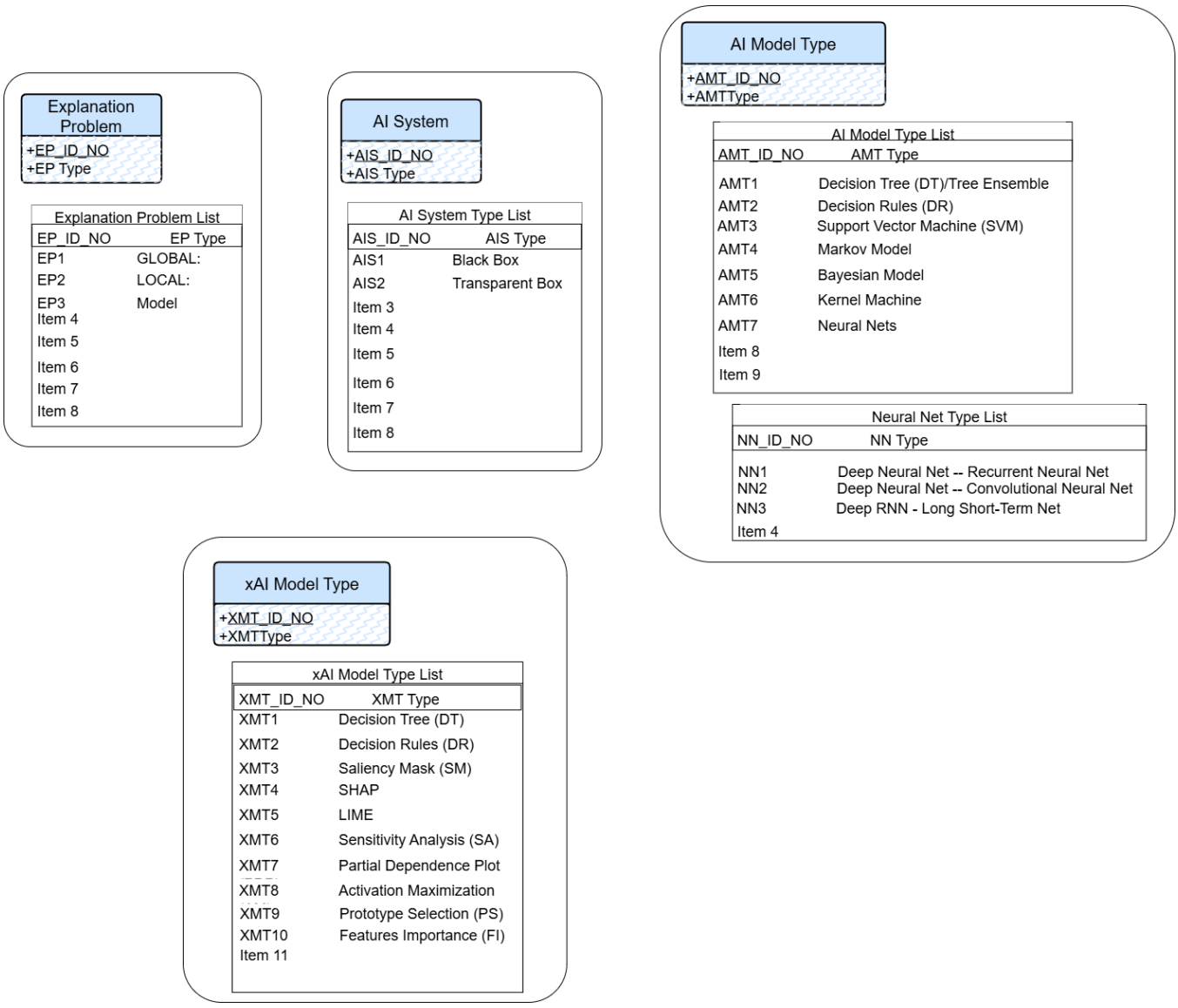
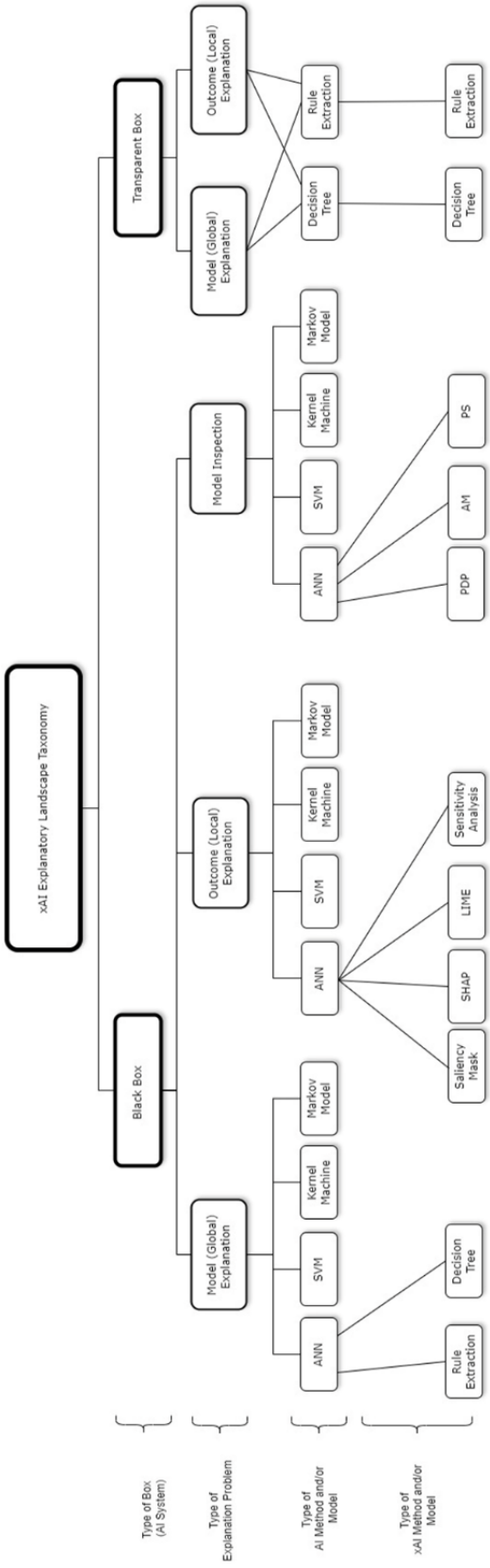


Figure 17:

Original Hierarchical Taxonomy of the xAI Landscape



Appendix C: Usage of Large Language Model (LLM)

Objective of Using Automated Software Tool:

As outlined above in section 2.2.3.1.2 above, to conduct qualitative data analysis, I chose to augment my manual analysis of the transcribed text with an automated software tool. I planned to harness a software tool with the objective of automatically generating data analysis reports in accordance with my specific instructions within a one-to-two-week timeframe. I sought to leverage the tool with the expectation that it would serve as a supplementary analyst which could perform cross-validation of manual analysis. In particular, I expected the LLM would improve efficiency and accuracy of the qualitative data analysis efforts.

Originally, I had planned to analyze the transcribed text by augmenting manual analysis with the use of computer software programs such as NVivo. I had planned to use coding words associated with the types of explanatory information as outlined in the literature review of SSH-based classes of explanation outlined in section 3.1. However, I found NVivo to be limiting and instead opted to use readily available Large Language Models (LLMs). Nvivo's limitations included its rules-based approach which resulted in its inability to understand the context of text data particularly if it involves nuances of language as well as its inability to identify patterns and themes in text; and its requirement for manual coding and categorization of text data which can be a time-consuming and tedious process.

In comparison, LLMs can process and handle large volumes of qualitative textual data with ease; better understand the context and subtleties of language; identify complex relationships between concepts such as entity recognition; automate the coding and categorization process and as a result, free up my efforts to focus on higher-level analysis and interpretation. Therefore, I opted to use an LLM to perform qualitative data analysis.

Upon consideration of various LLMs, I opted to use the readily available Google's Gemini (**Gemini 2.5 Pro Preview 05-06 gemini-2.5-pro-preview-05-06)**.

Choosing the LLM:

To facilitate the activities involved in qualitative data analysis of the transcribed interview text, the requirements that LLM are as follows:

1) The LLM must adhere to the guidelines and requirements of University of Ottawa's Research Ethics Board as they pertain to handling of sensitive, confidential and proprietary information. Due to the fact that this research project involved human participants, the data analysis must ensure that the participants' rights privacy and confidentiality guarantees are kept. Therefore, the analysis of all data pertaining to the transcribed interview text must be remain confidential and not be shared with or used by third parties for any purposes.

2) The computing capacity of the LLM includes the ability to process ~10-30 published academic papers plus conversational text (Word document ~10-25 pages in length).

3) The LLM must minimize the researcher's effort and time and enable the execution and completion of both manual and automated data analysis within 1-2-month timeframe maximum.

I set out in March 2025 to find an LLM that could satisfy the above requirements. In order to meet these requirements, I contacted uOttawa's IT department and was informed that uOttawa has a Microsoft CoPilot license with a version that resides within uOttawa's computing space and is a text-based LLM in that it does not ingest files.

Since this option did not meet criteria #2 above, I then investigated using an open source LLM on a local hardware host. This included using a local hardware platform such as MacStudio or using Alliance Canada's computers. To achieve results that would be close enough to commercial models such as ChatGPT, Google's Gemini, Anthropic's Claude, etc., I would have to manually build or augment existing hardware infrastructure that could handle the data I required, and this might include additional funds to do so. Also, I would have to spend a considerable amount of time to train the readily available open-source model such as LLAMA or DeepSeek. Since this option did not meet criteria #3 above, I investigated using a commercial LLM.

Upon consideration of various LLMs, I settled on Google's Gemini (**Gemini 2.5 Pro Preview 05-06 gemini-2.5-pro-preview-05-06)**. This LLM was readily available, free, and had a strong privacy policy. To safeguard all data and protect privacy, I turned off Google Gemini Apps activity in my Google account's settings. This ensured my conversations with Google's Gemini would not be reviewed by humans or used to improve Google machine-learning technologies. On top of adjusting the Google Gemini Apps settings, I took the following steps of manually editing the interview transcripts data: I deleted any and all comments and personal opinions expressed by the interviewee and/or the interviewer that were not within the scope of the research, namely, the topic of explanation. Once the pared down transcript was approved by the interviewee, I parsed the transcript further by a) removing all references to real people and places and then b) extracting only a subsection of that as it pertains to the topic of explanatory information that each stakeholder would be seeking. This final pared down transcript underwent full data analysis.

LLM Prompt Optimization and Tuning & Final Results

Prior to commencing the automated qualitative data analysis using Google's Gemini, I engaged in prompt engineering (prompt tuning and optimizing) and in-context learning activities to get the best possible output from this pre-trained LLM. The objective of prompt engineering was to ensure the LLM could generate accurate and detailed data analysis reports within a one-to-two-week timeframe. This involved iteratively refining the input prompts to improve the model's output to adapt it to the particular task of analyzing the transcribed text to identify categories of explanations sought by each individual. I crafted precise, strategic prompts to get better outputs and provided additional context and examples within the prompt to guide the model's response. I provided multiple prompts, fine-tuning them over time, to build confidence in the tool's ability to produce accurate results and analyze all the data. I compared my manual analysis of sample transcripts with the LLM's automated analysis to identify errors or discrepancies. The initial results of these activities were promising. Unfortunately, when I used the

tool to analyze all 26 transcripts and verified the accuracy of a random subset of its outputs, I found too many discrepancies. Despite 13 attempts with multiple prompts, including explicit instructions to follow prescribed analysis, the LLM produced spotty results. The main issue was its unreliability in terms of precision and accuracy. Gemini ventured off prompt and analyzed data in different formats, mixed up quotations of different stakeholders and attributed them to other stakeholders, etc.

The following prompt (Prompt#7) and input data (in pdf format) yielded the most usable results:

- 12 research papers that provided definitions for each category of the SSH-based faceted taxonomies: Grice (1975), Hesslow (1988), Hilton (1990), Hilton et al. (2005), Kass and Leake (1987), Keil (2006), Kahneman and Tversky (1982), Lewis (1986), Lipton (1990), Miller (2019), Thagard (1989).
- 2 files of faceted taxonomy of SSH-based explanation – pdf files of Figures 2a and 2b

PROMPT#7:

Attached are two files in pdf format titled "SS-Based _ Faceted Taxonomy of the What of Explanation" and "SS-Based _ Faceted Taxonomy of the How of Explanation" that provide a faceted taxonomy of explanation with each category having and "…ID_NO". Also, attached are 12 research papers in pdf format that the faceted taxonomy was based upon. Please use the 12 research papers and the two faceted taxonomy files to analyze the transcript of a conversation in the pdf format file titled "Stakeholder *#xx_Stakeholder Role(change this for each individual file fed into the LLM)* for Analysis.pdf".

As part of your analysis, propose categories and types of explanations. The categories must use "ID_NO"s from the faceted taxonomy that apply to the conversation, along with direct quotes from the conversation that justifies the categories and types of explanations that apply to the conversation.

Note that several of the 13 prompts that were used for prompt engineering activities included one additional file to exemplify an explanation analysis of a conversation involving Stakeholder1_Engineer that was in the format of Table 4, individual stakeholder analysis as described above in section 5.1.1. The LLM was given explicit results to follow the format provided and not deviate from the tabular format and ensure that it filled in as many fields and types of explanations as possible. Unfortunately, prompts that included an exemplar analysis, the results were worse than Prompt#7 above, therefore, I used the outputs of Prompt#7 to compare with and augment my manual analysis.

Because the LLM failed to repetitively perform tasks as prescribed, it yielded inaccurate and unreliable results. During the final data analysis activities of this project, this unreliability required me to manually check every detail of each stakeholder's transcript analysis data produced by Google Gemini, in addition to conducting my own manual analysis of each transcript. Initially, I expected the LLM to improve efficiency by automating analysis, freeing up time for higher-level tasks like interpreting results and making decisions. Unfortunately, due to its unreliability, I ended up manually checking every detail of each output from the multiple prompts, as well as conducting my own manual analysis of each stakeholder transcript. This added time to the process, but I considered it a collaborative effort between myself and another analyst, albeit an unreliable one, but one that produced some results. In the end, I decided to harness the benefits of human-computer collaboration by using the LLM as a cognitive assistant to support my qualitative data analysis. This approach enhanced the rigor and reliability of my analysis by providing fresh perspectives, validating my findings, and leveraging the LLM's ability to rephrase results in a distinct manner, ultimately augmenting the quality of my work. Notably, it did not reveal any significant new patterns or themes that I had missed, which reaffirmed my taxonomy of SSH-based Explanations.

Future Research Opportunities

Although not part of the scope of this research project, the unexpected findings of the limitations of the current Large Language Model (LLM) offerings for conducting qualitative data analysis provides future research opportunities.

Politicians, policy makers, and industry leaders have made sweeping pronouncements about generative AI's potential to improve productivity across all economic sectors, promising to assist designers, engineers, scientists, medical professionals, and researchers alike. The reality, at least for high-precision cognitive work, tells a more complicated story.

When I attempted to use LLMs to automate the qualitative data analysis portion of the FGCR testing, I expected them to improve efficiency by handling the repetitive processing of interview data, freeing me for higher-level interpretation and decision-making. Instead, they yielded inaccurate results, failed to perform tasks repeatably as prescribed, and ultimately required more time and human oversight than a traditional human collaborator would have. These are the opposite of the productivity gains promised.

These findings are not isolated. Karma (2025), Reevely (2025), and Kwa et al. (2025) all highlight the current inability of LLMs to deliver consistently reliable analytical results. A related and compounding problem is model drift, the tendency of LLM behavior to shift across sessions and over time, documented by Chen et al. (2023) and more recently by Dongre et al. (2025), which further undermines the repeatability essential to rigorous qualitative research.

The gap between the hype and the reality of LLMs for high-precision cognitive labour is wide and closing it presents fertile ground for future research. Further investigation is needed into how AI systems can be fine-tuned to assist with complex qualitative analysis without sacrificing the reliability and precision that rigorous scientific and engineering research demands.

Appendix D: Full Data Analysis Tables with Stakeholder Quotes

Table 8b

Individual Stakeholder Full Data Analysis for Stakeholder#1_Engineer (Non-AI Expert)

STAKEHOLDER (SH# & Description):	SH1_Engineer (Non-AI Expert)									SH1_Engineer (Non-AI Expert)										
	THE WHAT OF EXPLANATION									THE HOW OF EXPLANATION										
ACTIONS Executed by AI Models/ Algorithms Category	Quote(s)	P_ID_NO & TYPE	M_ID_NO & TYPE	BHR_ID_NO & TYPE	BHUC_ID_NO & TYPE	BG_ID_NO & TYPE	Comments			Quote(s)	GSE_ID_NO & TYPE	CC_ID_NO & TYPE	CCTF_ID_NO & TYPE	CR_ID_NO & TYPE	SL_ID_NO & TYPE	EVF_ID_NO & TYPE			Comments
PERCEPTION/ Detection	**Quote (Engineer 6:21):** "I think the first obvious question is how come ... uh... the AV couldn't see this stationary uh... let's say obstacle in general." **Quote (Interviewer 8:02):** "OK, so your question for the AI system is why didn't you see the obstacles?" **Quote (Engineer 9:24):** "...But the problem is why it doesn't detect it, why it didn't detect these ... these obstacles."	P1=Material Causes	M1.1 Device Fails to Adequately Perform Function	N/A	N/A	N/A	**DOES ASK ABOUT potential upstream failures in the AV's perception system (sensors) i.e. HARDWARE ASPECT OF PERCEPTION 15:42**			**Quote (Engineer 6:21-7:30):** "I think the first obvious question is how come ... uh... the AV couldn't see this stationary uh... let's say obstacle in general. In this case it's vehicles. But in general because the whole system I do understand is based on that I mean... It's identifying obstacles and of course avoid them. So, how come they didn't do, I mean, they didn't avoid these obstacles. It doesn't matter that doesn't matter if there are vehicles, first responder, vehicles, whatever, it's a volume on the road...Now my second question is why first responder vehicles? Because I understood that the whole cases are first responders, responder vehicles, no? ... Which is the primary function and ADAS should have. **Quote (Engineer 6:41-7:15):** "how come ... uh... the AV couldn't see this stationary uh... let's say obstacle in general. In this case it's vehicles. But in general because the whole system I do understand is based on that I mean...It's identifying obstacles and of course avoid them...Which is the primary function and ADAS should have." (Actual vs. Designed/Expected) **Quote (Engineer 53:57 - 54:04):** Inquiry about perception and then the "conclusion of this interpretation" implies a sequence.	GSE1=Causal Chain Type GSE2=Causal Reasoning Type GSE3=Structures/Levels of Explanation Type GSE4=Evaluation Factors GSE5=Other	CCT3=Unfolding	CCTF2=Proximal	CR2.6=Contrastive Designed/Specified/Expected Case	SL1=Aristotle	EVF2=Truth EVF1=Coherence		Could also be Data1-- questioning design Multiple contrastive questions: Quote: "I think the first obvious question is how come ... uh... the AV couldn't see this stationary uh... let's say obstacle in general. [Why X vs. not X?] Contrast=could not see (X) vs. could see Quote: " ..Now my second question is why first responder vehicles? [Why X vs. Y?] Contrast=First responder vehicles (X) vs. other types of vehicles Quote: "Because I understood that the whole cases are first responders, responder vehicles, no? Which is the primary function and ADAS should have. " [Why X vs. not X?] Quote (Engineer 7:5): "But how come a system that is mainly built to avoid obstacles, go crashing into obstacles?" [Why X vs. specified/designed/expected value for X?"] Contrast=not avoiding obstacles (X) vs. designed/expected to avoid obstacles Normal causes (ifs design and programming). Can also be characterized as SL4 Expert vs. Layperson or SL2 Dennett's design stance or SL3 Marr's Algorithmic Level

STAKEHOLDER (SH# & Description):	SH1_Engineer (Non-AI Expert)							SH1_Engineer (Non-AI Expert)							
	THE WHAT OF EXPLANATION							THE HOW OF EXPLANATION							
ACTIONS Executed by AI Models/ Algorithms Category	Quote(s)	P_ID_NO & TYPE	M_ID_NO & TYPE	BIIR_ID_NO & TYPE	BIUC_ID_NO & TYPE	BG_ID_NO & TYPE	Comments	Quote(s)	GSE_ID_NO & TYPE	CC_ID_NO & TYPE	CCTF_ID_NO & TYPE	CR_ID_NO & TYPE	SL_ID_NO & TYPE	EVF_ID_NO & TYPE	Comments
Motion Control (Steering and Braking) -- Decision Making & Taking Appropriate Action Failures: i.e. Decision-making & Action (given potential perception) Failure	**Quote (Engineer 9:24):** "The system, necessarily detected obstacles and avoid the . . . them, but not these ones...But the problem is why it doesn't detect it, why it didn't detect these ... these obstacles. And it went crashing then." (This shows a query into detection, but also implies a failure in the decision/action chain). **Quote (Engineer 21:28-21:57):** "OK. Why did you see? OK, I saw them. OK. Why didn't you avoid? ... How many, many, many possibilities, 2 main ones of course is changing direction or or braking , Why it doesn't happen? These are the questions.How many, many, many possibilities, 2 main ones of course is changing direction or or braking , Why it doesn't happen? These are the questions." **Quote (Engineer 54:04):** "And why the conclusion of this interpretation was just to.. to give up?" (Referring to the AI not acting).	P1=Material Causes	M1.3 Device Failure in Decision Making	N/A	N/A	N/A	Focus is on decision to continue driving and not taking action to avoid	**Quote (Engineer 52:03):** "I need just a list of of what happened. I mean chronology of OK, #1 sensors 1 send this information #2." (Although the interviewer states this log isn't available for a black box, the engineer's desire reflects wanting an unfolding causal chain). **Quote (Engineer 17:59):** how many obstacles this system avoided before getting to this one that it ignored. **Quote (Engineer 21:28):** "OK. Why did you see? OK, I saw them. OK. Why didn't you avoid?" **Quote (Engineer 54:04):** "And why the conclusion of this interpretation was just to.. to give up?" (Referring to the AI not acting).	GSE1=Causal Chain Type	CC3=Unfolding	CCTF2=Proximal				Engineer wants to understand the step-by-step process (or breakdown in process) from perception to action. Most questions focus on immediate moments leading up to the crash -- why didn't it see, why didn't it react, why it disengaged late? Also asking Distal questions about how many other obstacles were avoided by this system in the past as opposed to these 16?
			M1.2 Device Fails to Perform Appropriate Action						GSE2=Causal Reasoning Type			CR2.6=Contrastive Designed/Specified/Expected Case			Multiple Contrastive Reasoning: CR2.6=Actual inaction to avoid as opposed to expected action of avoiding. Could also be CR2.3 Non-common effect when asking distal questions about not avoiding these 16 vs. avoiding all other obstacles in the past ? This could also be construed as CR2.1 non-occurence of effect "why did not avoid vs. avoding obstacles? Could also be CR2.2 Normal the question highlights the abnormality of not avoindg and just giving up by not acting contrasting with normal/expected behaviore Also CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain
									GSE3=Structures/Levels of Explanation Type				SL1=Aristotle		Formal causes (its design and programming). Can also be characterized as SL4 Expert vs. Layperson. or SL2 Dennett's design stance or SL3 Marr's Algorithmic Level
									GSE4=Evaluation Factors					EVF1=Coherence	
									GSE5=Other						
System Disengagement (Timing, Process)	**Quote (Engineer 14:43):...** I don't understand how a second before it realises that there is a problem and try to give control to the driver by aborting the the control." **Quote (Engineer 54:04):** "And why the conclusion of this interpretation was just to.. to give up?" (Referring to the AI disengaging).	P1=Material Causes	M1.4 Device Fails to Engage/Disengage						GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal				
		P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions	N/A	N/A			GSE2=Causal Reasoning Type			CR4=Abnormality			CR4=Abnormality -- late (a second before crash) disengagement is abnormal and decision to abort when faced with a problem is also abnormal Also CR9=mechanistic causal reasoning seeking to identify necessary and salient components in the causal chain
									GSE3=Structures/Levels of				SL1=Aristotle		
									GSE4=Evaluation Factors					EVF1=Coherence	EVF1 & EVF2
									GSE5=Other						

STAKEHOLDER (SH# & Description):	SH1_Engineer (Non-AI Expert)							SH1_Engineer (Non-AI Expert)							
	THE WHAT OF EXPLANATION							THE HOW OF EXPLANATION							
ACTIONS Executed by AI Models/ Algorithms Category	Quote(s)	P_ID_NO & TYPE	M_ID_NO & TYPE	BIIR_ID_NO & TYPE	BIUC_ID_NO & TYPE	BG_ID_NO & TYPE	Comments	Quote(s)	GSE_ID_NO & TYPE	CC_ID_NO & TYPE	CCTF_ID_NO & TYPE	CR_ID_NO & TYPE	SL_ID_NO & TYPE	EVF_ID_NO & TYPE	Comments
Communication/ Warning to Human	**Quote (Engineer 10:51):** "Hmmm....uh ... Well for that, why? I don't know how it works, but does it notify the driver before ... before aborting the aborting the autopilot?" **Quote (Engineer 13:48):** "...why considering the fact that these obstacles are stationary and... and...supposing that the system saw them or detected them enough early before getting to this point, why it didn't notify the driver that yes, we're getting to a situation where probably the driver should take control." **Quote (Engineer 54:04):** "And why the conclusion of this interpretation was just to.. to give up?" (Referring to the AI not acting).	P1=Material Causes	M1.5 Device Communication Failure	N/A	N/A	N/A		**Quote (Engineer 13:48):** "...I mean normally it should be programmed to just uh...Just give control to the driver and notify the driver." **Quote (Engineer 15:42):** "Yeah Exactly. So there are there are multiple, there are multiple system component that are probably involved. First of all the sensors before getting to the algorithm." **Quote (Stakeholder1_Engineer 25:19):** "I ... I would be questioning everything because this is the system is what is. The algorithm, the learnings, the sensors, these, these are the three things that that collaborate all the time."	GSE1=Causal Chain Type	CC3=Unfolding	CCTF2=Proximal				Quote 13:48: CCTF2 Could also be characterized as CCTF1= Distal, namely, design time frame.
									GSE2=Causal Reasoning Type			CR2.2=Contrastive The Normal Case			Quote 13:48 = CR2.2
		P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions			Quote (Engineer 13:48): "...I mean normally it should be programmed to just uh...Just give control to the driver and notify the driver." Questions intention of designers & choices made by designers		GSE3=Structures/Levels of Explanation Type				SL1=Aristotle		Quote 15:42 and 25:19: The engineer, as an engineer, is not satisfied with a surface-level explanation. They want to understand the internal workings, sensors, algorithms, and learning processes. This echoes Miller's (2019) discussion of Aristotle's modes, where the engineer seeks efficient causes (how the system processed information) and formal causes (its design and programming). Can also be characterized as SL4 Expert vs. Layperson. or SL2 Dennett's design stance or SL3 Marr's Algorithmic Level
									GSE4=Evaluation Factors Type					EVF1=Coherence	EVF1 & EVF2
									GSE5=Other						
Design/Training of the Algorithm (More Distal Task Performed by AI developers)	**Quote (Stakeholder1_Engineer 22:32):** Now for me, and there's also another... variable.. which is the learning process." **Quote (Stakeholder1_Engineer 25:19):** "I ... I would be questioning everything because this is the system is what is ... the algorithm, the learnings, the sensors, these, these are the three things that that collaborate all the time."	P1=Material Causes	M2=Failure due to Design Flaw	N/A	N/A		How was this system trained before being put into the field?	**Quote (Stakeholder1_Engineer 28:26):** There is a there is a mismatch between the reality. I mean the situation and the learning.	GSE1=Causal Chain Type	CC3=Unfolding	CCTF1=Distal				Questions design
									GSE2=Causal Reasoning Type			CR9=Mechanistic Causal			Overall, seeking CR9=mechanistic causal chain information for all AI functions performed. Could also be CR2.3= non-common effect, comparing training situation vs. AI learning/training
									GSE3=Structures/Levels of Explanation				SL1=Aristotle		SL1=Aristotle (Formal cause – the nature/content of the training data)
									GSE4=Evaluation Factors Type					EVF1=Coherence	EVF1 & EVF2
									GSE5=Other						

Table 9b

Individual Stakeholder Full Data Analysis for Stakeholder#25_Police Officer

STAKEHOLDER (SH# & Description):	SH25_Police Officer																					
		THE WHAT OF EXPLANATION									THE HOW OF EXPLANATION											
ACTIONS Executed by AI Models/ Algorithms Category	Quote(s)	P_ID_NO & TYPE	M_ID_NO & TYPE	BIIR_ID_N O & TYPE	BIUC_ID_N O & TYPE	BG_ID_N O & TYPE	Comments	Quote(s)	GSE_ID_N O & TYPE	CC_ID_NO & TYPE	CCTF_ID_NO & TYPE	CR_ID_N O & TYPE	SL_ID_NO & TYPE	EVF_ID_N O & TYPE	Comments
PERCEPTION/ Detection	Quote (SH25_Police Officer 0:09:53): "...I guess the light and sirens were on or the lights were on so that the AI can interpret that the lights are on. But I guess maybe is the AI programmed to see lights and sirens and what that means to them, to the programming." Quote (SH25_Police Officer 00:11:35): "An AI computer can't comprehend if a fire truck is in a 45° angle in lane one, lane, two, lane three and then it kind of just shuts off 'cause... It's like, oh, I've never seen this." Quote (SH25_Police Officer 0:27:47): 'What about the camera system? Like you know the... you know how the lens or like the how far it can project?"	P1=Material Causes	M1.1 Device Fails to Adequately Perform Function M2=Failure due to Design Flaw	P2=Behavioura l/ Individual Intentional Reasons	BIIR3=Intenti ons		DOES ASK ABOUT potential upstream failures in the AV's perception system (sensors) i.e. HARDWARE ASPECT OF PERCEPTION Also questioning design choices made by designers of AV's perception system	Quote (SH25_Police Officer 0:09:53): "...They even say the natural... natural person when they're driving, when they see lights and sirens, their eye brings them to that area, and they drive into the cars. That's why we're taught to be safe when we're doing it. But I'm wondering is... it is it programmed in AI Autopilot and what does it tell it to do if there's lights and sirens on the road? Quote (SH25_Police Officer 00:11:35): "And then basically, when firefighters and police officers go to a scene, they're taught to go 45° and take up.... An AI computer can't comprehend if a fire truck is in a 45° angle in lane one, lane, two, lane three and then it kind of just shuts off 'cause... It's like, oh, I've never seen this."	GSE1=Causal Chain Type GSE2=Causal Reasoning Type GSE3=Struct ures/Levels of Explanation Type GSE4=Evalua tion Factors GSE5=Other	CCT3=Unfol ding	CCTF1=Distal CCTF2=Proxim al	CR2.3=Con trastive/ Non-Common Effect	SL4=Expert vs. Layperson	EVF1=Coher ence	Questions on capabilities of AI algorithm and whether it was designed properly Could be CR2.2 Norm = ability to recognize flashing lights and sirens as opposed to other normal conditions on road Also compares to humans having tendency to drive towards lights where their eye focuses on Also, CR2.3 is about contrasting to very specific way that first responders park at 45 degree angle at active collision scene and blocking more than one lane – is AI programmed to recognize such objects? Also, questioning physical hardware's capability (camera) to see far enough ahead could be construed as CR4=abnormality or CR1=abductive

STAKEHOLDER (SH# & Description):	**SH25_Police Officer**															
	THE WHAT OF EXPLANATION							THE HOW OF EXPLANATION								
ACTIONS Executed by AI Models/ Algorithms Category	Quote(s)	**P_ID_NO & TYPE**	**M_ID_NO & TYPE**	**BIIR_ID_NO & TYPE**	**BIUC_ID_NO & TYPE**	**BG_ID_NO & TYPE**	**Comments**	Quote(s)	**GSE_ID_NO & TYPE**	**CC_ID_NO & TYPE**	**CCTF_ID_NO & TYPE**	**CR_ID_NO & TYPE**	**SL_ID_NO & TYPE**	**EVF_ID_NO & TYPE**	**Comments**
Design/Training of the Algorithm (More Distal Task Performed by AI developers)	Quote (SH25_Police Officer 00:09:53): "...I guess the light and sirens were on or the lights were on so that the AI can interpret that the lights are on. But I guess maybe as the AI programmed to see lights and sirens and what that means to them, to the programming." Quote (SH25_Police Officer 00:11:35): "An AI computer can't comprehend if a fire truck is in a 45° angle in lane one, lane, two, lane three and then it kind of just shuts off 'cause… It's like, oh, I've never seen this."	P1=Material Causes	M2=Failure due to Design Flaw				Probing design failures		GSE1=Causal Chain Type	CCT6= Common-Cause Chain	CCTF1=Distal				
			M1.6 Device Failure in Information Processing						GSE2=Causal Reasoning Type			CR2.3=Contrastive Non-Common Effect			Could also be CR4=abnormality and CR1 Abductive looking at multiple hypotheses especially with respect to how first responders park their vehicles at collision scene
		P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions			Questions intention of designers & choices made by designers Does training data have flashing lights, sirens and also the way first responders park their vehicles at crash sites?		GSE3=Structures/Levels of Explanation Type				SL4=Expert vs. Layperson		
									GSE4=Evaluation Factors					EVF1=Coherence	
									GSE5=Other						
Motion Control (Steering and Braking) – Decision Making & Taking Appropriate Action Failures	Quote (SH25_Police Officer 00:13:04): I would say did you see the lights, sirens or did you see the lights and did you see the apparatus blocking the roadway and if so, why didn't you stop or I would be like, yeah, what was the reasoning you continued to go forward into the vehicles, did you not see them? And if you didn't see them, what were you doing? Quote (SH25_Police Officer 00:13:51):And that's what I… that's what I'm trying to… my question would be, I would think there'd be something wrong with the algorithm of why it didn't 'cause obviously it could. I've seen videos where it veers off quickly if someone falls onto the roadway, and the Tesla veers over, but why didn't at this moment is my question. What? What was it doing? What was it computing at that time that it didn't recognise this as a as a "brake" kind of thing or?	P1=Material Causes	M1.3 Device Failure in Decision Making						GSE1=Causal Chain Type	CCT3=Unfolding	CCTF2=Proximal				
			M2=Failure due to Design Flaw						GSE2=Causal Reasoning Type			CR2.2=Contrastive The Normal Case			CR2.2: Contrasting to normal case of pulling over and stopping or steering away as Teslas have done many times before. Could also be CR2.6 contrasting to designed/expected/specified case of previous actions by Tesla Could also be CR2.1 contrasting to not stopping.
			M1.6 Device Failure in Information Processing						GSE3=Structures/Levels of Explanation Type				SL4=Expert vs. Layperson		
			M1.2 Device Fails to Perform Appropriate Action						GSE4=Evaluation Factors Type					EVF1=Coherence	
		P2=Behavioural Individual Intentional Reasons		BIIR3=Intentions					GSE5=Other						