Universidad de Valladolid

Escuela de Doctorado Universidad de Valladolid

PROGRAMA DE DOCTORADO EN FILOSOFÍA

TESIS DOCTORAL:

TOMA DE DECISIÓN RESPONSABLE EN UNA INTERACCIÓN HUMANO-IA FIABLE: UNIENDO EXPLICACIONES CONTRAFÁCTICAS Y ARREPENTIMIENTO

Presentada por ROSA ESTHER MARTÍN PEÑA
para optar al grado de

Doctor/a por la Universidad de Valladolid

Dirigida por:

Sixto J. Castro Rodríguez

Christoph Benzmüller

Beishui Liao

Universidad de Valladolid

Escuela de Doctorado Universidad de Valladolid

DOCTORAL PROGRAM IN PHILOSOPHY

DOCTORAL THESIS:

RESPONSIBLE DECISION-MAKING IN TRUSTWORTHY HUMAN-AI INTERACTION: LINKING COUNTERFACTUAL EXPLANATIONS AND REGRET

Presented by ROSA ESTHER MARTÍN PEÑA

to apply for the degree of

Ph.D. from the University of Valladolid

Supervised by:

Sixto J. Castro Rodríguez

Christoph Benzmüller

Beishui Liao

" If you want the present to be different from the past, study the past"

Baruch Spinoza

DEDICATION

In loving memory of the two women of my life: my grandmother, María Calvo, who taught me the importance of observing, reading, being humble, and being thankful. My friend, Jing Gao, taught me the importance of fighting a disease until death with the mind, heart, and body. For their life teachings, thus, without them, I am not who I am.

DEDICATORIA

En cariñosa memoria de las dos mujeres de mi vida: mi abuela, María Calvo, quien me enseñó la importancia de observar, leer, ser humilde y agradecida. Mi amiga, Jing Gao, quien me enseñó la importancia de luchar contra una enfermedad hasta la muerte con la mente, el corazón y el cuerpo. Por sus enseñanzas de vida, pues sin ellas, no soy quien soy.

ACKNOWLEDGMENTS

Getting a Ph.D. is not easy, and those with the experience know it well. In fact, during these years, I thought I could not finish it on many occasions. Ultimately, it has been possible because of the people I have had by my side during these years. In addition, I am also very thankful for medicine, particularly the medication that has given life back to my lungs that deteriorated by the COVID-19 pandemic. These last three years have been full of people who have influenced my life deeply; some have passed away or suffered serious illnesses.

Dad, Tasio, I dedicate this research to you because you have always inspired me with your great heart, fight, and hard work. Mam, Encarna, thanks so much for always being by my side and believing unconditionally in me and my ideas. Tata, Palmira, thanks for keeping me fed during the pandemic with the ham and chorizo boxes and for all those handwritten letters encouraging me to keep writing. Matthias, thank you for giving a practical sense to this thesis, and I hope one can learn to manage the anticipated regrets. Uncle Jose Antonio, thank you for returning and teaching me that repairing machines can make people happy and humble.

Thanks to all my supervisors, who have not been few, not only for your academic teachings. Thank you, Professor Sixto Castro, for encouraging me in 2016 while walking along the river Pisuerga to do a PhD in Philosophy with you, even if the path was unexpected. Setbacks changed the course and direction of this doctoral thesis, and the pandemic gave it a specific order. Thank you, Professor Christoph Benzmüller, for being open-minded and welcoming me to the Free University of Berlin at the Institute for Mathematics and Computer Science in 2019. As a philosophy student with a very different background than yours but with something in common: our significant interest in AI and ethics. In addition, for allowing me to teach at the university, with a contract at the University of Bamberg, to finish my dissertation. Thank you, Professor Beishui Liao, for your teachings on Explainable AI at your lecture at the Free University of Berlin and our conversations on Confucianism, which have inspired part of this research.

I must also thank Professor Raúl Rojas, who introduced me to Christoph Benzmüller. Although Rojas has not officially supervised my Ph.D., he has been a tremendous mentor for me all these years, from 2019 until now, showing me the seriousness of mathematics and the importance of focusing on a specific aspect without drifting too much.

Thank you also for your financial support and for letting me be part of the Dahlem Center for Machine Learning and Robotics at the Free University in Berlin. I hope my relationship with robotics has only just begun!

In addition to these supervisors, I want to thank three professors from the University of Valladolid for their support and teachings. Thank you, Professor Ángel M. de Frutos Baraja, for having accompanied me in the first years of my doctoral studies since issues such as chaos theory have inspired my framework and my way of contemplating human nature and physics. Thank you, Professor José Mª Adán Sus Durán, for having also contributed with your interdisciplinary training in philosophy and physics to the direction of this extended research project. Finally, to finish with the supervision part, in this case, more informal, I want to thank Professor Javier Finat from the University of Valladolid as well for his attempt to explain to me in mathematical language the interdisciplinary theoretical framework that articulates this doctoral thesis, together with his moral support. Thanks to all of you! To my friends and colleagues. Thank you, Juana and Maik, for your support in Berlin and Cantalejo and for not losing confidence that I could finish this rocky journey. Thank you, Amy, for accompanying me on this linguistic and vital journey and for reviewing the language of this thesis. Thank you, Fang, for your encouragement and always believing in me even though I did not. See you in Singapore. Thank you, Iris Merget, for making AI and ethics a safe place for us and believing in my capabilities in these areas. Francisco, thank you for being a telecommunications engineer concerned about more human aspects and encouraging me with my path. We will be watching your progress.

All these events, together with the global pandemic, have shaped this dissertation ultimately like a source of inspiration and, at the same time, have brought about a massive change in me and how I want to invest my future. A big thank you from my heart!

AGRADECIMIENTOS

Hacer un doctorado no es fácil, y quienes tienen la experiencia lo saben bien. De hecho, durante estos años, en muchas ocasiones pensé que no podría terminarlo. Al final, ha sido posible gracias a las personas que he tenido a mi lado durante estos años. Además, también estoy muy agradecida a la medicina, en particular a la medicación que ha devuelto la vida a mis pulmones deteriorados por la pandemia de COVID-19. Estos tres últimos años han estado llenos de personas que han influido profundamente en mi vida; algunas han fallecido o han sufrido enfermedades graves.

Papá, Tasio, te dedico esta investigación porque siempre me has inspirado con tu gran corazón, tu lucha y por ser un gran trabajador. Mamá, Encarna, muchas gracias por estar siempre a mi lado y creer incondicionalmente en mí y en mis ideas. Tata, Palmira, gracias por mantenerme alimentada durante la pandemia con las cajas de jamón y chorizo y por todas esas cartas manuscritas animándome a seguir escribiendo. Matthias, gracias por darle un sentido práctico a esta tesis, y espero que uno pueda aprender a gestionar los arrepentimientos anticipados. Tío José Antonio, gracias por volver y enseñarme que reparar máquinas puede hacer feliz y humilde a la gente.

Gracias a todos mis supervisores, que no han sido pocos, no sólo por vuestras enseñanzas académicas. Gracias, profesor Sixto Castro, por animarme en 2016 paseando por el río Pisuerga a hacer el doctorado en Filosofía contigo, aunque el camino fuera inesperado. Los contratiempos cambiaron el rumbo y la dirección de esta tesis doctoral, y la pandemia le dio un orden específico. Gracias, profesor Christoph Benzmüller, por tener una mente abierta y acogerme en la Universidad Libre de Berlín, en el Instituto de Matemáticas y Ciencias de la Computación, en 2019 siendo un estudiante de Filosofía con una formación muy diferente a la suya, pero ambos con un interés significativo en la IA. Además de haberme permitido dar clases en la universidad, y haberme dado financiación con un contrato en la Universidad de Bamberg para terminar mi tesis. Gracias, profesor Beishui Liao, por sus enseñanzas sobre IA explicable en su conferencia en la Universidad Libre de Berlín y nuestras conversaciones sobre confucianismo que han inspirado parte de esta investigación.

También tengo que dar las gracias al profesor Raúl Rojas, que me presentó a Christoph Benzmüller, y aunque Rojas no ha dirigido oficialmente mi doctorado, ha sido un tremendo mentor para mí todos estos años desde 2019 hasta ahora, mostrándome la seriedad que aportan las matemáticas y lo importante que es centrarse en un aspecto concreto sin desviarse demasiado. Gracias también por su apoyo económico y por dejarme así formar parte del Dahlem Center for Machine Learning and Robotics de la Universidad Libre de Berlín. ¡Espero que mi relación con la robótica no haya hecho más que empezar!

Además de a estos supervisores, quiero dar las gracias a tres profesores de la Universidad de Valladolid por su apoyo y enseñanzas. Gracias, profesor Ángel M. de Frutos Baraja, por haberme acompañado en los primeros años de mis estudios de doctorado ya que temas como la teoría del caos han inspirado mi marco y mi forma de contemplar la naturaleza humana y la física. Gracias, profesor José Mª Adán Sus Durán por haber contribuido también con tu formación interdisciplinar en filosofía y física a la dirección de este extenso proyecto de investigación. Finalmente, para terminar con la supervisión, en este caso más informal, quiero agradecer también al profesor Javier Finat de la Universidad de Valladolid su intento de explicarme en lenguaje matemático el marco teórico interdisciplinar que articula esta tesis doctoral, junto con su apoyo moral. ¡Gracias a todos! A mis amigos y compañeros. Gracias, Juana y Maik, por vuestro apoyo en Berlín, en nuestro Cantalejo, y por no perder la confianza en que podría terminar este rocoso viaje. Gracias Amy, por acompañarme en este viaje lingüistico y vital y revisar el idioma de esta tesis. Gracias, Fang, por tus ánimos y por creer siempre en mí aunque yo no lo hiciera. Nos vemos en Singapur. Gracias, Iris Merget, por hacer de la IA y la ética un lugar seguro para nosotros y por creer en mis capacidades en estas áreas. Francisco, gracias por ser un ingeniero de telecomunicaciones preocupado por aspectos más humanos y animarme con mi camino. Estaremos atentos a tus progresos.

Todos estos acontecimientos, junto con la pandemia mundial, han dado forma a esta tesis doctoral en última instancia como una fuente de inspiración y, al mismo tiempo, han provocado un gran cambio en mí y en cómo quiero invertir mi futuro. ¡Muchas gracias de todo corazón!

SUMMARY

This dissertation addresses the issue of responsible decision-making in trustworthy human-AI interaction from an interdisciplinary perspective. This approach is motivated by the growing use of AI systems with increasingly autonomous decision-making capabilities. However, designing and implementing these AI systems with cognitive and affective human-like abilities to replace them in their decisions is not exempt from limitations and challenges as well as new opportunities.

In this way, to understand more about the risks and possibilities that this emerging scenario brings us, this dissertation presents, describes, and analyzes the fields of machine ethics and explainable AI, along with other advances in the areas of neuroscience and affective computing for their studies on the impact of emotions in human behavior. Thus, all these disciplines comprise the design proposal of the multi-ethical interdisciplinary framework for responsible AI. Thus, the proposed framework is divided into three levels in which human and artificial agents cooperate within goal-driven XAI. The aim is to create a theory of mind about the normative value of regret to prove whether the somatic marker hypothesis driven by the counterfactual component of the anticipated regret could serve as a recommendation norm for preventing unspecified errors before they occur.

Chapter 1 introduces the dissertation topic and justifies the motivations for conducting it. Chapter 2 provides an overview of the state of AI ethics. Chapter 3 discusses the research methodology employed. Chapter 4 discusses and presents the opportunities and constraints of the field of machine ethics for designing responsible AI. Chapter 5 deals with different approaches to Explainable AI and the phenomenon of biases in algorithmic systems and human behavior. Chapter 6 focuses on the current debate on theories about emotions and the possibilities and risks of using emotional data in affective computing. Chapter 7 concerns the ethical power of the human imagination in creating counterfactual scenarios by repressing those associated with negative emotional charges, such as regret. The last chapter of this research, Chapter 8, closes this dissertation by presenting the multi-ethical interdisciplinary framework for responsible AI in trustworthy human-AI interaction. The conclusions are in chapter 9.

Keywords: Machine ethics; Explainable AI; Goal-Driven Explainable AI; emotional data; bias and noise; counterfactual imagination; anticipated regret; trustworthy Human-AI interaction

RESUMEN

Esta tesis aborda la cuestión de la toma de decisiones responsable en una interacción de confianza entre el ser humano y la IA desde una perspectiva interdisciplinar. Este enfoque está motivado por el creciente uso de sistemas de IA con capacidades de toma de decisiones cada vez más autónomas. Sin embargo, diseñar e implementar estos sistemas de IA con capacidades cognitivas y afectivas similares a las humanas para sustituirlas en sus decisiones no está exento de limitaciones y retos, así como de nuevas oportunidades.

De este modo, para comprender mejor los riesgos y posibilidades que nos depara este escenario emergente, esta tesis presenta, describe y analiza los campos de la ética de las máquinas y la IA explicable, junto con otros avances en las áreas de la neurociencia y la computación afectiva por sus estudios sobre el impacto de las emociones en el comportamiento humano. Así todas estas disciplinas conforman la propuesta de diseño del marco interdisciplinar multiético para una IA responsable. Así, el marco propuesto se divide en tres niveles en los que los agentes humanos y artificiales cooperan dentro de la XAI orientada a objetivos. El objetivo es crear una teoría de la mente sobre el valor normativo del arrepentimiento para probar si la hipótesis del marcador somático impulsada por el componente contrafactual del arrepentimiento anticipado podría servir como norma de recomendación para prevenir errores no especificados antes de que ocurran.

El capítulo 1 presenta el tema de la tesis y justifica las motivaciones para llevarla a cabo. El capítulo 2 ofrece una visión general del estado de la ética de la IA. En el capítulo 3 se analiza la metodología de investigación empleada. El capítulo 4 analiza y presenta las oportunidades y limitaciones del campo de la ética de las máquinas para diseñar una IA responsable. El capítulo 5 aborda los distintos enfoques de la IA explicable y el fenómeno de los sesgos en los sistemas algorítmicos y el comportamiento humano. El capítulo 6 se centra en el debate actual sobre las teorías acerca de las emociones y las posibilidades y riesgos de utilizar datos emocionales en la informática afectiva. El capítulo 7 se refiere al poder ético de la imaginación humana para crear escenarios contrafactuales reprimiendo los asociados a cargas emocionales negativas, como el arrepentimiento. El último capítulo de esta investigación, el capítulo 8, cierra esta disertación presentando el marco interdisciplinar multiético para una IA responsable en la interacción humano-IA digna de confianza. Las conclusiones figuran en el capítulo 9.

Palabras clave: Ética de las máquinas; IA explicable impulsada por objetivos; datos emocionales; sesgo y ruido; imaginación contrafactual; arrepentimiento anticipado; interacción humano-IA fiable

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AMA	Artificial Moral Agent
BDI	Belief, Desire, Intention
DL	Deep Learning
KES	Kansei Engineering System
ML	Machine Learning
PeCox	Perceptual-Cognitive Explanation
RL	Reinforcement Learning
SCM	Structural Causal Model
SIM	Social Intuism Model
SMH	Somatic Marker Hypothesis
TM	Turing Machine
ToM	Theory of Mind
XAI	Explainable Artificial Intelligence
XGDAI	Explainable Goal-Driven Artificial Intelligence
XRL	Explainable Reinforcement Learning

LIST OF TABLES AND FIGURES

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION AND MOTIVATION

1.1 Artificial and natural intelligence

One of humanity's most significant challenges has been and continues to be understanding how the human mind works and what factors determine intelligence. Many definitions of intelligence as disciplines try to define it, but none can define it in universal terms. The reason might be the very nature of intelligence, which we can understand as an emergent phenomenon.

Delimiting it means diminishing it in some way. In this long and uneasy historical path, one of the most pragmatic solutions comes from artificial intelligence (AI). Researchers have been working for decades to replicate human intelligence, or at least certain aspects, with different objectives.

In recent years, I have gone from an interest in understanding human intelligence in general and how it might be understood in terms of artificial systems to a more pragmatic interest in ethical decision-making or what could be considered ethical and responsible behavior. Observing human behavior, I realized that our ability to make informed decisions is one of the hallmarks of our species. Why, in general terms, when speaking of intelligence, one refers to decision-making or behaviors that are considered correct in the way we solve (complex) problems.

However, how can we measure the correctness of a decision? Since decision-making occurs in a society, culture, or community of members, one way to understand this level of correctness of decisions could be under social norms. Cialdini and Trost define social norms as: "rules and standards that members of a group understand and that guide and constrain social behavior without the force of laws" (Cialdini & Goldstein, 2004).

Melnyk provides another definition that includes the relationship between social norms and values. According to this author, "these rules and standards entail the expected value of others that can be identified by observing their behavior" (Melnyk, 2011). However, although these social norms tend to regulate the uniform behavior of a given social group, they could often differ substantially among groups (Young, 2008).

Therefore, according to Young, "over time, norms changes could happen, due to objective circumstances or changes in subjective perceptions and expectations" (Young, 2008). But then, what happens when a norm or standard has ceased to have the value it had in guiding the behavior of a particular group? What kind of information or data can trigger the creation of new social norms?

The fact is that, since the advent of the internet, the information available in practically all spheres of our lives has exploded. Consequently, the range of behaviors in analog and digital worlds is becoming more complex because more data is available to make decisions.

However, the substantial shift is no longer caused by how such human agents deal with complex decision-making but by how AI-based algorithms increasingly operate autonomously, making decisions in fields previously only the territory of human intelligence or capabilities. Therefore, these AI technologies have started to replace us in specific domains, such as driving, through the introduction of autonomous cars or support human agents in their decision-making in a wide range of fields, such as medical diagnosis, hiring tools, and used for bail and sentencing decisions, to mention some examples.

However, the challenges we face as a civilization are not few since there is a clear gap in how these intelligent systems should behave so that they can be integrated and become part of our social structures, cultures, or communities. For AI systems to integrate more autonomously into society, it is required that they learn to identify what kind of behaviours are expected by them in the societies in which they are integrated. In short, the underlying question is to what extent or where these systems are required and how they should be implemented.

1.2 Combining ethical principles with AI principles

The issues regarding introducing AI systems, such as robots that replace or support humans in specific work areas, are not new, as they have been a recurrent theme in science fiction literature and films. One of the writers who has reflected on this topic to the point of linking the issue of machine behavior with a certain degree of autonomy and the ethics these intelligent systems must possess is Isaac Asimov.

He was a visionary who predicted some ethical challenges that autonomous artificial intelligence systems could face if we introduced them into societies.

His "Three Laws of Robotics" (Etzioni, 2017) deals with the contradictions related to robots acting in human physical and social environments.

The laws proposed by Asimov function as the fundamental obligations that every robot has towards humans and, therefore, towards society, and which, in order of priority, are as follows: 1). A robot may not injure a human being or, through inaction, allow a human being to come to harm; 2). A robot must obey orders given to it by human beings except where such orders would conflict with the First Law; 3). A robot must protect its existence if such protection does not conflict with the First or Second Law.

However, although Asimov's laws are often intrinsically vague, it is worth remembering that these rules have inspired AI researchers and philosophers who want to approach this challenge practically. As a result, the field of machine ethics, also called normative ethics principles for AI systems, appeared on the scene.

This field of machine ethics caught my attention because it is one of the research domains in which interdisciplinary teams of AI researchers and philosophers can work together. Therefore, machine ethics is concerned with adding an ethical dimension to machines. This crossover of disciplines highlights old discussions in the field of ethics and morality that remain unresolved. On the other hand, there is the technical side of implementing these ethical theories to respond to the growing challenges in AI.

One of the challenges that is also part of this dissertation's interests is to find ways to combine ethical principles permeated by the dominant philosophical doctrines of Deontology and Consequentialism in the West with other ethical principles from the East.

Arguments justifying this need to create multi-ethical frameworks that go beyond the most commonly used ethical frameworks in the West are studies such as the one conducted by Gold et al. These researchers based their research on finding cultural differences in responding to real-life and hypothetical trolley problems. The study of these researchers was based on testing whether moral intuitions should be generalized and whether psychological morality is universal. These researchers point out that trolley problems have been used in developing moral theory and the psychological study of moral judgments of behavior. However, most past research on this topic has focused on people from the West.

This approach may be why Gold et al. found apparent cultural differences in the approach to hypothetical trolley problems associated with differences in moral judgments and behaviors (Gold et al., 2014).

Therefore, the following question needs to be solved: How do ethical principles fit in a more contextualized way, considering or integrating the variety of existing behaviors and cultures in which they are embedded and operate? One way to address these limitations inherent in machine ethics is the proposal offered by authors such as Floridi and Cowls (2019). These authors dispense with the reference to ethical normative frameworks and prefer the term AI principles.

At this point, clarifying these two critical terms in the current state-of-the-art configuration of AI ethics is helpful. Then, according to Kim, Hooker, and Donaldson, ethical principles as defined here are thus established philosophical theories, which can be operationalized in reasoning capacities, as they imply certain logical propositions which must be valid for a given action to be ethical (Kim et al., 2020). On the other hand, AI principles are more general themes that an ethical principle may aim for in its application.

The following example proposed by Binns (2017) illustrates well this distinction between the two terms: while the principle of Egalitarianism supports the notion that human beings are in some fundamental sense equal, to work towards the AI principle of fairness, Egalitarianism may be operationalized by increasing efforts to avoid inequality. This example could be a rule that opportunities must be equally open to all applicants (Seng et al., 2021).

However, these concepts, which are encompassed as AI Principles, range from beneficence, non-maleficence, autonomy, justice, fairness, non-discrimination, transparency, responsibility, privacy, accountability, safety and security, explainability, human control of technology to the promotion of human values, among others, should underpin the design of AI technologies?

This question brings us back to the previous question: do these concepts have the same semantics in any context, understood as societies, cultures, or specific domains, or should we look for solutions where these general concepts have a specific function in a particular domain?

Examples make it easier to understand what is meant. As mentioned above, how the British and Chinese participants resolved specific unresolved dilemmas (Gold et al., 2014) makes it clear that the concept of responsibility that both cultures use and that permeate the social norms of these societies have different meanings. So, suppose we are to establish a principle of AI that has such an ethical component. In that case, the first thing to do is to understand how responsibility is gestated in different societies, bearing in mind that responsibility is an end in itself, but how it is expressed or manifested differs culturally.

1.3 Responsibility in Goal-Driven Explainable AI

As we have already mentioned in the previous section, of these AI principles that are part of the global framework of AI ethics, the one in which this dissertation is specifically interested is that of responsibility. However, as seen in the previous example, the perception of this same concept carries a different semantic load depending on the context in which it is applied. In normative language or from legal theory, this responsibility falls on whoever breaks the law.

However, there are more subtle ways to evade that responsibility, and on this issue, the work done by Kahnemanon cognitive biases (D. Kahneman, 2011) and later in his new book *Noise* (D. Kahneman et al., 2021) deserves mention. It is not because the hypothesis it shuffles is identical to the one this dissertation refers to but because it is a good starting point to understand the shortcuts and pitfalls human reasoners use to justify our behaviors to third parties. These human capabilities materialize in the two most common sources of error in deliberative processes, described to date as bias and noise. Kahneman et al. define bias as the average of errors and noise as variability (D. Kahneman et al., 2021).

Or said through several examples: a decision is biased when taken systematically against a thing, person, or group of people. In the case of noise, which is much more subtle and therefore more complicated to detect, it occurs when, for example, several doctors provide different diagnoses with the same information and data about the same patient.

Therefore, we should not be surprised that the concept of AI bias has become popular in such a short time. Setting AI systems to learn for themselves based solely on data provided by us has turned out to be quite a sociological experiment that opens the door to understanding more about how humans formulate their explanations to justify their behaviors.

The challenge now arises: Should AI systems, which work like black boxes, justify their decision-making like humans? The fact is that achieving explainability by imitating how humans offer explanations to date is not free of impediments or barriers. On the one hand, some AI systems operate as black boxes. Nevertheless, this lack of transparency is not only a phenomenon of artificial agents; humans also lack it. This argument is called "the limitations of human cognition" (Maclure, 2021).

When a human agent justifies or explains a decision, they produce a post hoc story supporting their choice. This argument is similar to the critique by the authors Dan Sperber and Hugo Mercier in their book *The Enigma of Reason: a New Theory of Human Understanding*. They write: "We produce reasons to justify our thoughts and actions to others and to produce arguments to convince others to think and act as we suggest" (Mercier & Sperber, 2017).

Based on these limitations, the first challenge is what kind of explanatory capabilities should be implemented in artificial agents. For example, Sado et al. state this argumentation in the same line: "What a user considers a good explanation is not the same as how helpful the explanations are" (Sado et al., 2020). In the same vein, the second challenge we have to solve is how this explanatory component in an AI system can contribute to reducing or making these sources of errors visible in human judgments. The objective is to reinforce certain aspects of goal-based behavior to improve ethical decision-making in human-AI interaction.

Specifically, the emerging subfield of Explainable AI, eXplainable Goal-Driven Artificial Intelligences (XGDAIs), seems much more suitable than the traditional way of understanding explainability as data-driven XAI to achieve this aim. In the domain research of goal-driven XAI, the AI systems also seek to justify their behaviors to lay users (Anjomshoae et al., 2019) with the peculiarity that the explanations contain a well-defined theory of mind oriented to achieve a specific behavior.

In the specific case of this research, the aim is to implement strategies where the explanatory component of the AI system can communicate the need for action planning to prevent unspecified errors before they occur.

At this point, it is worth highlighting why this research differs in how emotions are understood as the primary sources of error in human judgments (D. Kahneman et al., 2021).

In other words, the ignorance and even aversion to considering that emotions play a significant and essential role in human deliberative processes have led to their being labeled, in most cases, only as sources of errors in human judgment. Although this argument may be understandable and partly shared, it is not desirable for all cases involving emotions in the behavioral sciences.

On the other hand, it is not the purpose of this thesis to test whether one of the most well-founded theories of emotions so far is true, but what ethical challenges we face in each of them.

Moreover, to be able to carry out this research, we need to deal with only one emotion: regret. Regret has been studied extensively by researchers in social cognition and behavioral decision-making. One of the many experiments about regret was conducted by Gilovich and Medvec (1994), who concluded that "a sense of personal responsibility is central to the experience of regret."

Therefore, if the experience of regret is linked to a sense of personal responsibility, how could it be helpful to manipulate this emotion to reinforce certain behaviors and avoid others? In addition, as explained above, responsibility is considered one of the AI principles.

However, how would it be materialized as an ethical normative principle? Could, in some sense, the emotion of regret be the cause of the emergence of a new social recommendation norm?

1.4 The rule of reinforcing anticipated regret as AI principle

Above, it has been commented, on the one hand, that the same concepts and even ways of solving trolley problems and ethical dilemmas seem not as universal as we previously thought (Gold et al., 2014).

This research project hypothesizes that the role of emotions in deliberative processes must be considered in the development of morality.

The somatic marker hypothesis developed by the neuroscientist Antonio Damasio (1994) is a good starting point for understanding how emotions directly influence human deliberative processes.

"Since the emotion of regret is associated with having a sinking feeling, thinking about what a mistake one has made and about a lost opportunity, feeling the tendency to kick oneself and to correct one's mistake, actually doing something differently, and wanting to have a second chance and to improve one's performance. Hence, we conclude that the experience of regret involves focusing on the self as a cause of the event and on possibilities for undoing the regret by changing the unfavourable outcome by improving future performance" (Zeelenberg, Van Dijk, Manstead, et al., 1998).

Hence, the somatic marker hypothesis in the specific case of this research might be formulated as follows: these somatic markers work with our secondary emotions, which in this case is regret, and serve to guide our decision-making by discarding specific scenarios and choosing others. Zeelenberg et al. explain very well in this paragraph the value that regret has in decision-making and its relation to responsibility:

Nevertheless, the somatic marker hypothesis may also explain the decision-making impairment characteristic of patients with brain damage (Aday et al., 2017). This hypothesis supports that a defective activation of somatic states (biological ingredients of emotional signals) is the main reason for impaired behavior.

According to Aday et al., "the functional role of these somatic states is that they attack value to given options and scenarios and mark them as having potential positive or negative consequences in the future. These emotional (somatic markers) are covert or overt biases for guiding decisions. Deprived of these emotional signals, patients may resort to deciding based on the immediate reward of an option" (Aday et al., 2017).

However, the relevant aspect for this dissertation of the somatic marker hypothesis described by Damasio is that these somatic markers function only under two conditions: they can function properly if the individuals have no brain damage or, in other own words: "they are normal" (Damasio, 1994). On the other hand, the society and culture in which they are integrated must be normal, in the sense of healthy, in Damasio's terms. Although this statement is too generalistic and is mentioned only in a few lines in his book, this issue deserves special attention.

How do we measure these somatic markers objectively? (Feldman Barret, 2017) Is it possible today? (R. W. Picard, 1995) What does the somatic marker have to do with a normality associated with a culture? These questions lead us to the next important element of the emotion of regret.

According to Buchanan et al., this emotion has two components: an affective component, which is the one explained in the previous paragraph (Zeelenberg, Van Dijk, Manstead, et al., 1998), and a cognitive component, which is the one that corresponds to the capacity for counterfactual reasoning (J. Buchanan et al., 2016). This counterfactual reasoning seems almost part of our daily routines because, according to Kahneman and Tversky, "people often imagine how things could have turned out differently, especially after a bad outcome" (Tversky & Kahneman, 1986).

However, while the content and form in which these counterfactual thoughts are generated are not the same in all people, they have at least one affective component in common. This type of reasoning amplifies negative emotions such as regret and guilt. And the question here is: who is willing to pay this emotional price? Are we trained in this matter? Are we even aware of what this implies for learning processes, repairing mistakes, and feeling responsible for a bad outcome? In fact, according to Byrne, "people sometimes opt to inhibit counterfactuals" (Byrne, 2019). Furthermore, they decided not to be informed about the outcomes of unchosen options more often after significant losses than small ones (Tykocinski & Steinberg, 2005).

In any case, if this dissertation aims to try to remedy a set of errors before they occur, this way of counterfactual thinking does not work to achieve this goal. Nevertheless, regret is also a forward-looking emotion. According to Sanna (1996), prefactual thinking arises when making crucial decisions. That is when we tend to think about how we would evaluate our outcomes in light of forgone outcomes.

In particular, we focus here on regret because feeling it in a "simulated" way before a decision concerning future scenarios opens the door to changing our decision-making in various ways. However, one way to solve this challenge is to use model-reinforcement learning techniques to influence the goal-based aspect of the behavior, which we want to improve.

So, one of the research issues that this dissertation tries to answer is formulated as follows: Given that human agents tend to block part of the counterfactual thoughts that have a negative affective charge, such as the feeling of regret or responsibility for a bad outcome, how can an explanatory component of the goal-driven AI system reinforce the anticipation of possible regrets as a mental simulation before they occur?

CHAPTER 2

STATE OF ART IN AI ETHICS

2.1 Authority of (automated) decision making

The steam engine was one of the most important milestones of the Industrial Revolution and human history. At the time, it was a type of device consisting broadly of a box with levers and gears, which provided power to factories, locomotives, and ships. Simple power governors regulated these machines; they were the prelude to those that would follow.

Around the 1940s, we arrived at the next level of automation with computers and the abstract model of the Turing machine (TM). The Cambridge mathematician is the father of computer science precisely because the TM model incorporates the hardware and software of modern computers at an abstract level. It has inspired the design of what can be considered today's intelligent machines (Muggleton, 2014).

Nowadays, machines are achieving a new status as they have moved from supporting human decision-making processes to making decisions themselves (Tamò-Larrieux, 2021).

This almost Copernican shift in the conception of the machine and its expected behavior has begun to have a severe social impact at the practical, functional, theoretical, and legislative levels. Therefore, the fundamental question is based on the relationship between big data and algorithms. That is, how algorithms empower this data with direction and purpose. In Pasquale's words, "critical decisions are made not based on the data per se, but based on data analyzed algorithmically" (Pasquale, 2015).

Algorithms constitute the core of the decision-making parts of code. The current discussion focuses on the possibility of algorithms to make decisions without any human intervention, and Beer launches this appropriate question: "Should we treat the algorithms as lines of code, as objects, or should we see them as social processes in which the social world is embodied in the substrate of the code?" (Beer, 2017).

As human constructions, algorithms are part of the social machinery, and "their existence and design are a product of social forces, as are their implementations and redesigns" (Beer, 2017).

Therefore, algorithms as agents, machines, intelligent systems, or others already have authority because we perceive them as such, and we have introduced them in our social networks for at least the last two decades.

The power of algorithms lies in their ability to offer choices (recommender systems), classify, order what is essential, and, above all, decide what information should be visible to the user. This concern is not new. Pariser explains in *Filter Bubbles* that algorithms, instead of contributing to an open society of possibilities, continually expose people to the same kind of limited ideas, experiences, news, and culture (Pariser, 2012).

The authority of algorithms, we can prove it in their ability to produce truths through the kind of empowerment that they have on society (Bauman & Haugaard, 2008). This overall need made algorithms emerge as mediators and organizers, almost creators of new values, permeating the concept of normality.

In the same vein, we have the theory of Taleb, which he developed in his book *The Black Swan: The Impact of the Highly Improbable.* He starts with the idea that we cannot predict the future but that there are several ways of reducing the uncertainty of the forecast. These are the artistic and scientific paths, together with the techno-scientific axis. Taleb even speaks of a fourth path, which is the media. Taleb says all these initiatives are fueled by the human need to reduce and condense everything around us because it contains millions of details. In the author's words: "Without these channels, one is condemned to navigate in a universe of absolute and unbearable uncertainty." Moreover, these pathways would be narratives materialized in political, scientific, economic, and sociological discourses to spare us from the world's complexity and protect us, as he says, from its "randomness" (Taleb, 2007).

If Taleb is correct in his hypothesis, we can say that the authority given to algorithms lies in the human need to reduce uncertainty, amplified in recent decades by the sheer amount of information available. However, such authority seems to depend on reducing uncertainty and how automated decision-making can be perceived.

As Beer points out (2017): "the notion of the algorithm is part of a wider vocabulary, a vocabulary that we might see deployed to promote a certain rationality, a rationality based upon the virtues of calculation, competition, efficiency, objectivity, and the need to be strategic (…)". In this

way, the algorithm's power may not just be in the code, but in what way it becomes part of a discursive understanding of desirability and efficiency in which the mention of algorithms is part of "a code of normalization" (Hook, 2007).

The trust in machines' decisions strengthens our belief that they have made the right choice. On the one hand, automated decision-making is seen as superior to human decision-making and is more strongly relied upon. On the other hand, the lack of "humanness" challenges how individuals feel about an automated decision-making process" (Tamó-Larrieux, 2021).

Therefore, a paradox arises: although we can perceive machine decision-making as objective and free from the prejudices and biases characteristic of humans, the virtues that we associate with interhuman relationships, such as the ability to listen to each other and empathize, are removed (Zarsky, 2016). This dilemma leads directly to another of the critical aspects of automatic decision-making, which is accountability.

For the time being, although the long-term goal of artificial intelligence systems is to imitate and augment human intelligence, we should not forget that these intelligent systems are no more than the result of preprogrammed mathematical operations. This aspect is essential because otherwise, "we might wrongly "blame" automated decision-making systems, instead of holding the institution in which they are developed responsible for their programming and data selection choices" (Re & Solow-Niederman, 2019).

2.2 Machine ethics: designing ethical decisions

Autonomous machines are replacing humans in many tasks. The consequence is that our expectations about these intelligent systems are that they can act "in morally appropriate ways" (Tolmeijer et al., 2020) when faced with important decisions that have repercussions for third parties, or put in other words: "The greater the freedom of a machine, the more it will need moral standards" (R. W. Picard, 1995).

The term machine ethics was first used in 1987 by Mitchell Waldrop in the AI Magazine article: *A Question of Responsibility*. He provided the most accepted definition, which in some way constitutes the starting point of machine ethics as the discipline "concerned with the consequences of machine behavior toward human users and other machines" (Waldrop, 1987).

So, while the discipline is gaining popularity, other authors cannot guarantee that they will identify the inherent limitations. For example, Gert (2004) argues that humans and machines cannot improve their moral behavior in many situations by following the prescriptions of something akin to an algorithm. Indeed, there is widespread agreement on some moral rules (Gert, 2004).

In the same line are the assumptions developed years later by Brundage (2014). He states that "these moral rules are often ambiguous, and sometimes we need to break them, but there is a persistent disagreement about the conditions in which we must make such exceptions. There is a broad consensus that some specific ethical domains remain problematic despite the best efforts of philosophers" (Brundage, 2014).

This "unsolved" nature of ethics is not due to insufficient rational analysis but is a reflection that the intuitions at the base of our ethical theories are unsystematic at their core. This particular aspect leads to the question of the feasibility of machine ethics, and we need "a thorough reconceptualization of who or what should be considered a legitimate center of moral concern and why" (Gunkel, 2020). Another limitation is obeying rules without the possibility of exceptions, as Winograd has exemplified with his "the bureaucracy of mind."

He argues:

> "When a person views their job as the correct application of a set of rules (whether human-invoked or computer-based), there is a loss of personal responsibility or commitment. The "I just follow the rules" of the bureaucratic clerk has its direct analogue in "That's what the knowledge base says. "The individual is not committed to appropriate results, but to the faithful application of procedures" (*Winograd - Thinking Machines*, 1990).

Coeckelbergh (2010) follows the same argument as Winograd (1990). Still, he illustrates this phenomenon as a mental disorder, specifically in some manifested in psychopathological personalities. This term designates a type of personality disorder characterized by an abnormal lack of empathy whose condition is difficult to notice because it is complicated to detect by an ability to appear normal in most social situations. “These psychopathic machines would follow the rules but act without fear, compassion, care, and love. This lack of emotion would render them non-moral agents who cannot discern what is valuable. They would be morally blind” (Coeckelbergh, 2010).

Another of the weaknesses of machine ethics, or instead of the literature about Artificial Moral Agents (AMAs), is that most of the work done is by either deontological or utilitarian frameworks (Vallor, 2018; van Wynsberghe & Robbins, 2019). In contrast to the Eastern philosophical approaches to robot or machine ethics that focus on “defining what the good is” and worry about “how one can come to know the good” (Zhu et al., 2019), Chinese philosophers represented by Confucian scholars are more interested in the problem of “how to become good” (Ivanhoe, 2000).

In the face of such challenges, even if these artificial moral agents were verifiable and verified, that would not be enough to trust an autonomous system. Therefore, it needs to be designed with the ability to explain its decisions and should thus be supplemented with a “Machine-Explanation component” to control for possible bad outcomes” (Gunkel, 2017).

2.3 Explainability: the challenge of persuading with reasons

Some of the challenges automated decision-making systems face include that they may be systematically biased “against a particular class of people” (Jain, 2018). Another example would be AI algorithms that “can invade our privacy by inferring information about aspects of ourselves that we did not wish to disclose by correlating data points that are not legally considered personal information” (Wachter & Mittelstadt, 2018).

Trying to provide some response to these emerging issues, the field of Explainability in Artificial Intelligence (AI), or explainable artificial intelligence (XAI), has emerged as a topic of research “by the need of conveying safety and trust to users in the “how” and “why” of automated

decision-making in different applications such as autonomous driving, medical diagnosis, or banking and finance" (Confalonieri et al., 2021).

Article 29, "Data Protection Working Party's draft guidance" of the European Union's General Data Protection Regulation (the GDPR), from May 25, 2018, states that "a complex mathematical explanation about how algorithms or machine-learning work," though not generally relevant, "should also be provided if this is necessary to allow experts to verify further how the decision-making process works[1]" (Zerilli et al., 2019).

The field of explainable AI or explainability comes to respond to the problem of the lack of transparency of automatic decision, hence the question: "What sorts of explanations can we expect from automated decision systems, and will such explanations be good enough"? (Zerilli et al., 2019). Explainability or opacity was not a significant issue for traditional algorithms, or at least not of the same type that current machine learning (ML) models and artificial neural networks pose. That is because traditional algorithms had their rules and weights prespecified "by hand" (B. D. Mittelstadt et al., 2016), and there was nothing the system could do that was not already factored into the developer's design for how the system should operate given certain input conditions[2].

Philosophers and cognitive scientists drawing on phenomenology say, "the success of symbolic AI was mainly limited to virtual and contained environments such as games and logical puzzles" (Dreyfus, 1978; Varela et al., 2016). The inherent limitations of traditional hand-crafted algorithms led to a growing interest in whether machines could learn independently, with the ability to perform different cognitive tasks inspired by how the human brain works.

A leap took place. Machine learning models and artificial neural networks emerged from the "connectionist" paradigm in cognitive science and AI (Stuart & Peter, 2016). Maclure (2021) describes this new paradigm in AI systems in this way: "Artificial neural networks are in at least a superficial sense inspired by how neurons activate and are connected through synapsis in biological brains."

[1] This recommendation pertains to Article 15 of the GDPR. Article 15 requires the disclosure of "meaningful information about the logic involved" in certain kinds of fully automated decisions.

[2] Traditional algorithms, like expert systems, could be inscrutable after the fact: even simple rules can generate complex and inscrutable emergent properties.

Machine learning algorithms are massively inductive. “During training, a deep learning (DL) system adjusts the weights of these links to improve its performance. If trained on a decision task, it essentially derives its decision-making method, much as we would expect of an intelligent system. But there is the rub. In neural networks, these processes run independently of human control, so transparency inevitably becomes an issue: it is not known in advance what rules will be used to handle unforeseen information” (Zerilli et al., 2019).

While machine learning subsymbolic models are winning the battle for accuracy, they are losing the battle for transparency. However, this opacity problem is not only characteristic of deep ML systems because, as Maclure (2021) points out: “deep artificial neural networks are not significantly more opaque than the human brains/minds.”

He calls this the argument “from the limitation of human cognition,” adding as well that “as was abundantly shown by researchers in fields such cognitive science, social psychology, and behavioral economics, real-world human agents are much less rational than imagined by either some rationalist philosophers or by rational choice theorists in the social sciences.”

Therefore, this lack of transparency and opacity is also a hallmark of human reasoning or the human mind, and it is worth exploring further the kinds of explanations that can contribute to such transparency.

Pointing in the same direction about the challenges facing explainability, Páez offers new directions to understand the term Explainable AI in his article *The Pragmatic Turn in Explainable Artificial Intelligence (XAI).* He argues, "The purpose of explaining a model or decision is to make it understandable to stakeholders. But without a last grasp of what it means to say that an agent understands a model or a decision, the explanatory strategies will lack a well-defined goal” (Páez, 2019).

Another weak point in current explanatory models of AI black boxes is that many researchers build explanatory models for themselves rather than the intended users, a phenomenon called “the inmates running the asylum” (Miller et al., 2017). In the same vein, we have this hypothesis about understanding versus explanation: “Much of empirical inquiry consists in physical and intellectual activities that generate causal information, such as observation, experimentation,

manipulation, and inference (…) And these activities are distinct from giving and receiving explanations" (P. Lipton, 2009).

In addition, if we examine the information that drives machine learning today, we find it almost entirely statistical. Learning machines improve their performance by optimizing parameters over a stream of sensory inputs from the environment, and "No learning machines in operation today can answer reliably questions about situations not encountered before" (Pearl, 2018).

2.4 Behavior regulation: linking counterfactuals and regret

Many may wonder whether the pandemic's devastating effects might have been milder if we had acted differently earlier. When we ask ourselves more personal questions, such as: if I had studied something else, I would be happier now; if I had listened to my mother, I would not have come to this country. This way of thinking "about what might have been, about alternatives to our pasts, is central to human thinking and emotion" (Epstude & Roese, 2008).

They constitute mental representations of alternatives to past events, actions, or states (Byrne, 2016; Epstude & Roese, 2008). This kind of reasoning about "what might have been" implies the contraposition of an imagined against a factual situation. Counterfactual derives from philosophical writings in which the logical status of possibility and probabilistic reasoning have been closely scrutinized (Epstude & Roese, 2011; Evans et al., 2007; Roese & Epstude, 2017; Stern, 1981). In their usual embodiment, counterfactuals form a conditional proposition in which the antecedent corresponds to an action, and the consequent is an outcome. Crucially, "counterfactual thoughts are often evaluative, specifying alternatives that are in some tangible way better or worse than actuality. Better alternatives are termed "upward counterfactuals; worse alternatives are termed downward counterfactuals" (Markman et al., 1993; Roese, 1994). When upward counterfactuals focus on personal choice, the resulting emotion is termed regret (more about that below).

So, what is the purpose of counterfactuals? Counterfactual thinking's primary function centers on managing and coordinating ongoing behavior. Thinking about "what might have been" influences performance and expedites improvement, activating several distinct mechanisms.

Counterfactual thoughts are deeply connected to goals and are a regulatory component that keeps behavior on track, particularly within social interactions (Epstude & Roese, 2008; Markman et al., 2006; Roese & Epstude, 2017).

However, the definition of counterfactuals now is different from the past. The first theory that tried to explain the role of counterfactuals in human reasoning was "norm theory" (D. Kahneman & Miller, 1986). Norm theory belongs to the tradition of heuristics and biases, portraying counterfactual thinking (a "simulation heuristic") as a form of biased judgment and decision-making.

However, this dissertation focuses on the functional perspective of counterfactual reasoning as theorized by Epstude and Roese (2008) "as a useful, beneficial, and utterly necessary component of behavior regulation."

In addition, "a failed goal typically activates counterfactual thoughts, and they specify what one might have done to have achieved that goal" (Markman & McMullen, 2003; Roese et al., 1999). This way of regulating decision-making has even reached the field of AI; in particular, it has landed with force in the field of explainability (Byrne, 2019).

However, the number of counterfactuals we can manipulate to explain any event is potentially limitless, and it is a non-trivial problem to identify which counterfactuals or, better said, what kind of mechanism influences how humans construct counterfactuals and which of them could be helpful to ~~in~~ Explainable AI. These discoveries might be valuable in the design of a proper Theory of Mind (ToM) in an explanatory goal-driven AI system (Sado et al., 2020). According to Wellman Cross and Watson, ToM is defined "as the general ability to understand one's mental states, such as intentions, emotions, and desires, among other functions, that preface our actions, and to comprehend that others possess the same ability" (Wellman et al., 2001). Of all these elements that make up what is known as the theory of mind, emotions are of interest to this research work.

In this dissertation, we focus only on regret, and the main reason is that this emotion is strongly related to avoiding future errors from the self-agency perspective or, in other words, the ability to feel responsible for the outcome of a decision.

At this point, it should also be argued why the emotion of regret and not that of disappointment have been chosen, as both emotions seem to be interpreted as similar when, in fact, they are not. Hence, while the feeling of regret involves a focus on the self as a cause of an event and is related to the self-responsibility of repairing errors, the emotion of disappointment, also known as a counterfactual emotion, differs from regret in its "experiential content" (Zeelenberg, Van Dijk, Manstead, et al., 1998).

Disappointment arises from frustration but gives people the feeling that "they are not always able to control their destiny, and that they perceive a lack of control" (Zeelenberg, Van Dijk, Manstead, et al., 1998).

Then, regret shapes multiple aspects of decision processes, from avoidance of decisions to shifting responsibility for the decision to reframing decision alternatives (Pieters & Zeelenberg, 2007). Regret stimulates information search about decision alternatives and motivates choice switching (Igou et al., 2018; Shani & Zeelenberg, 2007).

However, this complex emotion can have seemingly paradoxical consequences: on the one hand, it can help guide thoughts and behaviors, and on the other hand, it can lead to decision avoidance because people exhibit regret aversion (Byrne, 2019). It is precisely this paradoxical aspect that needs to be investigated and resolved.

2.5 Mechanisms of recommendation norms

The idea of mimicking how humans generate their moral judgments in intelligent systems is subject to solid barriers due to the very nature of those judgments. This debate, far from being new, can be traced back to a dispute between the role of reason and emotion in moral judgments. Regarding that interaction: "What drives your judgment? Have you reasoned your way to the conclusion that something is morally wrong? Or have you reached a verdict because you feel indignation or outrage?" (May & Kumar, 2018).

Thus, while many researchers agree that reason and emotion play roles in moral judgments (Damasio, 1994), the question of which one plays the fundamental role remains open.

This thesis, with its pragmatic approach, does not try to give a single answer to this question but instead shows the different theories (Feldman Barrett, 2017; R. W. Picard, 1995; White & Katsuno, 2022; Woodley, 2020) that deal with emotions and why they are so relevant in the creation of AI systems (Fölster et al., 2014) to improve ethical decision making in a joint work of human and artificial agents.

Then, suppose we start from the assertion that behavior is considered ethical only because it obeys a social norm and that behavior is, in turn, automated, becoming routine. What happens when the objective context changes and certain norms should be adapted to these new circumstances? According to Johnson (2015), it is important to stress that an action done out of habit is not necessarily moral because "When we act in routine, habitually engrained ways, there is little or no actual deliberation or reflection involved" (M. Johnson, 2015).

In our case, our starting assumption is that emotions drive deliberative decision-making processes or, in other words, how human agents behave. But then, how can this new knowledge be integrated into how new social norms are updated or created? Would such a development help to generate new social norms that align with the specific needs of a society, culture, or group of members? Are the current mechanisms to create new norms sufficient to meet this challenge? According to the available literature (Boella & Torre, 2007; Finnemore & Sikkink, 1998; Küchle & Ríos, 2008; Tony et al., 2009) in the real world, norms are created from three methods which are: natural emergence from social interaction, the decree by a powerful agent, and agents negotiation within a group.

In the case of creating a norm in AI agents, although authors differ in some aspects, they are broadly based on three approaches. Two of the three approaches (Hollander & Wu, 2011) are grounded in offline design and norms of autonomous innovation. In offline design, the designer encodes norms directly in agents, which enact the norms. The designers update any new norms required by the system in the future. According to Hollander and Wu (2011), "This approach could be practically implemented in simple systems, but, in complex reasoning systems, offline design could fail in capturing the details required for realistic performance" (Hollander & Wu, 2011). In the same vein, Savarimuthu and Purvis (2007) point out that "offline design assumes that all agents adopt the norm in a society, which might not be realistic in open communities when different norms compete to present as the society's norm" (Savarimuthu & Purvis, 2007).

The other approach proposed by Hollander and Wu (2011) is called norms autonomous innovation, in which agents create new norms without any external interference. Current researchers on norms creation based on innovation have focused on machine learning and game theory (Tony et al., 2009). However, this approach contains the same problems of explainability as the rest of the systems that function as black boxes. The last approach that can also be an essential notion in establishing norms is social power. According to López, "an agent can express its social powers via its ability to change the belief, motivations, and goals of other agents" (López, 2003). In addition, Savarimuthu and Purvis (2007) add that the sources of power can either be leadership mechanisms (encourage and motivate followers to adopt a particular norm) or punishment mechanisms (enforce others to follow a particular norm) (Savarimuthu & Purvis, 2007). Nevertheless, although these concepts are relevant to this research work, norm enforcement is our fundamental concept. In communities of agents or real situations, norms regulate the agents' behavior.

However, what to do when the agents decide not to comply with the norms if this benefits them? According to Hollander and Wu (2011), social enforcement is often used to force an agent to adopt the behavior of other agents. Young (2008) explains that this enforcement can be performed externally, internally, or motivationally. Achieving social enforcement "requires a process that can detect the norms' activity and probable violations and handle this violation" (Vázquez-Salceda et al., 2004).

2.6 The frame problem: “What might have been” of COVID-19

But then, how do we detect that in specific contexts, the creation of new norms is necessary? Is it always obvious? In the literature on norms and normative multiagent systems, norm emergence is gaining more relevance as a new type of norm creation. According to Hollander and Wu (2011), "the norm is considered to emerge when it has been adopted by an adequate number of agents in a society" (Hollander & Wu, 2011).

Finnemore and Sikkink provide another definition of norm emergence as "persuasion by norm entrepreneurs which try to convince a critical mass of states (norm leaders) to embrace new norm" (Finnemore & Sikkink, 1998). Both definitions have subjective components when understanding the concept of "emergency" when creating new norms.

Therefore, looking at the global pandemic of COVID-19 from the point of view of the information available before it happened or on what kind of mechanisms lead decision-makers to make choices or behave is relevant to understanding the issues of bias and noise in depth. However, this dissertation doesn't focus on COVID-19 precisely but on specific open questions regarding the pandemic, which could serve as a good starting point for addressing the responsibility gap, a central term in this research. Therefore:

1. Why was the available information that a pandemic could not be sufficient to better prepare for it?
2. Why, being a global event, did each country act differently?
3. How can the decision-making or the mechanisms that led to specific decision-making be evaluated to establish responsibility?
4. Below is some information from before the pandemic that has already warned us of the seriousness of the issue. Why didn't anyone react years before?

2007, the article *Severe Acute Respiratory Syndrome Coronavirus as Agent of Emerging and Reemerging Infection* appeared. In the introduction, we find:

> "Severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV) is a novel virus that caused the first major pandemic of the new millennium. The rapid economic growth in southern China has increased demand for animal proteins, including those from exotic game food animals such as civets. Large numbers and varieties of these wild game mammals in overcrowded cages and the lack of biosecurity measures in wet markets allowed the jumping of this novel virus from animals to humans. Its capacity for human-to-human transmission, the lack of awareness of hospital infection control, and international air travel facilitated the rapid global dissemination of this agent. Over 8,000 people were affected, with a crude fatality rate of 10%. The acute and dramatic impact on affected countries' health care systems, economies, and societies within just a few months of early 2003 was unparalleled since the last plague. The small reemergence of SARS in late 2003 after the resumption of the wildlife market in southern China and the recent discovery of a very similar virus in horseshoe bats, bat SARS-CoV, suggested that SARS can return if conditions are fit for the introduction, mutation, amplification, and transmission of this dangerous virus" (Cheng et al., 2007).

Although this introduction might remind us of the beginning of the pandemic in Wuhan in December 2019- January 2020, we should pay attention to the question these researchers raised in their conclusions: "Should we be ready for the reemergence of SARS"?

In 2012, similarly, the Robert Koch Institute (RKI) developed the scenario of a global coronavirus outbreak. The document (Rki, 2012), published on the Internet, became famous through an article published on April 07, 2020, in the German magazine *Der Spiegel* titled "The Pandemic Simulation Game" (Das Pandemie-Planspiel). In the subtitle of the article, the following question is posed: “Should they have been better prepared for the current crisis? It's not that simple“ (Hätten sie besser auf die aktuelle Krise vorbereitet sein müssen? So einfach ist es nicht) (Merlot, 2020).

While these documents represent examples of warnings we needed to take seriously, the actual detection of the virus was made possible by novel technologies, including artificial intelligence (AI), machine learning, natural language processing, big data, the Internet of Things (IoT), and others.

According to Niiler (2020), "BluDot employed the services of AI-driven algorithms to analyze data gathered from sources such as news reports, air ticketing, and animal disease outbreaks to predict that the world was facing a new type of virus outbreak" (Niiler, 2020).

Besides those achievements, this startup and another called Metabiota independently and correctly predicted some regions where the virus would arrive in force. We have Japan, Taiwan, South Korea, Singapore, Thailand, and Hong Kong among the regions and countries with a correct prediction (Allam, 2020).

However, it is worth noting that predicting an event or having data available is not the same as knowing what to do with the information obtained. Thus, the fundamental question is whether the world could have been better prepared for the arrival of the pandemic or not. If the answer is no, the issue of responsibility becomes less critical, but if we could have been better prepared, the matter would have taken on another dimension. Therefore, an important point is whether we can refer to the COVID-19 pandemic as a "black swan."

This concept, developed by the Lebanese-American philosopher and researcher Nassim Nicholas Taleb, is a metaphor he developed in his 2007 book *The Black Swan: The Impact of the Highly Improbable.*

As explained in the book, a black swan is "a surprising event (for the observer) that has a great socioeconomic impact, and after it happens, it is rationalized retrospectively (making it seem predictable or explainable and giving the impression that it was highly probable that it occurs)." In addition, these types of incidents, considered extremely atypical, collectively play much more significant roles than regular events. On March 31, 2020, in remarks he made to Bloomberg about whether the coronavirus is "a black swan," Taleb answered: "A white swan event like the coronavirus pandemic was preventable" (Avishai, 2020).

Looking at the event of COVID-19 from the point of view of demanding responsibilities, one of the problems we find is the so-called "frame problem," also known as the "relevance problem." Artificial intelligence experts have addressed this issue in artificial agents, but it also occurs in human agent behavior and decision-making.

The "frame problem" was initially described by AI scientists (Mccarthy & Hayes, 1969) and is reformulated by more contemporary authors in this way:

> "If a programmed artificial agent is expected to keep track of persisting facts when reasoning about change, how can the programmer make the artificial agent do this without representing the numerous non-effects of the focused change?" (Shanahan, 2008).

According to Shanahan (2008), "Within Classical AI, various solutions to the frame problem have been developed, and it is not considered a serious obstacle." In the 1980s, the "Frame Problem" became a matter of interest to philosophers and cognitive scientists, but in the 1990s, interest dropped. "A resurgence of interest in these more robust forms of AI and more generally in the properties of autonomy, robustness, and flexibility that intelligent biological agents exhibit -warrants revisiting the Frame Problem.

Moreover, this problem interests cognitive science since it is widely believed that information processing is crucial for understanding mental processes" (Miracchi, 2020).

According to Miracchi, "the way that the Frame Problem is standardly interpreted, and so the strategies considered for attempting to solve it, must be updated." In addition, she adds:

> "We must take seriously the possibility that how intelligent agents use information is inherently different. Whereas intelligent agents are plausibly genuinely causally sensitive to semantic properties as such (to what they perceive, desire, believe, and intend.), computational systems can only be causally sensitive to the formal features that represent these properties. Indeed, this same substitution of formal generalizations for genuinely semantic ones is responsible for how AI systems are brittle, inflexible, and highly specialized. Formal causal relationships cannot reproduce the functional properties of genuinely intelligent systems, except in highly specialized and restricted circumstances". (Miracchi, 2020)

For this research, Glymour's warning is the most appropriate. He says: "The frame problem is not one problem, but an endless hodgepodge of problems concerned with how to characterize what is relevant in knowledge, action, and planning" (Glymour, 1987).

These points are essential not only for designing artificial intelligence systems but also for the behavior and decision-making of human agents. For this reason, this doctoral thesis focuses on searching and justifying what knowledge must be acquired and developed to plan our actions following the principle of responsibility.

The starting assumption is that we must first search for other mechanisms to understand new properties of reality that provide us with more knowledge about a specific behavior that we need to change or reinforce. This dissertation aims to theoretically understand the benefits of regret's affective and counterfactual components in achieving the principle of responsibility in a trustworthy human-AI interaction. In this roadmap, AI systems might be the principal support for human agents in this task, particularly the latest advances in the research domain of Explainable Goal-driven AI (XGDAIs).

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research paradigm and research question

The first step to consider before focusing specifically on the methodology used in this dissertation is to analyze the philosophical research framework that guides it. In other words, depending on the ideas one has about reality and the nature of knowledge, one paradigm or another or a mix between them is the most appropriate to achieve our goals. Thus, a paradigm represents how humans make sense of the world and their environment. The three main research paradigms in Philosophy are positivism, interpretivism or constructivism, and pragmatism. Due to the nature of this research, this dissertation falls into the latter group. Thus, the pragmatic paradigm aims to cover practical issues related to achieving specific goals, which can be theoretical or empirical. Thus, it is helpful to explain the other two paradigms mentioned above to corroborate the importance of the pragmatic paradigm since they are a mixture of them.

In short, from the positivist point of view, reality is considered objective and fixed, where things exist independently of us. Therefore, researchers can access it objectively and reproducibly via established tools and methods. On the other hand, the interpretative or constructive paradigm considers reality multiple because it is a social and cultural product created by individuals. According to Collis and Hussey, knowledge creation is generated through a deep understanding of phenomena and is shaped by our perceptions (Collis & Hussey, 2014).

Since the pragmatist paradigm is a mixture of both, let us see below how it is reflected in concrete terms for this research. The pragmatist paradigm states that reality is perceived as single and multiple simultaneously. This way of understanding and interpreting reality requires top-down and bottom-up approaches. Thus, although this research does not involve experimental research typical of the natural sciences and engineering, different theories and frameworks of these specific areas are objects to study to prove if they continue to respond to the needs of a globalized world where technological and social contexts are not always considered. An interpretive or constructivist approach to reality, typical of the social sciences and philosophy, is needed because of these realities or the diversity of contexts.

Because of this complex nature of reality or realities understood as local contexts or scenarios, the constructivist or interpretative paradigm of reality has a great weight in deciding why the pragmatic paradigm is the most suitable for this research work.

Thus, the knowledge that this interdisciplinary research in the area of philosophy and ethics of AI tries to provide is, on the one hand, objective as it tries to test what theories and frameworks are still relevant to contribute to an in-depth study of complex phenomena without losing objectivity and rigor. However, on the other hand, these specific contexts may lead to the need to rework and redefine some theories and frameworks because they do not solve current problems or issues we face.

The particular reason for that is that the problem or the question to be solved or understood arises from the visible dissonances established between methods or theories that were thought for other circumstances or contexts and today are not valid in the same way because we are facing new challenges.

These views of reality supported by theories and the meaning we attach to them reflect how we make decisions, the topic that is the core of this investigation. Now, how do humans make judgments? Do we always decide based on a system of rules? Do emotions play an essential role in making decisions, and how do we explain them? Could we create new ethical frameworks by integrating advances from different disciplines concerned with the research of human judgments or how humans make and justify/explain decisions? Could combining various techniques from various disciplines and cultures be a way to move towards more integrative approaches to understanding the contents of our decision-making? Given this complex scenario, my research question would be formulated as follows:

How could we design an interdisciplinary theoretical framework for responsible AI, where developing different levels of explanations in human-AI interaction validates whether the somatic marker hypothesis could serve as a recommendation norm for preventing unspecified errors before they occur?

3.2 Research approach

In line with the pragmatic research philosophy, this dissertation's research approach is qualitative and combines inductive and deductive techniques. Hence, it is imperative to highlight that deductive elements are very present since we work with theories that try to demonstrate patterns of reality or systematize aspects of it.

Therefore, this dissertation has an analytical character based on top-down and bottom-up techniques. In addition, this research adopts a critical attitude due to the sensitive subject matter. The qualitative-based analytical approach allows the creation of new meanings from the interactions and relationships of concepts and theories, letting us identify new contexts, patterns, and ways of acting. The result of such analysis and information processing is formulating a new theory.

This study explores the phenomenon of enhancing responsible decision-making in human-AI interaction from an interdisciplinary perspective, emphasizing what aspects should reinforce cooperation between human and artificial agents. To this specific goal, it is necessary to take the deductive approach to understand which theories or frameworks are being implemented for this purpose and where the limitations and barriers occur.

Thus, the pragmatist paradigm, the inherent techniques of qualitative research, and the inductive and deductive methods of analysis at the same time make this approach the most appropriate to achieve the answer to the research question posed in the previous point. Although this research could be approached in terms of quantitative analysis, the fact that it is not is because it is a preparatory prior stage to a practical approach. That is, collecting data in the way of a qualitative approach in a flexible manner, in turn, allows us to understand complex phenomena without simply reducing them to individual components.

This approach means that the data are not collected all at once, but it is a cumulative work where the premises and theses are revised in the different phases that correspond to the different chapters that make up this thesis. Thus, being able to introduce changes according to the new contextual information that is being added gives the research a dynamic and flexible character. In addition, aspects that were crucial at the beginning of the research may become less relevant in the later phases.

3.3 Research design

From the point of view of design, we must remember that research in philosophy is supported not only by the paradigm chosen but also by the ideas one has about the nature of reality and knowledge. To design this research, the dialectical method (Forster, 1993) tries to establish an internal coherence to this research's aspects or thematic blocks. Unlike purely scientific or engineering theses, the information that can be worked with is broader and not so focused, so establishing a clear structure to validate the hypotheses is the key to focusing a topic in the social and human sciences. Furthermore, one should add that the information provided in the previous Chapter on the state of the art has contributed to the decision to choose the dialectical method of the three phases, as described below.

It is convenient to highlight that this research does not aim to prove whether dialectics is a suitable method or has contradictions. The justification of this chosen design is that it helps in the formal and content aspects to cover the objectives and the hypotheses and the way of jumping from one to another according to the presented chapters. Thus, in the five chapters that make up the bulk of this research, both the structure and the content correspond to the dialectical method expounded by Hegel, which is characterized by a process that repeats itself endlessly in the search for truth and contains these three phases: (Forster, 1993): thesis, antithesis, and synthesis. Again, it should be noted that the concept with which we work on reality is dynamic, not static. In other words, the context in which this research is inserted is the hypothesis that all reality or realities understood as ways of thinking, specialized knowledge, history, evolution, technology, and political systems, to cite some examples (Taleb, 2007) arise from a need to organize ourselves (Bauman & Haugaard, 2008).

In other words, the possibilities of organizing ourselves in all these areas are infinite, but only some are established as realities that can be experienced. Thus, describing this dialectical method in broad strokes, we can say that the first phase corresponds to an initial thesis; however, as soon as another need or needs emerge, that can be technical or driven by psychological or sociological needs, in terms of Kuhn's book *The Structure of Scientific Revolutions*, a new paradigm is required (Van Dyck, 2018).

From this point on, we enter the second phase of this method, the antithesis. The antithesis arises from this emerging need to find a new paradigm that aligns more with the demands of the different contexts or realities. From the confrontation of the old thesis, which would be the original state that no longer answers what reality demands, arises the antithesis, which is almost a negation or an attempt to overcome the contradiction of the initial state.

That is to say, this second phase tries to argue why another type of paradigm is required. In other words, theories, laws, or ways of doing or behaving typical of the initial state known as "thesis" are called into question. The last stage of this triad of the dialectical method corresponds to what is known as synthesis. This final part tries to resolve the conflict between the thesis and the antithesis. This approach fits well with Kuhn's theory about the requirements for a new scientific paradigm. We should remember that in the case of Kuhn's view, these changes in science are motivated not only by the variables of scientific methods themselves but also by sociological and psychological issues, which are, after all, the topics we are working on in this research.

Another reason why this dialectical method has been chosen to conduct this research project is because it is part of the philosophical idealism influenced by the German philosopher Kant (Maybee, 2020). After all, the subject's consciousness creates the real world or what is the same: our narrations (Bauman & Haugaard, 2008; Harari et al., 2011; Taleb, 2007).

In other words, the human agent is active in determining what is possible, necessary, and desirable from the point of view of his particular context since we, as human agents, decide what variables are relevant when creating, in this case, one type of technology or another or using one theory, model, or framework instead of others.

It should be emphasized that this emergent approach is determined by the limits imposed on us by our physical nature. That is to say, the more we discover the laws that govern phenomena, the more information and knowledge we have for designing other types of realities more in accord with emerging needs.

The box below shows the dialectic process that permeates the three phases of the research project, and that would be structured by its contents as follows:

First stage: Thesis
One way to integrate AI systems into our society is to turn them into artificial moral agents, i.e., by explicitly implementing ethical frameworks or moral theories to constrain their behavior. In this way, they can learn to identify rules or norms and adapt to what society's demands expect of them. Given the increasing autonomy of AI systems to make decisions in high-risk situations and due to the progressive complexity of their internal structure referred to as black boxes, a sub-area in AI called Explainable AI has emerged. This new field aims to justify how these systems have arrived at their results or inputs to increase trust in those systems.
Second stage: Antithesis
The ethical frameworks or moral theories most commonly used in Machine Ethics are based almost entirely on the dominant Western approaches: deontology and consequentialism. This fact leads to an "oversimplification". Furthermore, anthropocentric moral approaches have failed to answer human dilemmas. In that case, it is necessary to review the shortcuts and pitfalls related to the issues of bias and noise related to how human beings make decisions. In the field of Explainable AI, the most fruitful field at the moment is the one that corresponds to data-driven XAI. This method aims to explain and discover how, from input, the same output was obtained under the same circumstances. However, an appropriate technical explanation differs from an explanation that helps the user. Moreover, how humans construct their explanations should not be how these systems explain their decision-making. The hypothesis is that the human mind also works as a black box in an attempt to explain the mental contents that lead them to make decisions or to make moral judgments.
Third stage: Synthesis
One way to solve the "oversimplification" in the field of Machine Ethics could be to implement or constrain the behavior of these systems based on multi-ethical frameworks from the East and the West. On the other hand, these frameworks should also consider the emotional component that drives humans in their decision-making. It has been scientifically proven that certain emotions impact our behavior and the strategies we use to achieve our moral goals. Regret has captured the interest of different scientific disciplines regarding responsible decision-making. The most relevant aspect is its anticipatory, counterfactual dimension and ability to feel responsible for a bad outcome. However, counterfactual scenarios by human agents are subject to the aversion to experience regret. Thus, the emerging field of goal-driven XAI could help achieve this goal. In this case, the IA system aims to explain to the user the importance of anticipating possible regrets in advance and giving specific recommendations to prevent unspecified errors before they occur.

TABLE 1 | Adapting the "dialectical method" as a research guide

3.4. Interdisciplinary data collection and mixed analysis methods

Due to the interdisciplinary nature of this research and the design presented before, a mixed method of data analysis based on deductive and inductive reasoning techniques, all within the qualitative research method, has been chosen. The image below illustrates how these two data analysis techniques combine. In the case of this dissertation, both methods are part of an emergent process that constantly feeds back.

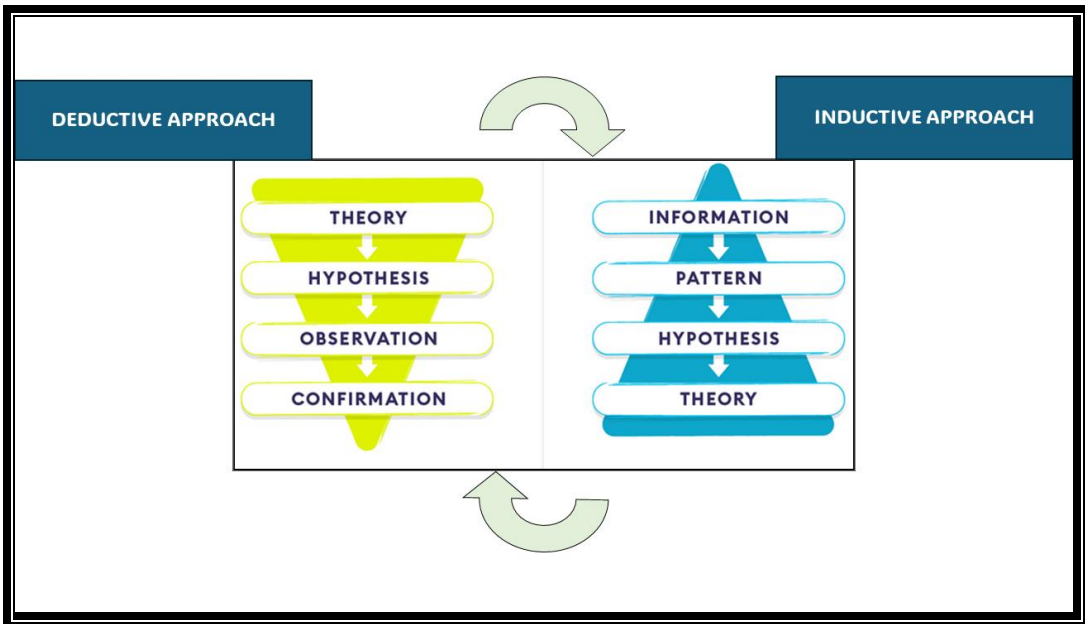


FIGURE 1 | Combination of inductive and deductive methods

The combination of these two methodological approaches is due to several issues that should be highlighted. As explained before, in the qualitative methodological approach, the researcher's background and experience might directly influence how reality is analyzed and how meaning is constructed from it. In this case, the desire to unite these two data collection and analysis techniques comes from the academic disciplines that have accompanied me and my theoretical knowledge about the two branches of AI that have the most strength in implementing this type of system, represented in the symbolic and subsymbolic approaches.

Thus, while in the philosophy of science, the deductive approach is used more because many of the theories with which we work have the status of premises and immovable truths, in scientific and investigative journalism, the collection of information and concrete cases is what then allows us to generate conclusions or a valid theory only from the data collected and analyzed. Undoubtedly, the best argument or the one that has served to design both the form and content of this thesis is Kuhn's hypothesis about what drives a paradigm shift (Van Dyck, 2018) in science, which is driven or motivated by sociological and psychological issues. Thus, the theories or ways of doing science in the different areas of knowledge have been sufficient so far.

However, due to the changes that, in this particular case, have introduced certain technologies in our current societies, a new redefinition and new theoretical and legal frameworks that respond more to the needs of emerging societies are required. Describing how the literature available for this dissertation is systematized and analyzed is essential. First, each thematic block is given by a design based on the dialectic triad (Maybee, 2020). Nevertheless, each of the phases that make up the dialectic method is structured according to the mixture of different queries or questions around how it has been changing the science of "ethical" decision-making until now, with the introduction of AI systems that support human decision-making.

Because this research is based on improving responsible decision-making strategies in joint work between humans and artificial agents, it is essential to understand how the processes of decision-making work, that is, what has begun to be done in this direction, what limitations and challenges remain open or have even been unleashed and even enhanced since the advent of AI systems or other advances in this kind of technologies. The following table analyzes and describes each part of this research in detail, considering which disciplines are included, the research questions, and analysis methods in each stage.

Questions	Disciplines involved	Analysis methods
First stage: Thesis		
Are the ethical frameworks and social norms governing human behavior sufficient and necessary to create an Artificial Morality Agent (AMA)?	Philosophy, Symbolic and Sub-Symbolic IA, Psychology and Neuroscience	Top-Down Bottom-Up
Can data-driven XAI explain the decision-making of an AI system considered as a black box?	Philosophy, Explainable AI, Psychology and Neuroscience	Top-Down Bottom-Up
Second stage: Antithesis		
Why are primarily only Western anthropocentric moral frameworks still used today to constrain the behavior of AI systems?	Philosophy, Symbolic and Sub-Symbolic IA, Explainable AI, Psychology and Neuroscience	Top-Down Bottom-Up
Why is the anthropocentric approach again present in how explanations are implemented in AI systems if the human mind also functions as a black box?	Philosophy, Symbolic and Sub-Symbolic IA, Explainable AI, Psychology and Neuroscience	Top-Down Bottom-Up
Third stage: Synthesis		
Could AI systems, with their explanatory component, serve as feedback by relying on the area of affective computing to understand what role emotions play in decision-making?	Philosophy, Symbolic and Sub-Symbolic IA, Explainable AI, Affective Computing, Psychology and Neuroscience	Top-Down Bottom-Up
How can we achieve a responsible framework where the explanatory component of an AI system can detect when its recommendations have not been followed?	Philosophy, Symbolic and Sub-Symbolic IA, Explainable AI, Affective Computing, Psychology and Neuroscience	Top-Down Bottom-Up

TABLE 2 | Description of disciplines involved, research questions, and analysis methods

3.5 Research plan

CHAPTERS	METHOD	ACTIVITIES	RESOURCES
CHAPTERS 4 & 5	**PHASE I** **THESIS** Identifying and contextualizing the problem	-Analysis and critical assessment of the techniques developed to explain the behavior of machines -Comparison and description of normative frameworks for constraining machine behavior -Limitations and challenges facing the area of data-driven Explainable AI and Machine Ethics	-The materials (literature) used come from English sources that have been compiled from different research databases such as Scope, Google Scholar, IEEE Xplore, Elsevier, and jstor -The material searched contained the following keywords or related categories in the title or abstract: "machine ethics," “normative Ethics Principles,” "roboethics," "artificial moral agents," "norm identification," "limitations of Machine Ethics," "explainable AI," XAI, "explainability," "black box model." -Courses, conferences, lectures, informal interviews, daily news readings, and analysis of decision-making situations have served as additional materials to complete this part.

CHAPTER 6	**PHASE II** **ANTITHESIS** Combining disciplines and categories to solve the problem	Description of the impact of the black box phenomenon in AI systems and the human mind Analysis and methods of measure of emotional data as an essential part of the decision-making process Introduction and assessment of the research domain of goal-driven Explainable AI	-The materials (literature) used come from English sources that have been compiled from different research databases such as Scope, Google Scholar, IEEE Xplore, Elsevier, and jstor -The material searched contained the following keywords or related categories in the title or abstract: "Explainable AI," "GDPR-Right to explanation," "data-driven XAI," "goal-driven XAI," "algorithmic and human bias," "noise" XAI, "trust," "Trustworthy," "requirements for a good explanation," "human judgment," "emotional data," "emotional theories," "somatic marker," "affective computing," "artificial emotional intelligence" -Courses, conferences, lectures, informal interviews, daily news readings, and analysis of decision-making situations have served as additional materials to complete this part.

CHAPTERS **7 & 8**	**PHASE III** **SYNTHESIS** Proposal of the different phases of the design of the framework	-Analyze the role of imagination and counterfactual reasoning in norms creation mechanisms -Identify and allocate the relationship between the ability to experience regret and the principle of responsibility in decision-making -Reinforce the prospective and counterfactual element of regret emotion as an explanatory component in the research domain of goal-driven explainable AI	-The materials (literature) used come from English sources that have been compiled from different research databases such as Scope, Google Scholar, IEEE Xplore, Elsevier, and jstor -The material searched contained the following keywords or related categories in the title or abstract: "imagination," "counterfactual reasoning in XAI," "causal inferencing," “social enforcement,” “norms creation mechanisms,” "prospective regret, "reinforcement learning," "latent learning," "regret aversion," "bias and noise." -Courses, conferences, lectures, informal interviews, daily news readings, and analysis of decision-making situations have served as additional materials to complete this part.

TABLE 3 | Description of the research plan

3.6. Limitations

Theoretical and interdisciplinary limitations
The design of an interdisciplinary methodology raises significant issues. Above all, the limitations are centered on the methods, sometimes even contrary, of collecting and analyzing the data, theories, or concepts with which one works. In the natural sciences and engineering, the use of materials and how these disciplines conduct their experiments is characterized by its universality. In the case of this research, finding a methodology that fits with each of the sub-disciplines present makes one doubt the scope of the chosen method itself, even if it is adequately justified. This research's interdisciplinary nature is a paradox, making it a limitation. On the one hand, to solve complex problems, one needs to understand what makes them complex, not by reducing them to individual parts but by trying to understand the problem, albeit in each specific context. This way of integrating global and local aspects is currently a limitation because it requires, most of the time, a combination of methodologies that may be antagonistic at first glance. Another limiting aspect of interdisciplinarity is using the same concepts but with different meanings according to the discipline in which they are included. The concepts of necessary and sufficient in philosophy differ from their uses in mathematics and logic. The concept of model and even framework in philosophy does not have the same function as in computer science or psychology. All these conceptual challenges make it difficult to accept specific hypotheses since understanding phenomena is done from a different angle. This limitation is due to a lack of a common language, a communication problem between disciplines that needs to be solved urgently. The most limiting aspect of this research is that it has not been tested in practice. Implementing a project that incorporates these characteristics into reality requires the collaboration of an interdisciplinary team formed by the research domains mentioned above. Only to form an adequate team to solve the problem in question does one first identify with the objective; one must understand in depth what the problem implies. From the beginning until almost the end of this dissertation, I doubted I could express my thoughts. In philosophy, one works with concepts, theories, and ways of seeing and understanding reality and creating new meanings, but this does not correspond to the rapid creation of a computer program or with something material that can be put into operation quickly.

TABLE 4 | Limitations identified before starting the research project

CHAPTER 4

WHO IS RESPONSIBLE IN RESPONSIBLE AI?

4.1 Facing the (automated) decision-making gap in terms of responsibility

4.1.1 The dilemma: technology as a means or as an end

Innovations in areas such as autonomous machinery, machine learning, and social robots are generating debates regarding who is responsible for the actions and decision-making of these technologies.

Thus, assigning responsibility is something intrinsically linked with morality. And, morality, as Hall (2001) points out, "rests on human shoulders, and if machines changed the ease with which things were done, they did not change responsibility for doing them. People have always been the only moral agents" (Gunkel, 2018).

This way of understanding morality and technology has to do with instrumental purpose. The "instrumental theory of technology is the most widely accepted view, and it is based on the common-sense idea that technologies are "tools" standing ready to serve the purposes of users" (Feenberg, 1991).

In the same line, Deborah Johnson adds that "no matter how independently, automatic, and inter-active computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision" (D. G. Johnson, 2006).

In addition, the view of machines as artifacts under strictly human control is the moral issue of their use. In other words, "holding a robotic mechanism or system culpable would be illogical and irresponsible" (Gunkel, 2020).

However, we must also consider criticisms of instrumental theory that reduce technology to the ontological status of mere tools or instruments. Marx reflected on the status of machines in his book *Capital*. For Marx, there is a tendency to confuse the concepts of tools as simple machines and machines as complex tools.

He states, "The machine, therefore, is a mechanism that, after being set in motion, performs with its tools the same operations as the worker formerly did with similar tools" (Marx, 1967).

However, a vital modification must be considered when discussing "autonomous technology" as "such mechanisms are not mere tools to be used by human beings but occupy, in one way or another, the place of the human agent" (Gunkel, 2020).

4.1.2 Real examples: when ontologically, the machine is an end in itself

A contemporary example illustrating the problem described by Marx is autonomous vehicles, also called self-driving cars. According to Gunkel, "the autonomous vehicle, whether the Google Car or one of its competitors, is not designed for and intended to replace the automobile. It is, in its design, function, and materials, the same kind of instrument that we currently use for personal transportation.

Therefore, the autonomous vehicle does not replace the instrument of transportation (the car); it is intended to replace (or at least significantly displace) the driver" (Gunkel, 2020).

We must emphasize this new technological landscape, where control over actions and decisions does not involve a human being.

The progressive automation of many of the tasks previously performed by humans does not imply that the instrumental theory becomes useless, but instead that it must be recognized that the autonomous technology of which Winner (1997) speaks occupies another ontological position (Van De Poel, 2020).

In this way, a significant gap arises in the concept of the "responsibility gap" developed by Matthias, which refers to the change in the role of a programmer from someone doing traditional software engineering to a builder of autonomous artificial intelligence systems. This author concludes that:

1. In the course of the progression of programming techniques, from the conventional procedural program via neural network simulations to generically evolved software, the programmer loses more and more control over the finished product. She increasingly becomes a 'creator' of 'software organisms,' the exact coding of which she does not know and cannot check for errors.

2. The machine's behaviour is no longer defined solely by some initial and henceforth fixed program. However, it is increasingly shaped by its interaction with the operating environment, from which the machine adapts new behavioural patterns that constitute solutions in the machine's problem space.

3. If the machine needs to adapt flexibly to new situations (which is necessary if it operates in dynamic and changing environments), automata must leave behind a clear separation between the programming, training, and operation phases. Practically useful technologies will have to be learned during operation, which also means that they will have to make 'mistakes' during operation (a "mistake" being just the exploration of the solution space by the machine itself, which enables it to arrive autonomously at new solutions).

4. There are an increasing number of situations in which the supervision of an operating machine by a human expert is either in principle or, for economic reasons, impossible. The supervision of an operating machine is impossible when the machine has an informational advantage over the operator. It is also impossible when a human cannot control the machine in real-time due to its processing speed and many operational variables involved.

5. Still, we cannot do without such systems because the pattern processing and systems control tasks we must accomplish in our highly dynamic and complex environments are so complicated that more uncomplicated, statically programmed machines cannot address them.

(Matthias, 2004):

Several years after the term “responsibility gap,” we have witnessed at least two good examples of the effects that these technologies could have that have made headlines worldwide.

4.1.3 Real examples of the "responsibility gap" in machine learning

The technological cases we consider involve machines that have avoided the control of their programmers or the actions they have executed because the programmers have been unable to understand their decision-making or the process that led them to decide that way.

The first example is AlphaGo. As the company Google DeepMind (2016) points out, “this technology combines Montecarlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play" (Moyer, n.d.)

One of AlphaGo's creators, Thore Graepel, explains the relevant aspect worth highlighting: "Although we have programmed this machine to play, we have no idea what move it will come up with. Its moves are an emergent phenomenon from the training. We create the data sets and the training algorithms. But the moves it then comes up with are out of our hands" (Metz, 2016).

The interesting question that would fill headlines in the media would be: who beat Lee Sedol? If this answer were known, accountability for the actions that led to the victory would be simple.

FIGURE 2 | AlphaGo defeated Lee Sedol / Lee Jin-Man / AP

The second example that highlights this "responsibility gap" (Matthias, 2004) and is more related to the ethical dimension of this research is Microsoft's "AI chatbot Tay.” As Gunkel (2020)

notes, if the instrumental theory of technology is strictly applied, Tai's programmers would be responsible for creating a racist machine. But again, the answer is not that simple because we do not know the type of data intended to be fed into the system or the type of training.

Tai was primarily a machine learning project, but it had more of a cultural and social experiment component than a technical one. According to some headlines and news, "Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways" (Fingas, 2016).

As a result, we have taken Tay offline and are making adjustments". In terms of responsibility, as Gunkel suggests, "Microsoft is only responsible for not anticipating the bad outcome; it does not take responsibility or answer for the offensive tweets" (Gunkel, 2020).

FIGURE 3 | Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation

Of course, Tay generated hate speech from the data of the user who interacted with it.

The provocative side of this experiment is not only to what extent machines are a means to harm third parties but also to understand whether humans' behavior is an excellent example of ethics for machines and other humans, determined by their decisions or actions.

To deal with the responsibility gap, Gunkel (2020) offers in his article *Mind the Gap: Responsible Robotics and the Problem of Responsibility*, these three possible methods:

1). *Instrumentalism 2.0,* in which these recent innovations in artificial intelligence and robotics work as instruments or tools. Therefore, operating robots is entirely our responsibility.

We determine their goals and behavior, either directly or indirectly, by specifying their intelligence or, even more directly, by specifying how they acquire it.

2). Developing *machine ethics is* a field that seeks to implement moral decision-making faculties in computers and robots.

3). *Hybrid responsibility,* as an attempt to distribute responsibility across a network of interacting human and machine components.

Of these three options, instrumentalism 2.0 is the subject of Chapter 4, as it is how we have begun to regulate the decision-making of AI systems.

On the other hand, machine ethics could be another proper avenue for regulating these increasingly independent and autonomous systems, but this requires further experimentation, which takes years of study. Point three may soon become a form of regulation.

However, terms and tasks still need to be defined, i.e., separating what roles within decision-making are still the human agent's domain and what kind of role the artificial agent has in that task execution and the final output.

4.2 Towards doing ethics in machines computable

4.2.1 Machine Ethics as an emerging interdisciplinary field

This section is of great relevance. The reason is not only the success with which different ethical theories can create a moral machine but the effect this ethical achievement could have on how humans understand moral and ethical issues.

The outstanding achievement is that this subfield of artificial intelligence has brought together the disciplines of engineering and philosophy. Specialists from both fields are jointly trying to address today's ethical challenges.

Before beginning an in-depth analysis of developments in machine ethics, it is essential to note that this research does not focus on clarifying its terminology.

That is, the way to deal with the concepts is similar to that used in the paper *Motivations and Risks of Machine Ethics* (Cave et al., 2019).

For example, these authors use the term machine in the broadest sense to include (among others) both ordinary physical machines, autonomous robots, and purely algorithmic systems.

The other terms frequently used interchangeably are "ethical" and "moral." Also, an "ethical machine" and an "artificial moral agent" (AMA) have the same semantic charge. In the case of needing a more limited context, the distinction takes place explicitly.

The rise of automation has brought the need to rethink decision-making in human-machine interaction. More and more intelligent systems are making decisions for and on behalf of humans.

This progress has led some disciplines, such as cognitive science, moral philosophy, intelligent system developers, neuroscientists, and cognitive psychologists, to consider extending particular human morality to machines.

Ethics remains a diffuse discipline because it attempts to respond to several dilemmas in a way that is not always satisfactory for the parties involved. Indeed, reducing ethics to a set of universal rules seems counterproductive.

On the positive side, this field of research is a good starting point to reflect not only on whether we can create an artificial morality but also to understand more about the pillars on which human morality or moralities are based and work together from that vantage point.

One way of justifying the possibility of creating ethical machines is the following statement: "The engineering task of building autonomous systems that safeguard fundamental human values will force scientists to break down moral decision-making into its parts, recognize what kind of decisions can and cannot be codified and managed by essentially mechanical systems, and learn how to design cognitive and affective systems capable of managing ambiguity and conflicting perspectives" (Allen et al., 2006).

More concisely and in direct relation to the impact that the automation of certain decisions would have, which could even put at risk aspects of the lives of those involved, Rosalind Picard, director of the Affective Computing Group at MIT, said, "the greater the freedom of a machine, the more it will need moral standards" (R. Picard, 1997).

However, this sentence could be understood as referring to the freedom of the machine and the freedom that humans possess to make decisions that affect the lives of others, including behavior toward other non-human agents.

Some authors who can be considered pioneers in the field of machine ethics suggest the following: "In fact, computers may be better than humans in making moral decisions in so far as they may not be as limited by the bounded rationality that characterizes human decisions, and they need not be vulnerable to the emotional hijacking[3]" (Allen et al., 2006). Thus, the ultimate goal of Machine Ethics as a discipline is "to create a machine that follows an ideal ethical principle or a set of ethical principles in guiding its behavior" (Anderson & Anderson, 2011).

In the field of Machine Ethics, one of the terms is Artificial Moral Agent or AMA, which refers to "a machine with ethics as part of its programming" (Tolmeijer et al., 2020).

[3] It should be emphasized that understanding emotions and the role they play in the decision-making process is a topic for the next chapter. The traditional view that moral judgements are strictly rational is a critical point in this thesis.

However, this general concept has acquired different meanings depending on what is meant by a moral agent. One of the most worked definitions has been the classification offered by Moor, which distinguishes the following four levels of moral agents:

> *Ethical-impact agents* are those that have an (indirect) ethical impact. An example would be a simple assembly line robot that replaces a human in a task. The robot itself does not do anything (un)ethical by acting. However, by existing and performing, it has an ethical impact on its environment; in this case, the human that performed the task is replaced and has to find another job.
>
> *Implicit ethical agents* do not have any ethics explicitly added to their software. They are implicitly ethical because their design involves safety or critical reliability concerns. For example, aeroplane autopilots should allow passengers to arrive safely and on time.
>
> *Explicit ethical agents* draw on ethical knowledge or reasoning in their decision-making process. They are explicitly ethical since normative premises can be found directly in their programming or reasoning process.
>
> *Fully ethical agents* can make explicit judgments and can justify these judgments. Humans are the only agents considered full ethical agents, partially because they have consciousness, free will, and intentionality.
>
> (Moor, 2006)

This classification does not delve into the exact definition of a moral agent. Later work, such as Himma's, suggests that a moral agent must possess consciousness and intentionality. Both requirements are lacking in current systems (Himma, 2009). In addition, these two concepts lead to considering different theories beyond the scope and objective of this research.

On the other hand, this analysis of the ethical theories by the scientific community provides a clue as to how humans understand morality. It thus allows us to realize the potential role machines can play in supporting and helping us understand how we express certain ethical or moral principles through our behavior and decision-making.

4.2.2 Between Kant and Bentham: the dominant Western ethical values

One way of getting autonomous machines to do "the right thing" is to endow them with normative ethical theories borrowed from philosophy. Currently, most normative ethical theories used for AMAs or to design moral or ethical machines are deontology, consequentialism, and virtue ethics.

This division corresponds to one of the most recent studies about what specific approaches drive the work on AMAs and ethical theory, written by Zoshak and Dew in their paper entitled *Beyond Kant and Bentham: How Ethical Theories are Being Used in Artificial Moral Agents.* For their paper, they performed a literature search and qualitative review of papers published about machine ethics in the ACM Digital Library and IEEE Xplore.

The most relevant part of their research is the conclusion that most of the research work was limited to "making analytical arguments and contributing technical implementation knowledge, with the minimal empirical study of AMA development and deployment in a natural setting" (Zoshak & Dew, 2021).

Therefore, it is appropriate to briefly describe the normative ethical theories mentioned above because they are the preferred ethical frameworks for endowing machines with artificial morality.

Deontology: this is an ethical framework in which acting according to specific rules is considered the right thing to do. It is closely associated with Immanuel Kant (1785) and his categorical imperative. The categorical imperative can be formulated as follows: "Act only following that maxim through which you can at the same time will that it become a universal law" (Manna & Nath, 2021)

Consequentialism: this is an ethical theory alternative to deontology because it focuses on the outcomes of behavior. In the classical view of the utilitarianism of Bentham (1789) and Mill (1861), which is the most prevalent branch of consequentialism, an act[4] "is morally right if and only if that act maximizes the good, that is, if and only if the total amount of good for all minus the total amount of bad for all is greater than this net amount for any incompatible act available to the agent on that occasion" (Roff, 2020).

Virtue ethics: this approach differs from the theories presented because "it does not focus principally on the consequences or rule-consistency of actions but on agents and more particularly on whether they exhibit good moral character or virtuous dispositions. A good action is consistent with a virtuous person's moral dispositions" (Tolmeijer et al., 2020).

Although these three approaches are the most studied in Machine Ethics, other literature can bring other perspectives. For example, Tolmeijer et al. refer to particularism in this way:

> "This approach focuses on the fact that there is no unique source of normative value, nor is there a single, universally applicable procedure for moral assessment. Rules or precedents can guide our evaluative practices. However, they are deemed too crude to do justice to many individual situations. Thus, according to particularism, whether a particular feature is morally relevant in a new situation -and if so, what exact role it plays- will be sensitive to other features of the situation" Tolmeijer et al. (2020).

[4] https://plato.stanford.edu/entries/consequentialism/

Thus, to broaden the range of possible normative ethical theories, the following section presents Confucianism's contributions to this field of Machine Ethics.

4.2.3 Confucian robot ethics: looking for "how to become good"

As mentioned above, most literature about machine ethics or artificial moral agents (AMAs) is devoted to the deontological and consequentialist frameworks. These theories, however, often struggle to "accommodate the constant flux, contextual variety, and increasingly opaque horizon of emerging technologies and their application" (Vallor, 2018).

Because of such limitations, the role-based Confucian approach brings a new perspective. In contrast to the Western philosophical methodologies of Machine Ethics that focus on defining "what the good is," Confucianism worries about "how one can come to know the good," Chinese philosophers represented by Confucian scholars are more interested in the problem of "how to become good," (Ivanhoe, 2000).

Also, Confucian ethics is not only focused on creating "a reliable and efficient human interaction but also a robot-mediated environment in which human teammates can grow their virtues" (Zhu et al., 2019).

A possible disadvantage that machine ethics could have if it only focuses on Western values is "if the philosophy of AI and robotics only comes from the West, that won't be enough because it won't always apply to non-Western countries" (Wong, 2019), "And you miss opportunities to think differently about technology" (Zhu et al., 2019). The same authors, Zhu et al. (2019), add, "Confucian ethics, along with other non-Western ethical resources, can enrich the moral imagination of roboticists and enhance their capabilities to define and engage ethical issues in designing robotics from culturally diverse perspectives."

Scholars conceive Confucian ethics as a role-based moral theory (Ames, 2011). This Confucian role ethics focuses on the premise that humans were born into a web of social relationships, and their interactions "have normative implications and they prescribe specific moral responsibilities for us in the communities we belong to" (Zhu et al., 2019).

According to Ames (2011), the term person is “relational and social[5].” Therefore, for Zhu and co (2019), the Confucian ultimate goal of becoming a good person depends on how well we live and practice the moral responsibilities prescribed by these social roles.

The most interesting point to remember is that to achieve this goal of becoming a good person, "Western ethical approaches focus on moral reasoning and justification while Confucian ethics emphasizes ethical practice and practical wisdom. The moral development model is central to Confucian ethics, consisting of three interrelated components: observation, reflection, and practice” (Zhu, 2018).

First, we consider these components by observing how people interact with others and reflecting on whether and how their daily interactions follow li 礼 (rituals[6] or ritual propriety). Zhu (2018) adds: "The appropriate practice of rituals manifests virtues, whereas virtues underlie and guide the practice.

Then, one needs to incorporate her reflective learning experience into future interactions with others and test to what extent she has grasped the appropriate practice of rituals and their underlying moral virtues" (Zhu, 2018).

Another important aspect of Confucian ethics is how personhood is understood. Whereas the Western value system resides in terms of innate and inalienable human rights, Confucians think instead that one's intentional efforts to strive to be a good person define one's personhood.

Applying this Confucian wisdom to the specific field of robotic or machine ethics leads to conclusions such as those expressed by the philosopher of technology Peter-Paul Verbeek (2006): "Design engineers do have an obligation to imagine the potential relationship that will be constructed between technology and its users and how such relationship affects the moral perception and behavior of the user (Verbeek, 2006).

Borrowing from the idea of Asimov's three rules of robotics, as known in the West, the philosopher JeeLoo Liu constructed three principles for Confucian robotic ethics on the moral role of robots:

[5] This vision of the human being as a "social being" is also present in Aristotelian ethics, although with other peculiarities.

[6] This way of understanding rituals would be equivalent to the concept of habit in the West.

[CR1] A robot must first and foremost fulfil its assigned role.

[CR2] A robot should not act in ways that would afflict the highest displeasure or the lowest preference of human beings when other options are available.

[CR3] A robot must help other human beings in their pursuit of moral improvement unless doing so would violate [CR1] and [CR3]. A robot must also refuse assistance to other human beings when their projects would bring out their evil qualities or produce immorality.

(Liu, 2017)

This type of technology design requires combining functional requirements with moral values. Thus, "designers not only can provide us with technical means but also address the values of people and society and think about expressing them in material culture and technology" (van den Hoven et al., 2015). One last aspect to consider is the different conceptions of autonomous moral agency. Western philosophers like Dumouchel and Damiano state that "only robots with no explicit purpose may have autonomous moral agency comparable to human personhood. Not having an explicit and predetermined purpose indicates that these robots can do anything they want" (Dumouchel et al., 2017).

However, Confucian ethics focuses on social robots, often designed to serve human purposes in specific situations. Those purposes are not conceived in Asian philosophy relative to the individualistic and liberal character of moral agency or the concept of personhood but emphasize the importance of social roles and their relationship to others.

4.2.4 Encoding ethical frameworks in machines: three different approaches

Developing AMAs, ethical machines, or ethical robots aligned with human values requires looking closely at how the normative ethical theories described in the previous section can be encoded. Allen, Smit, and Wallach (2005) are three authors who have significantly contributed to this work area.

For them, there are three ways of creating AMAs: top-down, bottom-up, and hybrid. A top-down approach "consists of taking a specified ethical theory and instantiating it to identifying particular states and actions as ethical or unethical concerning that theory" (Allen et al., 2005).

This view is the best option for ethical reasoning in limited domains, where decision-making is subject to following pre-established rules.

In such cases, the ethical theory can be implemented before deploying the system and does not change throughout its usage. "This allows for a thorough and well-informed decision-making process and the verification of the system before its practical application" (Charisi et al., 2017). The philosophical traditions more relevant to this approach are mainly consequentialism and deontology.

However, this approach is not without limitations. "General ethical principles are typically formulated on a very abstract level. Much philosophical discourse on ethics concerns problems when applying such general principles to concrete situations" (Charisi et al., 2017).

In contrast, the bottom-up approaches "provide an AMA with environments and instances to learn preferable behaviors and infer core values to arrive at moral decisions and actions" (Allen et al., 2005). This approach does not predetermine moral principles, ethical theories, or rules; instead, this point is questionable because every system that learns by itself must discriminate particular features of the data set and ignore others.

Therefore, it is difficult to say whether it is possible to create ethical-theory-agnostic AI systems since every choice of relevant data features is already, to some extent, a choice of moral theory.

If neither of these approaches meets the requirements "for designating an artificial entity as a moral agent, then some hybrid will be necessary" (Allen et al., 2005). The challenge consists of appropriately combining features of top-down and bottom-up approaches (Charisi et al., 2017).

However, other critical scholars are against the idea that machines can or should have some moral decision-making capabilities due to their interpretive flexibility (Rudin, 2019).

Other authors argue instead that we should focus on empirical studies of practices to uncover value tensions and levers to understand better how ethics gets worked out in situ (Rhim et al., 2021; Shilton et al., 2014)

As can be seen, the disagreements between the different experts involved in this challenge of creating artificial moral agents are evident due to the limitations and challenges of both approaches.

These constraints are mainly because ethical normative principles adopt different semantic charges depending on the context in which they are inserted. This point calls for reforming and adapting these normative ethical principles more integrative. This approach means, for example, combining different normative ethical principles and integrating them within a multi-level approach to ethics. Therefore, this dissertation proposes the combination of deontology with a norm recommendation based on the principle of responsibility with the Confucian maxim of "how to become good" over time. However, what disciplines, techniques, and knowledge are necessary to achieve at the forefront of designing such an interdisciplinary framework?

4.3 Moral machines and emerging ethical dilemmas

4.3.1 The emotional component of human judgments

The central argument advanced by machine ethicists about the need to have ethical or moral machines is that their reasoning ability and ethical decisions could be even better than those of humans. Along this line of thought, Dietrich points out that: "humans are genetically hardwired to be immoral…let us -the humans- exit the stage, leaving behind a planet populated with machines which, although not perfect angels, will nevertheless be a vast improvement over us" (Dietrich, 2007).

This line of argument assumes that an artificial moral agent could be better at moral decision-making than a human, given that it would be impartial, unemotional, consistent, and rational every time it made a decision. This approach fits with the rationalist view of moral judgments. In addition, this way of arguing, which is very much in line with other significant research works (D. Kahneman, 2011; D. Kahneman et al., 2021), clashes with the latest advances in the field of affective neuroscience and with other more recent research in the field of cognitive development of morality (Gold et al., 2014). Consequently, psychological and neurological research advances have demonstrated that emotions play an important role in moral judgment.

Significant advances are, for example, the somatic marker hypothesis postulated by Damasio (1994). In his neuroscientific studies, Damasio has stressed the importance of secondary emotions in morality. In parallel, Greene and others have conducted fMRI investigations of emotional engagement during moral judgment.

Gold et al. conducted research on their behalf to find cultural differences between Chinese and British participants' responses to real-life and hypothetical trolley problems.

These researchers' study was based on testing whether moral intuitions should be generalized and whether psychological morality is universal. These researchers point out that trolley problems have been used in developing moral theory and the psychological study of moral judgments of behavior. However, most past research on this topic has focused on people from the West. This approach may be why Gold et al. found apparent cultural differences in the approach to hypothetical trolley problems associated with differences in moral judgments and behaviors (Gold et al., 2014).

Because of this new knowledge about human judgments and emotions' role in deliberative processes, it is necessary to understand what they imply for human behavior and decision-making. These relevant aspects call for generating or adapting existing ethical frameworks, principles, norms, or rules or creating new ones based on this new scientific evidence.

4.3.2 Appearing ethical or being ethical

The previously commented aspect about the little information we have at the scientific level with rigor and systematization about the importance that emotions play in the moral judgments issued by humans leads to updated, philosophical questions such as: what does it mean to be moral? Or rather, what are the essential requirements for a moral agent? While ethics seems to be an intrinsically human domain because we must make moral decisions daily, proving that we can introspect how we arrived at our behavior or decisions is unclear. There is a clear gap in our knowledge about how we process information and what content drives how we behave or make decisions.

In fact, and returning to one of the salient issues of this research, humans tend to interpret emotions to the point that "we do not demand proof that another person has mental states or that they are conscious; instead, we interpret the other's appearance and behavior as an emotion" (Coeckelbergh, 2010). However, could AI systems help us move from the appearance of being moral to understanding even more aspects of our morality?

4.3.3 When a machine explains its ethical behavior

So, while on the one hand, experts in machine ethics are looking for solutions to limit or constrain the behavior of increasingly autonomous AI systems, on the other hand, it is becoming increasingly common for these AI systems to make decisions in place of humans, even in high-risk situations. As discussed in the following sections, this situation has led to regulatory changes in many countries, which call for greater algorithmic transparency (Bundy, 2017; Kaminski et al., 2019; Wachter, Mittelstadt, & Floridi, 2017).

For that purpose, Article 22 of the GDPR, for example, in Europe, has introduced "The Right to Explanation," also known as algorithmic accountability.

Understanding why these systems behave as they do has led the field of explainability to re-emerge as a solution to this challenge. In addition, this issue calls for the need to increase transparency in algorithmic decision-making, and in the first instance, it is assumed that this right to explanation increases the user's confidence in the AI system. However, what kind of explanations are black-box systems capable of providing? What kind of content should they provide? How can the quality of an explanation be evaluated?

As can be seen in these first introductory questions of the following thematic block, which is explainable artificial intelligence, the challenges of this emerging field leave the door open to introduce or further specify the explanations or orient them to a specific domain to mitigate their possible misuses. In this way, it is thoughtful to introduce this field by giving a historical overview of its current applications until reaching the field of explainable goal-driven AI, which is the thematic block that closes this first research phase.

CHAPTER 5

EXPLAINING EXPLAINABLE AI

5.1 A “right to explanation”: a controversial issue

While AI experts and philosophers continue to look for ways to create artificial morality in Machine Ethics, the legal system has taken a step forward. The necessity of new legislation arises because automated decision-making has started to influence private spheres, and it is a significant issue to know the reason that drives the AI systems to arrive at their outputs.

Facing the challenges, one contribution to this new technological paradigm is the legislation that since May 25, 2018, the General Data Protection Regulation has been replacing the EU’s 1995 Data Protection Directive (DPD) (Birnhack, 2008). The GDPR (European information commisisioners office, 2021) appeared as a need to regulate algorithm decisions.

FIGURE 4 | General Data Protection Regulation GDPR background concept. Vecteezy.com

This regulation calls for citizens’ rights to receive an explanation for algorithmic decisions that impact their lives, pursuing what is known as Algorithmic Accountability.

However, this legislation has weaknesses in fully automated decision-making: “Algorithmic decision-making can be opaque, complex, and subject to error, bias, discrimination, in addition to implicating dignitary concerns” (Kaminski et al., 2019).

The GDPR, since its inception, has caused several experts to doubt that this legislation, as described in Articles 22, as well as 13, 14, and 15, can pursue the objectives of Algorithmic Accountability that our societies need at both the individual and collective level (Kaminski et al., 2019).

Given that this dissertation is dealing with the principle of responsibility in a trustworthy human-AI interaction, it is relevant to turn to the articles that in this regulation have a direct bearing on “algorithmic accountability,” appealing to “meaningful information” and “right to explanation.”

The first part of this section focuses in detail on the description of particular articles of the GDPR -cited above- that have to do with Algorithmic Decision-Making and what this implies legally—followed by an overview of the field of Explainable AI that emphasizes its historical development, different methods and currently achievements technically speaking. Immediately afterward, the frameworks proposed by groups of experts in explainable AI are shown.

This second part focuses mainly on which aspects of human and non-human communication could serve as a basis for establishing such natural language explanations.

However, due to the intrinsic limitations that natural explanations have or what is also known as data-driven Explainable AI, a new emerging subfield called goal-driven XAI is presented as one of the fundamental research domains proposed in the final framework.

5.1.1 Automated decision systems decide for us: is there an explanation?

One of the controversies caused by the articles referring to “Algorithmic Accountability” in the GDPR legislation (Selbst & Powles, 2017) is that it does not discuss “a right to explanation.” However, articles 13-15 provide rights to “meaningful information about the logic involved” in automated decisions.

Therefore, while several scholars consider that this "right to explanation" should be interpreted broadly, functionally, and flexibly (Goodman & Flaxman, 2017), researchers, lawyers, and other experts argue that this "right to explanation" is formulated in this legislation in a vague and unclear manner. (Wachter et al., 2016). But then, what do the articles of this legislation say about "Algorithmic Accountability"?

5.1.2 The four Articles of the GDPR that address Algorithmic Decision-Making

Four Articles of the GDPR refer in detail to Algorithmic Decision-Making. Article 22 introduces the right "not to be subject to a decision based solely on automated processing" (Vollmer, 2023).

While this regulation was already present in Directive 95/46, the GDPR's predecessor, the current situation has led to an increase over the last 20 years in the use of profiling for practices that include advertising services, -among other uses- that may call into question the legality of such practices. Nevertheless, one should emphasize that Article 22 of the GDPR "does not explicitly provide for an individual's right to an explanation of an automated decision" (Tosoni, 2021). Consequently, this article allows companies to use algorithms for important decisions, "adjusting their behavior only if individuals invoke their rights" (Kaminski et al., 2019).

Then, what right/prohibition applies in Article 22? This article applies only when the decision is "based solely" on Algorithmic Decision-Making and produces "legal effects" or "similarly significant" effects on the individual (Tosoni, 2021).

On the other hand, articles 13, 14, and 15 of the EU legislatures contain "a series of individual notification and access rights" (Kaminski et al., 2019). In summary, Article 13 sets out rights and requirements when the systems collect information from individuals. Similarly, Article 14 refers to the rights and requirements of the previous article with the difference that the information is collected this time from third parties.

However, for our purpose, the most relevant article is 15 (1) (h) since it explicitly mentions the obligation imposed on data controllers by introducing the need to provide "meaningful information" on the logic involved in such decisions, considering the significance and envisaged consequences of such processing for the data subject.

This article is the source of the term “right to explanation.” The term does not appear precisely in the regulation, which has given rise to multiple discussions on the semantic scope of this wording difference.

However, the interpretation of “meaningful information” as the “right to explanation” is confirmed by Recital 71 of the GDPR in this way: “suitable safeguards …should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment to challenge the decision”.

Otherwise stated: “An individual has a right to explanation of an individual decision because that explanation is necessary for her to invoke other rights- e.g., to contest a decision, to express her view” (Kaminski et al., 2019).

So, although the term explainability concerning artificial intelligence has become more relevant nowadays, it is not a new concept but dates from some decades ago, “when AI systems were mainly developed as (knowledge-based) expert systems” (Confalonieri et al., 2021).

Nowadays, lawyers, AI experts, and other stakeholders are paying so much attention to this vital interdisciplinary research field and investing budget and effort in it, undoubtedly due to the European regulation, GDPR, with “the right to explanation,” or better to say, with articles 13-15 that provide rights to “meaningful information about the logic involved” in automated decisions.

Nevertheless, explaining why an automated system makes one decision and not another becomes more complicated when more sophisticated ML or AI systems come into play. Indeed, “the black box nature of these systems allows powerful predictions, but it cannot be directly explained” (Adadi & Berrada, 2018). Due to the relevance that this field is gaining in terms of its implications in legal, technical, and social areas, it is convenient to broadly review the term explainability.

5.1.3 Defining Explainable AI

In general, XAI, or Explainable AI, is an interdisciplinary field that aims to "make AI systems results more understandable to humans" (Adadi & Berrada, 2018). However, achieving this goal is taking different directions. The disciplines involved in this area and their approaches are different and even contradictory because of their complex nature. This section presents a broad spectrum of definitions and ways of understanding this emerging area, as shown in the literature reviewed and evaluated.

Adadi and Berrada commented that the use of the term explainability has common aspects in the same way that we currently understand the field of XAI has to do with the intent to explain the behavior of AI-controlled entities in simulation games applications (Van Lent et al., 2004).

Nonetheless, the issue of Explainability began to be present much earlier in the mid-1970s, "when researchers studied explanation for expert systems" (Swartout & Moore, 1993). The relevance and re-emergence of Explainability are undoubtedly the results of a set of events.

Undoubtedly, one of these is the unstoppable increase in the use of AI/ML for automated decision-making in industries and other critical areas like hospitals for diagnostics. This critical use leads to the increasing need to understand such decision-making processes as providing the chain of reasoning from the legal point of view that leads to confident decisions, recommendations, predictions, or actions made by it.

In addition, the legal aspect that arises because of the General Data Protection Regulation (GDPR) has added a significant plus that calls for developing new AI techniques that are accurate and efficient in their predictive power but also explainable and understandable.

Therefore, different communities try to define explainability in various ways. Although this research does not address the etymology of the term or the controversy itself, offering different definitions of the notion of Explainability depending on the concrete area serves as an illustrative fact to understand what this field of AI is facing.

One of the most common ways of understanding or defining the concept of Explainability is to relate it to other terms or AI Principles (Floridi & Cowls, 2019), such as transparency, interpretability, trust, fairness, and accountability.

For its part, The Defense Advanced Research Projects Agency (DARPA) provides some XAI aims in the way to "produce more explainable models while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (Gunning et al., 2021).

Lewis (1986) argues that "to explain an event is to provide some information about its causal history. In explaining, someone with some information about the causal history of some event -explanatory information- tries to convey it to someone else" (Stern, 1981).

Halpern and Pearl consider a reasonable explanation as a response to a Why question that "(a) provides information that goes beyond the knowledge of the individual asking the question and (b) is such that the individual can see that it would, if true, be (or be very likely to be) a cause of" (Halpern & Pearl, 2005). However, if there is one work worth mentioning, it is the work done by Miller, who connects the discussion in social science and Explainability in AI.

His three outstanding findings are: 1) explanations are counterfactual, and humans tend to understand why a particular event happened instead of some other events; 2) explanations are selective and focus on one or two plausible causes -instead of all potential causes – for a decision or recommendation; that is, explanations should not overwhelm the user with too much information; 3) explanations are a social conversation and interaction to transfer knowledge, implying that the explainer must be able to leverage the mental model of the explainee while engaging in the explanation process.

Miller's conclusions refer to the fact that while these three points are vital properties when building helpful explanations, the different notions of Explainability prevalent in XAI only recently started to consider them (Miller, 2019).

Psychology researchers claim explanations should be human-oriented. Therefore, a good starting point in Explainability could be distinguishing goals for this field. It is essential to associate semantic information with an explanation (or its components) for effective knowledge transmission to human users (Hilton, 1990).

To close this extensive list of candidate definitions of Explainability, we highlight the distinction between two types of explanations that arise within philosophy according to their completeness

or the degree to which an event's entire causal chain and necessity can be explained. Thus, we have "scientific" and "every day" explanations.

Miller (2019) points out that "everyday explanations" refer to "why particular facts (events, properties, decisions) occurred, rather than general scientific relationships (Miller, 2019; B. Mittelstadt et al., 2019).

As we have seen above, the list of explainability definitions is extensive and sometimes contradictory. Therefore, it is essential to delve into a corpus that explores the critical concepts encompassed by this vast landscape of Explainable AI.

5.1.4 Why do we need explanations?

Another way of defining the concept of explainability is by the type of objectives it pursues. This approach is the direction of the research work conducted by Adadi and Berrada in their paper entitled *Peeking Inside the Black Box: Survey on XAI*. According to these authors, the need for Explainable AI is evident "for commercial benefits, for ethics concerns or regulatory considerations, XAI is also essential if users are to understand, appropriately trust, and effectively manage AI results" (Adadi & Berrada, 2018). They suggest that Explainability has almost these four specific goals.

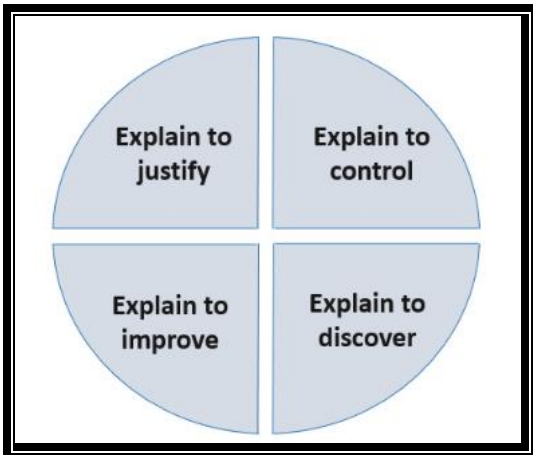


FIGURE 5 | Reasons for XAI (Adadi & Berrada, 2018)

1. **Explain to justify**

In the last few years, there has been an increase in the controversy surrounding the biased or discriminatory results of AI/ML-enabled systems (Howard et al., 2017). This phenomenon has increased the demand for explanations to avoid errors and biases.

When we talk about providing explanations, we are talking about claiming the reasons or justifications that are responsible for a given result or a particular outcome, "rather than a description of the inner workings of the logic of reasoning behind the decision-making process in general (Adadi & Berrada, 2018).

The XAI field becomes relevant here, especially when it is necessary to justify the results when the outcome or decisions are not as expected.

In addition, there are ways such as audits and other mechanisms to defend that algorithmic decisions are ethical and fair, which contributes to building a trustworthy relationship between AI systems and intended users. Moreover, this need responds to the demands of the General Data Protection Regulation (Kaminski et al., 2019) and its "right of explanation."

2. **Explain to control**

This objective has a preventive character and not so much to justify decisions. In other words, understanding the behavior of a system makes it more flexible regarding its vulnerabilities and flaws and helps to quickly identify and correct errors in low-critical situations (debugging).

3. **Explain to improve**

As shown in the following chapters, this goal is one of the most important for this research work. According to Adadi and Berrada, "a model that can be explained and understood can be more easily improved.

Because users know why the system produced specific outputs, they will also know how to make it brighter. Thus, XAI could be the foundation for ongoing iteration and improvement between humans and machines."

4. **Explain to discover**

This last goal or aim is the second directly affecting this dissertation. Thus, asking for explanations is one of the very effective ways of learning new facts, “to gather information and thus to gain knowledge” (Adadi & Berrada, 2018), and these authors add, “so it will come as no surprise if, in the future, XAI models taught us about new hidden laws in biology, chemistry, and physics.”

Undoubtedly, these four goals of Explainable AI are desirable. Moreover, achieving them is like achieving a historical milestone because of the challenges and limitations that the technique entails.

5.2 Enabling artificial explanations: the technical path

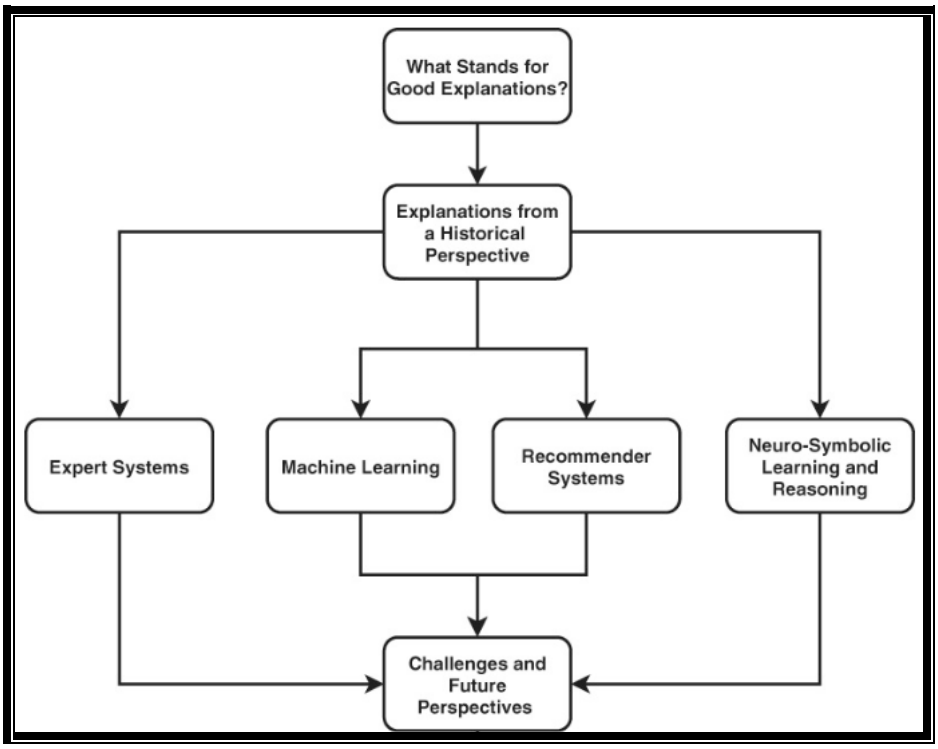


FIGURE 6 | A historical perspective of explainable AI (Confalonieri et al., 2021)

So, if the demand for Explainability is growing in several domains and legislation is becoming a conditional requirement, why isn't there a more systematic corpus, or is explainability not used across the board? The answer is not that simple. As noted in previous sections of this chapter, one of the multiple causes can be this field's interdisciplinary and multifaceted nature.

This section focuses on categorizing the domain of Explainable AI and the technical challenges involved in this field. Confalonieri et al. categorized three groups that "rely on the reasoning characteristics of the underlying decision system: symbolic, sub-symbolic, and hybrid. However, although the authors propose the core of this categorization mentioned above, other literature provides more points of view (Confalonieri et al., 2021).

5.2.1 Explanations in expert systems

The XAI research field originated in the early nineties in the field of expert systems (Xu et al., 2019). The key idea proposed by pioneers such as Swartout and Moore was that the results/ advice provided by expert systems to support humans in decision-making in several domains would be more acceptable if the expert systems could explain why they gave particular advice (Swartout & Moore, 1993).

An expert system consists of a vast collection of rules that try to encode the knowledge of an expert. Bruijn et al. state, "Rules are typically described as implications, from which new conclusions can be derived if specific premises hold." These authors immediately add that the first explanations, named trace explanations, look like this example: "The system came to the diagnosis because it applied these rules to these initial symptoms, thereby concluding that the patient has this sickness" (de Bruijn et al., 2022).

Explanations as a trace line of reasoning

PREMISE:	(AND (SAME CNTXT GRAM GRAMNEG)
	(SAME CNTXT MORPH ROD)
	(SAME CNTXT AIR ANAEROBIC))
ACTION:	(CONCLUDE CNTXT IDENTITY BACTEROIDES TALLY .6)
IF:	(1) The gram stain of the organism is gramneg,
	(2) The morphology of the organism is rod, and
	(3) The aerobicity of the organism is anaerobic
THEN:	There is suggestive evidence (.6) that
	The identity of the organism is bacteroides

TABLE 5 | Example of a MYCIN rule (B. G. Buchanan & Shortliffe, 1984)

One of the most famous systems that offered such explanations is MYCIN (B. G. Buchanan & Shortliffe, 1984). MYCIN was developed in the 1970s and is a knowledge-based system created to provide doctors with diagnostic and therapeutic advice about patients with an infection.

This expert system works with a static knowledge base containing domain-specific knowledge of an expert and factual knowledge about the particular issue to solve. The input concerned some knowledge about a specific patient, and "the system uses this knowledge to instantiate rules and make the diagnosis corresponding to the specific case" (de Bruijn et al., 2022). This system provides explanations using a general question-answering module and a reasoning-status checker. This module incorporates predefined questions that allow expert users to query the dynamic knowledge and contexts used in specific consultations. Thus, the pre-set module allows for two types of essential explanations, as Bruijn et al. describe it as "a *why* command, by which the user "can ascend the reasoning chain and explore higher goals, and a *how* command, by which the user can descend the chain of inferences exploring how a goal was achieved" (de Bruijn et al., 2022).

The point is that although such explanations as lines of reasoning and understanding logically how the system arrives at its outputs improve the interpretability of expert systems, the truth is that users do not benefit as much from technical explanations. Therefore, a new reconceptualization of explanations arises based on a problem-solving activity.

Explanations as a problem-solving activity

This type of explanation involves going beyond the mere reconstruction of lines of reasoning of a system. These explanations' strengths are that they consider various levels of abstraction and cover a broad spectrum ranging from technical explanations to other formats adapted to different users. Lacave and Diez use Rex's example to illustrate how explanations as problem-solving activities work (Lacave & Diez, 2004).

Therefore, Rex's design aimed to explain how an expert system moves from the data of a particular case to a conclusion (a line of explanation) by building a "story" as an abstract of the expert system's reasoning. Rex was not a part of the expert systems but rather a subcomponent that provided an interface and two knowledge bases: knowledge specification and explanatory knowledge.

The knowledge specification on transitions between hypotheses focuses on where any change requires the satisfaction of some goals and some reasoning cues. Thus, only some hypotheses can be inferred and available to the explanatory knowledge. For its part, explanatory knowledge works as a central component of the explanation process. It models cues, goals, and hypotheses. Thus, the explainer tries to find an "explanation plan" using only transitions whose hypotheses can be proven. The search for an explanation plan is carried out backward from the conclusion until reaching the empty hypothesis.

Once one has an explanation, the storyteller organizes it consistently from data to conclusions. After that, it presents the explanation as a story according to a grammar that models the memory structure built during human story understanding. The basic idea is to extract the information concerning the design of each hypothesis transition from the line of explanations.

Each change is story-free with a setting, theme, plot, and resolution. The story tree is then converted into textual description by the verbalizer that fills in a template with the problem description, goal description, movement description, and the conclusion of the expert system.

Rex is only one example of a vast literature on this topic. Some examples of explanation types that emerged during that time include justification, strategy, and terminological explanations (Gregor & Benbasat, 1999).

The two expert systems described so far have in common that they are based on the so-called symbolic representation. Bruijn et al. state that "symbolic systems use languages or, better to say, symbols, which are understandable by humans and can be used to verify the reasoning. For decision-making, the logic to reach a decision is simulating human reasoning, for instance, through rules "if…then" (de Bruijn et al., 2022).

The weakness of the explanations provided by expert systems is that "while an expert may understand a symbolic system's logic, the logic might not be easy to understand for non-experts" (Preece et al., 2018).

Consequently, the XAI field often concentrates on interpreting if the results are correct, and the term Interpretable Machine Learning is used instead of XAI to explain and present model behavior in understandable terms to humans (Du et al., 2018).

5.2.2 Explanations in Machine Learning

Explainable AI in Machine Learning tries to find an interpretable model that approximates the black-box model as much as possible. Although the literature about Explainable or Interpretable Machine Learning is also extensive (Confalonieri et al., 2021), it is convenient to mention two subclasses or methods that imply understanding an automated model: global interpretability, which has to do with understanding the entire model behavior, and the other one is known as local interpretability and concerns understanding a single prediction.

Global interpretability

This method facilitates the understanding of the whole logic of a model and follows the entire reasoning, leading to all the different possible outcomes. According to authors Yang, Rangarajan, and Ranka, "this class of methods is helpful when ML models are crucial to informing population-level decisions, such as drug consumption trends or climate change" (Yang et al., 2018). However, Adadi and Berrada (2018) argue that "global model interpretability is hard to achieve in practice, especially for models that exceed a handful of parameters.

Analogically, for humans, who focus on only part of the model to comprehend the whole of it, local interpretability can be more readily applicable."

Local interpretability

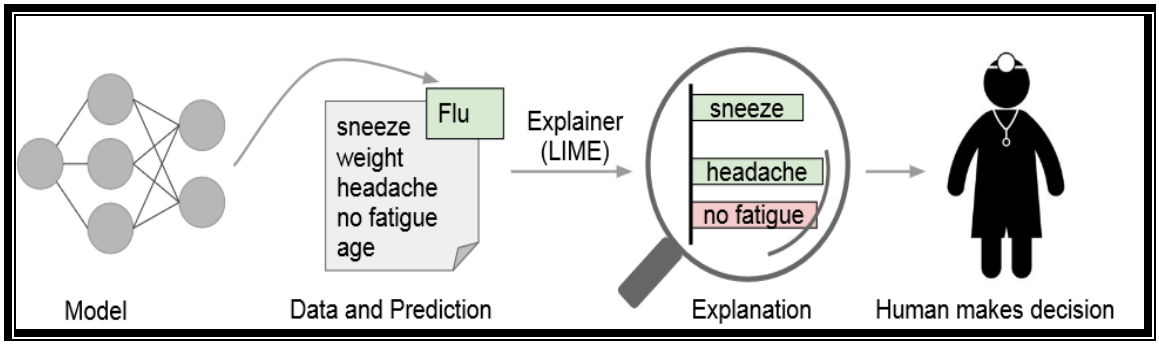


FIGURE 7 | LIME: Explaining individual predictions in the medical context (Ribeiro et al., 2016)

In local explanation methods, the main goal is understanding the reasons for a specific decision or single prediction. This scoop of interpretability intends to generate an individual explanation to justify why the model made a specific decision.

One of the most representative works mentioned in the revised literature is LIME, proposed by Ribeiro et al. The LIME approach exploits the fact that the trained black-box model can be queried multiple times about the predictions of particular instances.

By perturbing the data used for training, LIME generates a new dataset after feeding the black-box model with perturbed data and creates a new interpretable model from the predictions made over the new dataset (Ribeiro et al., 2016).

As explained at the beginning of this section, the ML models, also known as non-symbolic systems, reach decisions by connecting the inputs with the outputs. In addition, as Bruijn et al. (2022) mention, "the internal representation of these non-symbolic systems is more cumbersome, as translation steps are required to make them understandable to humans" (de Bruijn et al., 2022). That means that only post-hoc analysis could help to verify the results obtained by the systems. On the other hand, these models are highly dynamic as they are continuously being trained with new data and learn from their own decisions. Therefore, continuous training and decisions must be updated, and explanations must be constantly given.

5.2.3 Measuring and evaluating explanations

Another aspect that needs to mature in the XAI field is the evaluation of explanations. Doshi-Velez and Kim question this subject by asking, "Are all models in all defined-to-be-interpretable model classes equally interpretable?" (Doshi-Velez & Kim, 2017).

In the same vein, we have, more recently, the work of Miller (2019), and according to him, "most explanations rely on causal relations while people do not find likely causes very useful, and states that simplicity, generality, and coherence are at least equally important. Years before, Lipton pointed out in the same direction that "the question of correctness has been dodged, and only subjective views are proposed" (Z. C. Lipton, 2016).

Doshi-Velez and Kim (2017) go beyond criticism and suggest a way to evaluate the explanations based on Human-Computer Interaction (HCI) user tests.

The authors established three approach types of interpretability evaluation taking into account the economic cost from the most specific and expensive to the most general and cheapest: (1) application-grounded: evaluation with real humans and fundamental tasks or better said, it consists of putting the explanation into the application and let the end user (typically a domain expert) test it; (2) human grounded evaluation with real humans but simplified tasks or in other words it is about conducting simplified application-grounded evaluation where experiments run with lay humans rather than domain experts; (3) functionality-grounded evaluation does not involve humans, and it is only appropriate, in case we have a class of models that have already been validated, for example through human-grounded studies.

Taking as a starting point the work conducted by Doshi-Velez and Kim (2017), the authors Mohseni and Ragan proposed and presented a human-ground evaluation benchmark for evaluating instance explanations of images and textual data. Through their research, these authors demonstrated that comparing the explanatory results from classification models to the benchmark's annotation meta-data allows for evaluating the quality of local explanations (Mohseni et al., 2018).

On the other hand, the Explainable AI DARPA program (XAI) started in 2017 to cover different gaps in this topic and has since opened up new lines of research to meet the challenges of this domain. Some of the goals of this program are "to generate trust and facilitate the appropriate use of technology" (Gunning et al., 2021).

What this literature makes clear by now is the importance of establishing dynamic evaluative goals and metrics to consider what kind of systems are better than others and why. The following section presents the existing literature about how natural explanations, or how humans communicate, can improve the domain of Explainability.

5.3 Explanations inspired by the way humans/nonhumans communicate

Predicting, understanding, and explaining are three different actions. While ML systems are more accurate in making predictions, understanding and explaining them are vastly different.

According to Adadi and Berrada (2018), explaining depends on what is explained (i.e., the original model) and how the explanation is made (the interpretability method), while understanding depends, in addition to these elements, on who is receiving the explanation (i.e., explainee). Although it is not the subject of this research paper, the situation would be hugely different if one AI system required understanding and explanations from another artificial system. For now, we are interested in an ML model that can be explained to us; for this purpose, it must be understandable to humans.

One way to move in this direction is to take as a source of inspiration other human sciences, such as philosophy, psychology, information sciences, neurosciences, and communication, to cite some of those with a long tradition in the study of explanatory and decision-making processes. Based on this idea, this section is concerned with reviewing and describing representative literature on how humans produce explanations and how this process can contribute to the development of what is known as "human-centered explanations" (Adadi & Berrada, 2018) or "user-centered explainable AI (Ribera & Lapedriza, 2019).

5.3.1 Linking targeted explanations

This first part discusses the work done by Ribera and Lapedriza, who focus on two central axes to create better explanations. The first axis deals with providing more than one explanation, each targeted to a different user group.

The second axis is related to making explanations that follow cooperative principles of human conversation. According to these authors, "explanations are multifaceted and cannot be attained with one static explanation" (Ribera & Lapedriza, 2019).

In addition, they suggest creating different explanations for every need and user profile.

This separation and way of work permit contextualizing current developments in Explainability, which considers the explanations' communicative nature and the categorization of the explainees into three main groups based on their goals, background, and relationship with the product (Brasse et al., 2023). Based on this division, Ribera and Lapedriza (2019) also suggest three levels with content that supports every stage and the evaluations of each separately:

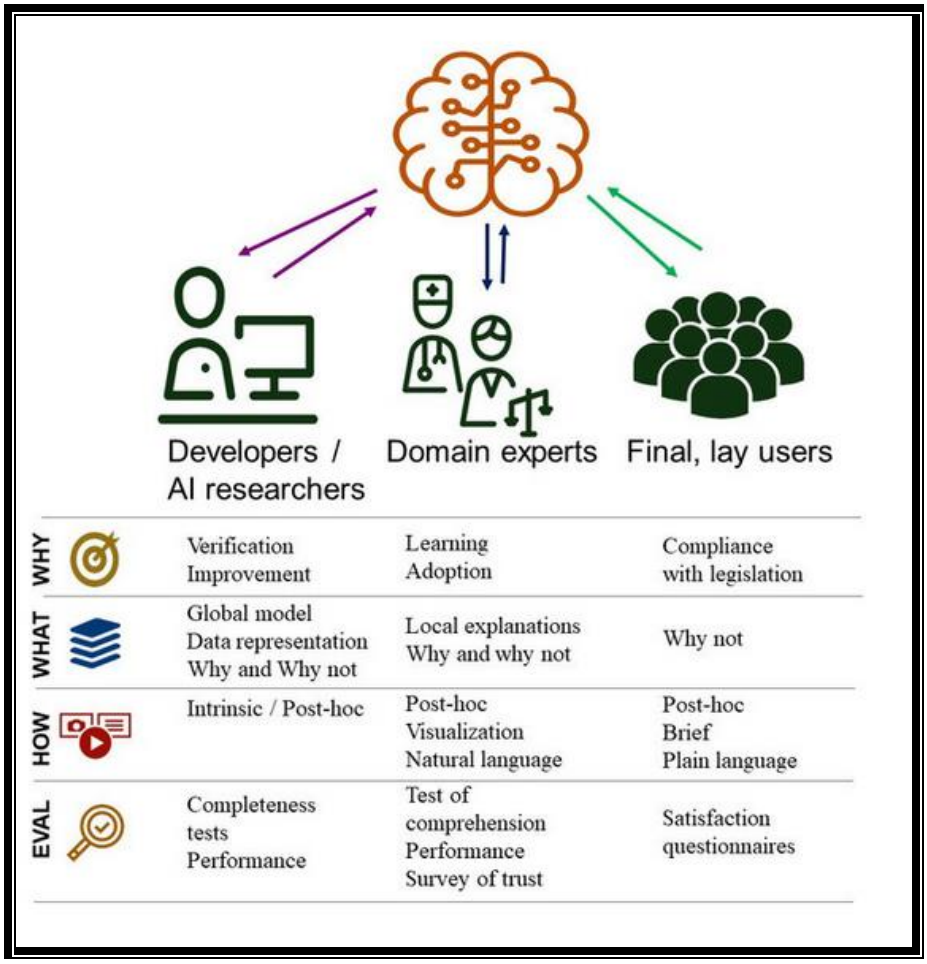


FIGURE 8 | The system targets explanations to different types of users (Ribera & Lapedriza, 2019)

Developers and AI researchers: at this level, the explanations of model inspection and simulation with proxy models are very well suited to verify the system, detect failures, and give hints to improve it. The mode of communication fits well with the audience, who can understand code, data representation structures, and statistical deviations. In addition, completeness tests covering different scenarios are used to evaluate explanations.

Domain experts: the type of explanation provided at this level could be through natural language conversations or interactive visualizations, letting the expert decide when and how to question the explanation and lead the discovery themselves. Explanations must be customized to the domain experts' discipline area and the context of their application, be it legal or medical decisions or any other, to be clear and use the discipline terminology. The evaluation of explanations can be managed with tests of comprehension, performance, and a survey of trust.

Lay users: Wachter et al. (2018) argue at this level the use of outcome explanations with several counterfactuals with which users can interact and select the one most interesting to their particular case. This explanation simulates how humans give explanations and are highly likely to generate trust. One way to evaluate this explanation type is through satisfaction questionnaires (Wachter, Mittelstadt, & Russell, 2017).

Depending on each profile, different needs and explanations can be detected at these three levels. The goals of verification and improvement appeal clearly to a developer or researcher profile, from whom it appears the goal is to improve the algorithm's parameters or optimization. These goals can be achieved with the help of domain experts, to whom the tool tries to help, as they could be the ones that detect possible system failures.

The domain experts' primary purpose is to learn from the system to understand the mechanisms of inference or correlation used to improve its decision methods or to hypothesize possible general rules.

The right to an explanation targets lay users because the system's decisions may have economic or personal implications. However, this goal can also be necessary for domain experts, who might be legally responsible for the final decision.

Related to the explanation content, Ribera and Lapedriza (2019) add relevant information to work conducted by Doshi-Velez and Kim (2017), who argue that different explanations are required depending on global versus local scope, thematic area, the severity of incompleteness, time constraints, and the nature of user expertise.

To these authors' approach, Ribera and Lapedriza (2019) suggest that we can delve a bit more into this idea, especially in the need to tailor explanations to the user's expertise. Tailor explanations consider that if we have a system that offers explanations at the representational level, describing data structures, these should not be communicated in the same language for developers as for domain experts. Even experts in different domain areas require different explanations.

However, according to Lipton (2016), humans do not exhibit transparency in types of explanations, sustaining that human explanations are always post-hoc. On the other hand, the black box issue in ML systems opens a big challenge for human reasoning to understand them (Wachter & Mittelstadt, 2018).

Nevertheless, the work conducted by Grice (1975) in *Logic and Conversation* describes in detail that explanations must follow the cooperative principle and its four maxims. He describes these four cooperative principles as follows: 1. Quality: Make sure that the information is of high quality: (a) do not say things that you believe to be false, and (b) do not say things for which you do not have sufficient evidence; 2. Quantity: Provide the correct quantity of information (a) make your contribution as informative as required, and (b) do not make it more informative than is required; 3. Relation: Only provide information that is related to the conversation. (a) be relevant. 4. Manner: Relating to how one provides information rather than what one provides. This fact consists of the "supermaxim" of "be perspicuous" and, according to Grice, is broken into various maxims such as: "(a) avoid obscurity of expression; (b) avoid ambiguity; (c) be brief (avoid unnecessary prolixity), and (d) be orderly (Grice, 1975).

Therefore, the first three maxims correspond to the content of the explanation, and the last one refers to the type of explanation.

According to Ribera and Lapedriza (2019), making Explainable AI from a user-centered perspective has two main benefits to these authors.

On the one hand, it makes the design and creation of explainable systems more affordable because the purpose of the explanation is more concrete and can be more precisely defined than when we try to create an all-size, all-audience explanation. On the other hand, this way of creating explanations could increase satisfaction among developers and researchers, domain experts, and users since each receives a more targeted explanation. Concerning the evaluation of explanations, it might be easier to achieve this goal if we have metrics specific to each case.

What stood out in the proposal elaborated by Ribera and Lapedriza (2019) and commented on here in detail is the commitment to develop Explainability at different levels, considering each person's particular needs and maintaining the principles of human communication. In addition, their work critiques explanations that offer global and unitary solutions for all levels.

5.3.2 The challenge of explaining Dennett's Intentional Stance

One of the pillars of democratic governance lies in the judiciary. Thus, superior courts usually have an "inherent" jurisdiction to review decisions made by lower courts, tribunals, and administrative agencies for errors of fact or law affecting the exercise of their jurisdictions. This "reviewability is a concomitant of the rule of law" (Zerilli et al., 2019). In addition, since one cannot appeal a decision without knowing the basis upon which it reaches, transparency or Explainability remains a crucial prerequisite to safeguarding democratic rights.

However, since algorithms have begun to make decisions that directly affect other aspects of human lives, it is not surprising that AI scholars are trying to find or simulate how humans formulate explanations to justify their decision-making.

It is worth remembering that traditional algorithms do not have the problem of transparency that the current deep learning networks pose (Zerilli et al., 2019). While traditional algorithms had their rules and weights prespecified "by hand" (Wachter & Mittelstadt, 2018), the neural networks that implement deep learning algorithms try to mimic somehow the brain's style of computation and learning: they take the form of large arrays of simple neuron-like units, densely interconnected by a vast number of plastic synapse-like links.

This way of processing information is no longer transparent because the system's method of arriving at an output operates independently from human control.

So, how do we address transparency in these deep learning algorithms? We should not trivialize the issue of algorithmic transparency because these systems are already in use to "recognize, detect, or predict speech, gestures, faces, objects, sexuality, politics, criminality, pathology, solvency" -to cite some examples. The demand for transparency in algorithmic decision mechanisms is more than supported. However, if the way we humans make decisions represents the desired standard for transparency, Zerilli et al. (2019) state that AI can already be said to meet it in some respects.

These authors support this assertion because although human beings can establish very elaborate reasons for their decisions, they are not always made based on deliberate reflection but are the result of intuition and personal impressions and are far from transparent. What Zerilli et al. (2019) refer to as "human-level opacity" is often a result of a misunderstanding or mistake about their real (internal) motivations and processing logic.

This issue is subsequently covered "by the ability of human decision-makers to invent post hoc rationalizations" (Zerilli et al., 2019).

Moreover, because of this human ability to create a story about a particular decision, according to Zerilli et al. (2019), "Often, scholars of explainable AI treat human decision-making as epistemically privileged." In the same vein, Mittelstadt et al. (2016) state that "algorithmic processing contrasts with traditional decision-making, where human decision-makers can in principle articulate their rationale when queried, limited only by their desire and capacity to explain, and the questioner's capacity to understand it" (B. D. Mittelstadt et al., 2016).

Therefore, the term black box attributed to how specific deep learning systems reach their conclusions also applies to how the human brain works. In the words of Muehlhauser: "We can observe its inputs (light, sound...), its outputs (behavior), and some of its transfer characteristics (swinging a bat at someone's eyes often results in ducking or blocking behavior), but we don't know very much about how the brain works. We've begun to develop an algorithmic understanding of some of its functions (especially vision), but only barely" (Muehlhauser, 2013).

If we focus only on the functional part of human decision-making, elaborately stating the reasons for a decision may be sufficient most of the time. However, there is a substantial difference between utility and truth.

An explanation may be adequate from the target audience's point of view and inadequate from others. In philosophical terms, practical reason is the domain of reason that is in charge of justifying actions and is different from the so-called "epistemic" or "theoretical reason" that focuses on the justification of belief.

One way of approaching the field of practical reason and Explainability is the "intentional stance" suggested by Dennett (1987). Thus, Dennett describes the physical, design, and intentional levels as three levels of abstraction from which we can explain an object's behavior. It is crucial to emphasize that he considers these levels "stances."

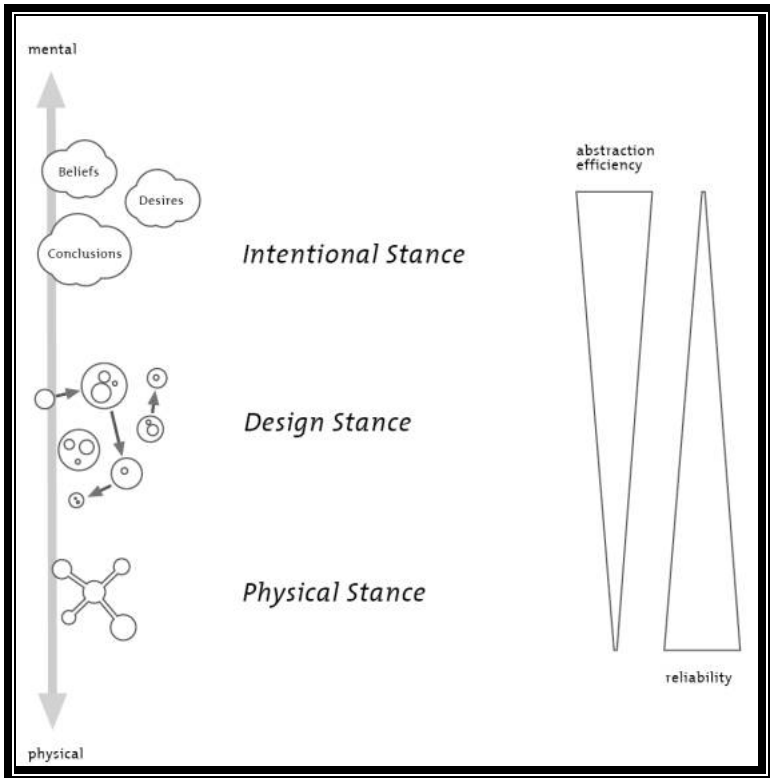


FIGURE 9 | Dennett defines three levels of abstraction, attained by adopting one of three entirely different "stances" (Kitcher & Dennett, 1990)

At the physical level, the structure's features and the object's physical constitution, such as mass, velocity, and molecular arrangement, constitute the system's behavior. In the case of adopting the design stance, the focus is on understanding how the object functions as an integrated mechanism, i.e., how its parts cohere to generate systematic behavior of a specific sort.

An example of this level is the study of human anatomy to understand the behavior of the human body. The last level, and the one of most interest in the question of Explainability, is the "intentional stance." This level explains the behavior of a system in terms of mental states, i.e., folk concepts, propositional attitudes, and belief-desire psychology. Through this level, we can understand what is known as ordinary human behavior and engage in practical reasoning. The following example helps to understand this "intentional stance" more straightforwardly.

If a close friend decides to stay home on a Friday when we had already closed the plan to go to the cinema in the afternoon and we know the person in question in the sense that this behavior has been repeated in the past, we could predict that a further cancellation of the plans again could be highly probable.

In this case, we appeal to the "intentional state," and in other instances, the "physical" and "design" levels do not provide more details to cover the goals of the practical reason (Zerilli et al., 2019).

The underlying question is whether it is necessary to cover the "physical" and "design" instances to meet the requirements of practical reasoning. In the case of the above example and the opinion of author Dennett (1987), "a little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do" (Kitcher & Dennett, 1990).

However, the controversy between these three levels is whether they have the same ontological status. That is to say, while the "physical" and "design" stances may be "real" objects, our mental states, such as intentions, emotions, and desires, still escape this classification.

Thus, Dennett's intentional stance is part of the theory of mind (ToM), which at the moment is based more on an interpretation of what others feel, think, and desire than on having the objective certainty of the mental contents, which in the decision-making processes push to opt for a particular behavior and to disdain others.

In the same way, this intentional stance obeys the form in which human agents construct post hoc explanations, which might be subject to sources of error such as bias and noise. These sources of human error are an issue to watch closely for their unintended effects on how natural and artificial explanations are constructed, as they directly impact how responsibilities are assigned.

5.3.3 Designing a multi-level explanations approach for achieving Broad XAI

To close this section devoted to the existing literature on how the field of Explainability draws on theoretical frameworks that refer to how humans and nonhumans communicate, the research introduced by the authors Dazeley et al. (2021) is more future-oriented, and their approach is innovative and challenging.

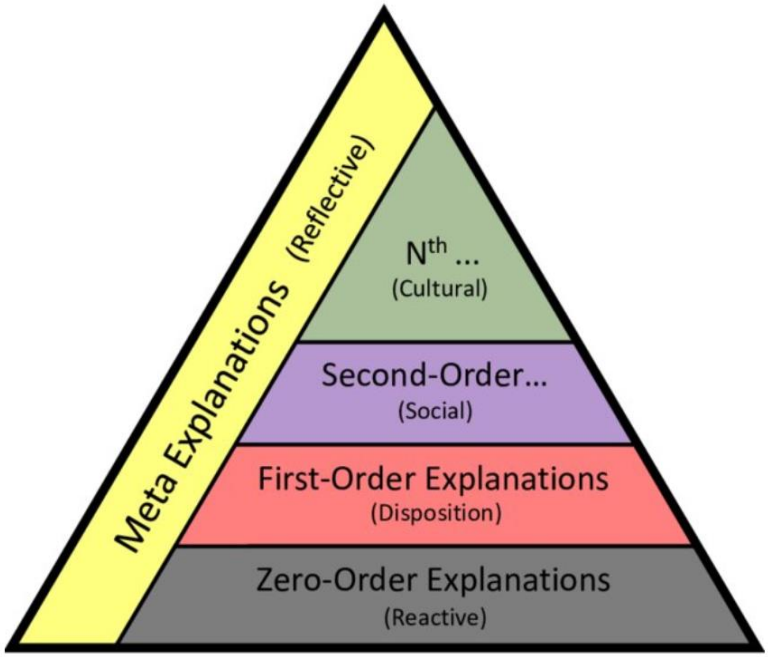


FIGURE 10 | Levels of Explanation for XAI, providing a bottom-up constructivist model for explaining AI agent behavior. This model is adapted from Animal Cognitive Ethology's level of intentionality (Dazeley et al., 2021)

Based on a theoretical approach, these authors discuss the possibility of integrating several technologies to achieve different levels of Explainable AI for Human-Aligned Conversational Explanations to move towards what they call: “strong high-level explanations” (Dazeley et al., 2021).

In some way, the research conducted by Dazeley et al. (2021) tries to merge the other approaches to Explainability described already in this section. Thus, we have, on the one hand, the work conducted by Ribera and Lapedriza (2019), which uses the principles of human conversations.

On the other hand, the second part of this section focuses on the research carried out by Zerilli et al. (2019), which centers on explaining the behavior of a system in terms of mental states and explains the importance of the area of theory of mind (ToM) in the science of decision making or behavior.

Another element that justifies the importance of this research work is that it is the prelude to what has come to be known as XGDAIs (Sado et al., 2020), which is, after all, the domain in which the proposed framework is inserted.

The following is a description of the fundamental aspects of this research work, divided into four levels, and a further one is called meta-explanations.

However, in addition to describing these levels, particular emphasis is placed on the aspects that raise doubts about whether explainable data-driven XAI can be considered an adequate instrument to ensure the principle of algorithmic accountability.

Zero-order (Reactive) Explanations

Most current research in XAI focused on interpreting a single decision point based on data provided to generate that decision. This explanation refers to local explanations (Linardatos et al., 2020). This type of explanation aims to provide a particular decision/conclusion/value/classification based on the data provided. Dazeley et al. (2021) introduce the terminology of Zero-order Explanations to formalize them and define the term in the following way: “is an explanation of an agent’s reaction to immediately perceived inputs.”

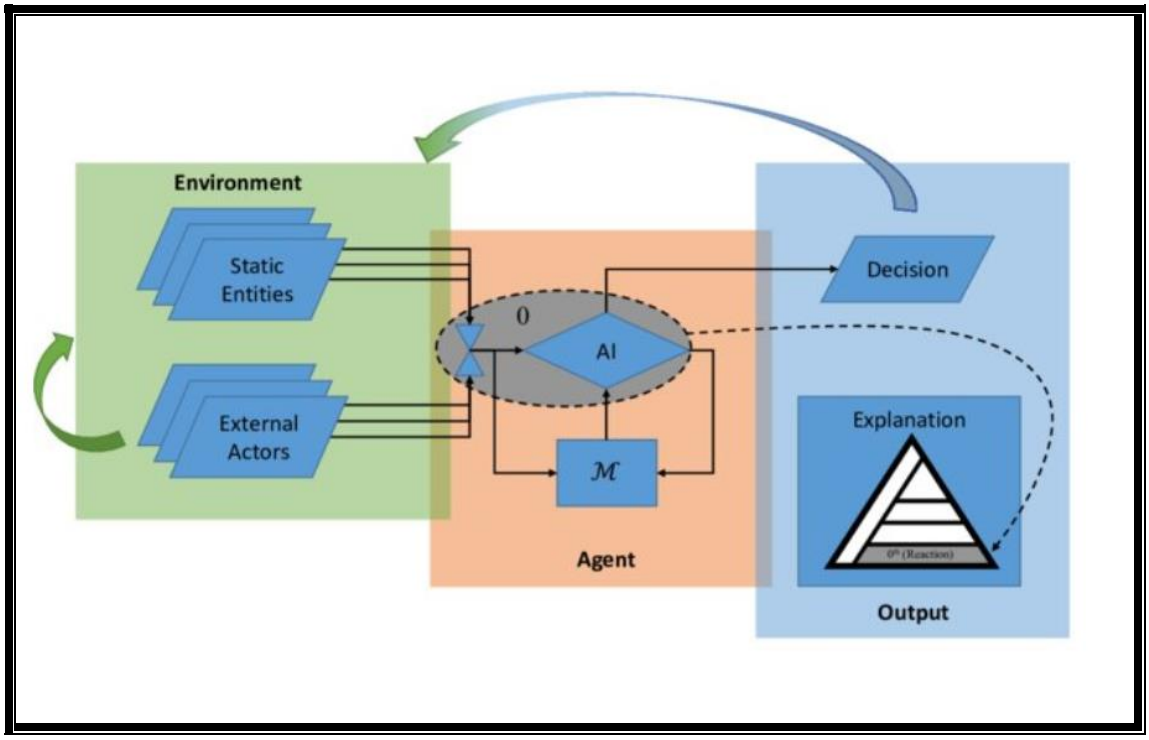


FIGURE 11 | Zero-order (Reactive) Explanation shown diagrammatically, where the grey oval, 0, indicates the focus area of Reactive explanations around the AI agent's decision based on the input interpretation

According to Cheney and Seyfarth (1990), this idea comes from the fact that animals may have no (zero) intentions when simply automatically reacting to a situation, and thus, an explanation of their behavior is based solely on their immediate environment.

For Dazeley et al. (2021), this level of explanation allows the agent to answer questions such as: "What input features were the most important when making a decision?"; "What objects were observed in the environment? or "What features of the state made it decide to turn left?".

In addition, the AI scholars Doshi-Velez et al. argue that this level of explanation is also appropriate to answer many counterfactual style questions such as: "Would changing a certain factor have changed the decision?"; "Why did two similar-looking cases get different decisions, or vice versa?" (Doshi-Velez & Kim, 2017).

These explanations are at the bottom and are the foundation of all other explanations built on top of this level. This research field is known as Interpretable Machine Learning (IML).

In addition, Huysmans et al. (2011) suggest "that people will either be presented with a transparent *model* of the black box that mimics the behavior, such as a set of rules allowing the user an interpretation of how the system will behave for individual cases" (Huysmans et al., 2011).

First-order (Disposition) Explanations

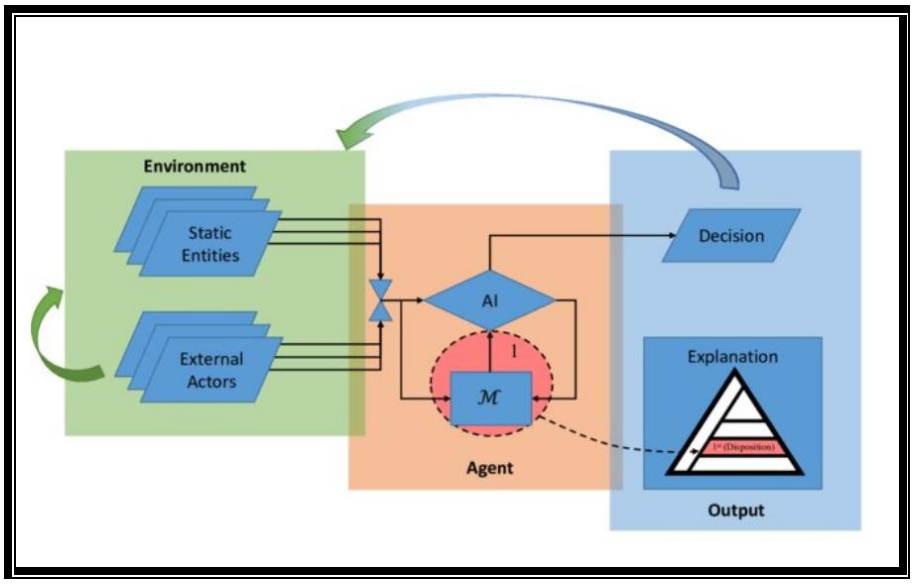


FIGURE 12 | First-order (Disposition) Explanation shown diagrammatically, where the red oval, 1, indicates the focus area of Disposition explanations around the AI agent's decision, based on the interpretation of the agent's Merkwelt, such as its beliefs, desires, or memories

AI systems advance, and the requirements for more meaningful explanations increase considerably. In turn, Tegmark argues that we need to understand an agent's goals, objectives, beliefs, emotions, and memory of the past event and how these affect its reaction to a set of inputs (Tegmark, 2017). Thus, belief, desire, intention (BDI) agents and reinforcement learning are designed to work as goal-oriented or intentional AI systems. Langley et al. describe this as an explainable agency and focus on the idea that many agents are goal-directed (Langley et al., 2017).

Dazeley et al. (2021) consider agency “the capacity of an agent to act independently – making its own free choices.” They referred to this explanation level as a First Order (Disposition) Explanation: “an explanation of an agent’s underlying internal disposition towards the environment and other actors that motivated a particular decision.”

In addition, the authors add that “a disposition explanation could be drawn on the foundational (reactive) explanation but will also incorporate details of the agent’s current internal disposition and how it influenced its reaction” (Dazeley et al., 2021).

Disposition explanations tend to make it possible for agents to answer questions, such as Dazeley et al. (2021) pointed out: “Why did you want to be in that lane?” and “Why did you want to clean the room?”; or counterfactuals like “Why didn’t you wish to play the rook to c7?” The most crucial aspect is to discern the differences between reactive and disposition explanations.

While reactive explanations explain how the agent's inputs contributed to the decision, a disposition explanation explains the agent’s internal motivations that caused it to react to the inputs in a particular way.

In addition, many AI approaches currently try to simulate some notions of belief and desire, such as reinforcement learning, bayesian optimization, dynamic programming, BDI agents, and evolutionary algorithms. These approaches do not have any inbuilt knowledge or control over these beliefs or desires. For the authors, Dazeley et al. (2021) also seem unsuitable in providing a disposition explanation in these simple systems.

Nowadays, there is not much practical research on incorporating this kind of explanation into this framework because of the limitations and the challenges.

However, some approaches, such as the BDI agents, model an agent’s beliefs and desires and how they change over time due to social interactions.

Examples are the works in reinforcement learning and Markov decision processes (MDPs) proposed by different researchers (Madumal et al., 2019), where causal models or predictions about the outcomes intend to explain an agent’s behavior.

The core issue is that these disposition explanations are not exclusively derived from an agent’s goals, objectives, and beliefs, as Dazeley et al. (2021) argued, because they may also be “based

on an agent's emotional state and memory of past events. That is that many learning systems rely not only on their current inputs but also on a memory of past events. Dazeley et al. state that "disposition explanations should be able to provide details of how these elements of their Merk-welt influence their decisions (Dazeley et al., 2021).

In this line of research, we find the conclusive research work developed by Kaptein et al. in the research domain of Emotion-aware XAI. These researchers argue that using emotional words in creating explanations is essential because humans explain their emotions, which is part of the self-explained behavior and a sign of awareness of others' emotions (Kaptein et al., 2018). However, later on, in the chapter dedicated to the different theories on emotions, it is shown whether using emotional categories to justify our behavior is sufficient to understand the intentions of others.

Second-order (Social) Explanations

Let's move more in the direction of Artificial General Intelligence (AGI). We have to say that from the approach of the authors Cheney and Seyfarth (1990) in Animal Ethology, it seems that some animal behavior may indicate an awareness of, or at least an interpretation of, other animals' internal beliefs and desires. The turning point unclear in Animal Ethology is, for example, if the animal is aware of the keeper's intentions or is simply predicting future behaviour. This approach requires higher reasoning about human mental states as metacognitive processes.

These processes are part of one of the primary disciplines in this research, which is the theory of mind (ToM) (Leslie, 1987; Wellman et al., 2001), also previously called mentalization (Wimmer & Perner, 1983).

However, these areas of knowledge are called by other names in popular psychology, such as empathy, emotional understanding, attribution, mind-mindedness, and self-awareness. For instance, according to Holmes, mentalization is "the ability to see ourselves as others see us, and others as they see themselves" (Holmes, 2008).

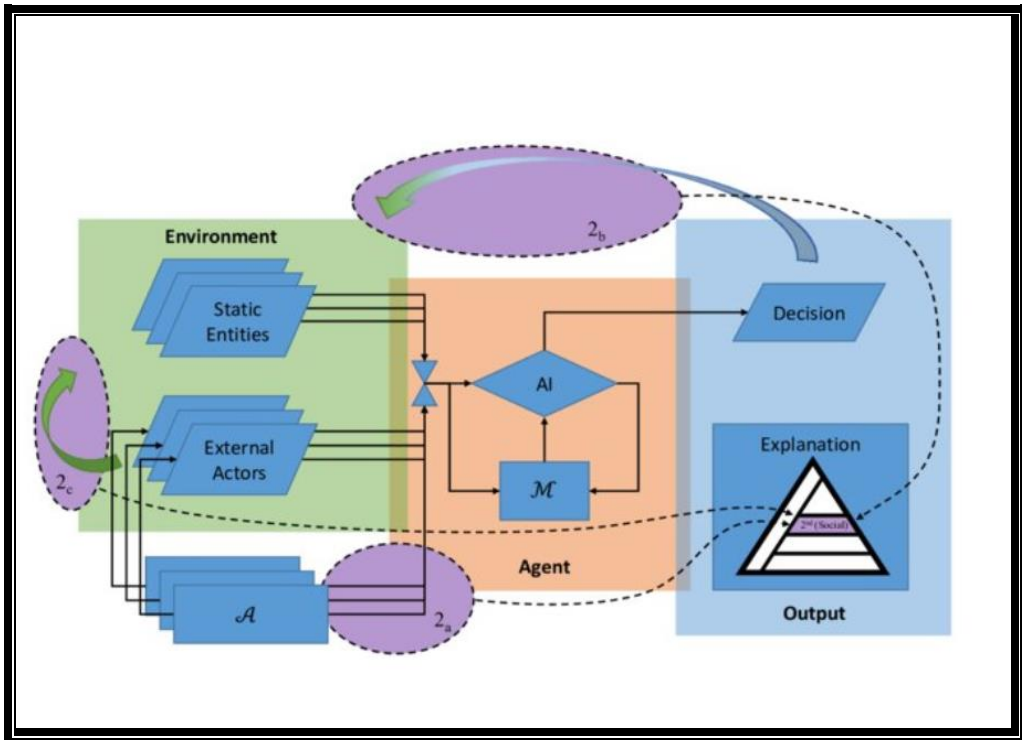


FIGURE 13 | Second-order (Social) Explanation shown diagrammatically, where the purple ovals show that the agent's and other actors' behavior affects the environment, which in turn affects the Merkwelt of the agent itself and the other Actors' Models. The agent models the other actors' potential internal state and memory to predict their future behavior

The vital aspect of this kind of explanation level is that there is the expectation that the system discusses not only the expected behavior but also that to get the actor's intention, "it will be required the model to predict the actor's internal state and likely memory" (Dazeley et al., 2021).

In addition, these same authors construct the following definition of these social explanations. Thus, Second-Order (social) Explanations reside on an awareness or belief of its or other actors' mental states. This level of explanation could give the agent a deeper grasp of the fact that other environmental actors have their own Merkwelt.

These explanations align with basic human reasoning in social interactions and make it possible for an agent to respond to its environment and reach targeted goals. Let's think about what kind of AI systems we should be able to develop.

We require this higher standard of reasoning, for example, for technologies such as robot carers, autonomous self-driving cars, and autonomous personal organizers.

Dazeley et al. (2021) propose this example for this explanation: "Why did you slow down when the pedestrian approached?" The answer could be as the authors said: "I slowed down because the pedestrian approaching the intersection appeared intent on the crossing." This explanation goes more in-depth than the simple reasoning behind actions and tries to explain how other actors have affected their behavior.

Thus, the relevant part of this type of social explanation is not the interpretation of the actor's model but the justification of how that outcome affected the final decision. Therefore, for this research, the work carried out by Kaptein et al. is of maximum interest and usefulness since the core of their work on Emotion-aware XAI (EXAI) (Kaptein et al., 2018) is to explain the agent's internal and external actor's emotions and how they affected the agent's decision-making.

N th-order (Cultural) Explanations

The social level of explanations concerns a relatively simple degree of interaction because, in reality, in human interactions, rather than obeying a fixed, pre-established standard, the rules of the game are set by expectations.

To illustrate this, Dazeley et al. take the following example: "Imagine that the autonomous car always gives way to all other cars in every situation. Such a car may never go anywhere and quickly become useless to human passengers. Furthermore, it will not behave like human drivers, creating confusion and risk as it does not meet other drivers' behavioral expectations" (Dazeley et al., 2021).

Expectations require another reasoning system equivalent to what the philosopher Dennett (1983) calls "third-order intentionality" (Kitcher & Dennett, 1990). In this way, expectations are not only because agent A is convinced that agent B will make a particular decision, but agent B expects agent A to do what is expected. This phenomenon would fall under what the authors Dazeley et al. (2021) call "an understanding of a set of cultural rules about behavior" (Dazeley et al., 2021).

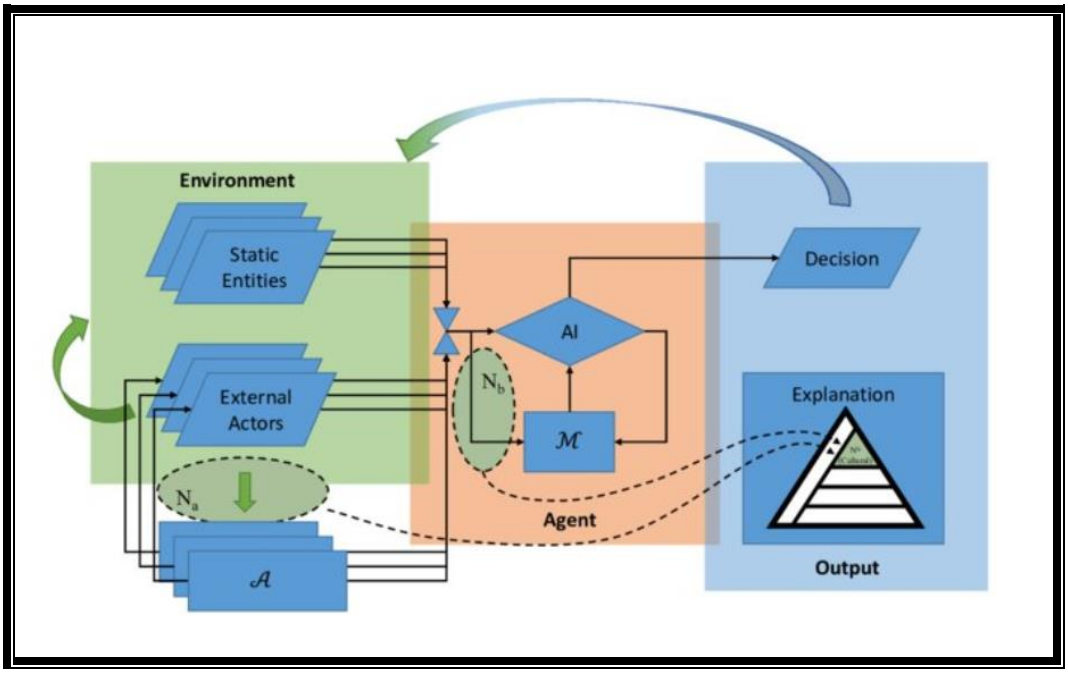


FIGURE 14 | N th -order (Cultural) Explanation shown diagrammatically, where the green oval indicates that we need to model how the actors' potential internal state and memory are affected by changes in the environment, while indicates the need to explain how the changing environment affects the agent's Merkwelt

The fundamental difference between social and cultural explanations is that in the former, the agent's objective is to predict the other agent's behavior. At the same time, the latter focuses on the agent's expectations and how to respond to the agent's behavior.

However, although this form of behavioral modeling has well-established literature, these models are still not operational in dynamic human environments, and therefore, what practical application these explanations might have for AI systems and society is still unknown.

Meta (Reflective) Explanations

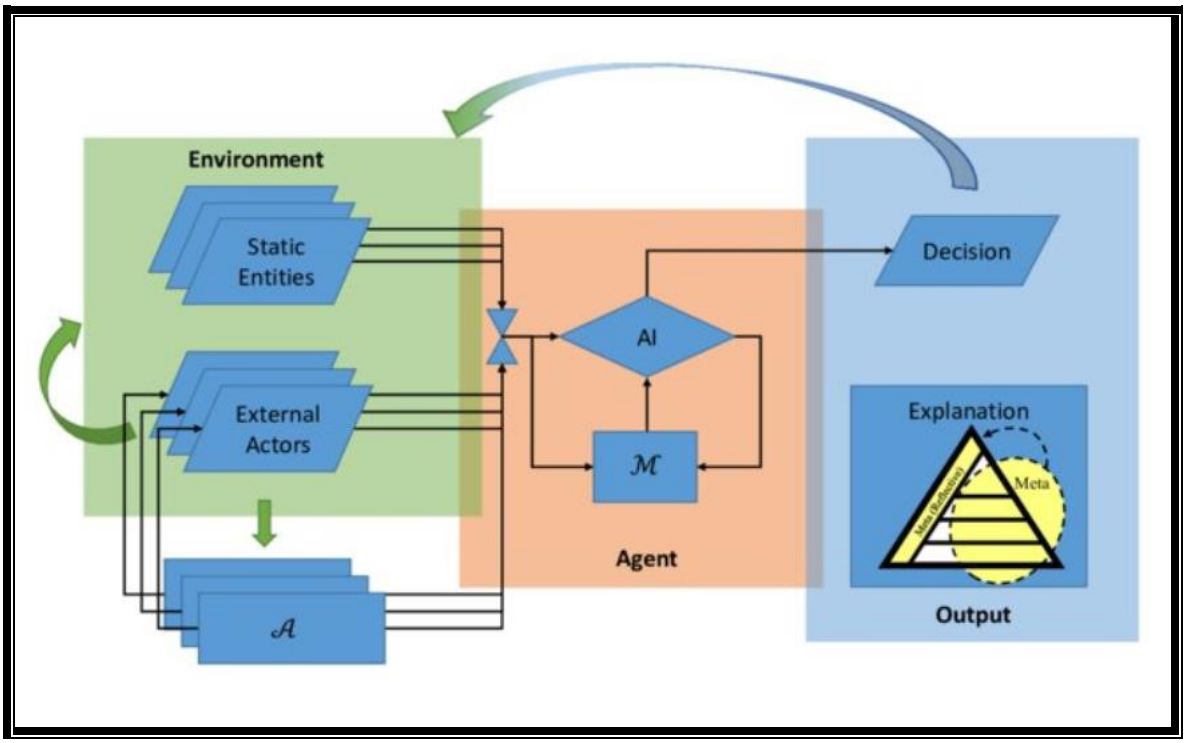


FIGURE 15 | Meta (Reflective) Explanation is shown diagrammatically, where the yellow circle, Meta, encompasses the explanation levels, indicating that a meta-explanation is concerned with inferring how each explanation level was mapped.

Arguably, one of the primary goals of Explainable AI is to make itself understandable so that users can be confident in their decision-making. However, in the words of the authors Dazeley et al. (2021), how can we trust a system's explanation of a decision we distrust?

According to the works of Abdul et al. (2018), human-centered interaction methods focus on the user's acceptance and understanding. In other words, these methods are more centered on "improving the understandability of the explanation at the detriment of accurately reflecting on the implementation details behind the decision" (Dazeley et al., 2021). Therefore, this human-centered approach to XAI could provide potentially untrustworthy explanations with challenging consequences.

Dazeley et al. (2021) consider in their paper these three critical points in the provision of comprehensible explanations:

1. Utility-driven or data-driven XAI -using human acceptance and understanding to assess the success of an agent's explanation.
2. Deceptive explanations -where an agent needs to hide its true objective from the explainee.
3. Simplified/generalized explanations -inadvertent deception through omission.

If one wants to enhance trust in the explanations that artificial systems give for their decision-making, implementing the explanation component in these systems should focus not only on the explanation but also on understanding the process used in formulating the explanation.

Reflecting on the conclusions performed by Dazeley et al. (2021): "A truly broad-XAI agent would need to be able to perform a reflective exercise on its explanation process." Therefore, it is categorized as meta-explanation and descriptively referred to as reflective explanation.

These authors define a Meta (Reflective) Explanation as an explanation detailing the process and factors used to generate, infer, or select an explanation. The novelty introduced in this paper is that, on the one hand, the concept of Meta-Explanation is not a term used in the XAI literature. On the other hand, the main goal is not to justify why a decision is the right or correct one but to explain "the problem-solving in solving it" (Dazeley et al., 2021).

Thus, while most of the efforts in the field of data-driven XAI are still focused on providing explanations of particular decisions or arguments for a specific output or behavior, the meta-explanation level contributes to referencing the structures behind the explanation or argument and may include historical decisions or justifications. The exact nature of a meta-explanation determines the chosen method that generates the original explanation.

This approach to meta-explanations is contrary to “regular explanations whose aim is to satisfy the explainee’s need to understand the reasons for the decision and not the processing method used” (Dazeley et al., 2021). Furthermore, the meta-explanation adds a component of objectivity to the regular explanation, which, although less understandable to the user, results from a more honest and accurate reflection of the decision and the provided explanation.

However, if there is a thorny issue in this meta-explanation landscape, it is precisely its regulation. As discussed above, the issue of “intentionally deceptive agents created by organizations will frequently require the exact decision-making process and its associated explanation generation to be hidden from clients” (Dazeley et al., 2021).

In addition, forcing companies to disclose the true causes of their decision-making has never been easy, so much so that companies themselves or other institutions or organizations would argue that regular explanations would be sufficient.

This argument that regular explanations are sufficient has two ethical consequences that should be emphasized: on the one hand, it is impossible to acquire a more rigorous and precise knowledge of human error sources, such as biases and noise. On the other hand, the inability or unwillingness to update knowledge about the motivations behind human judgments paralyzes attempts to create new mechanisms or social norms that advocate a principle of responsibility in more ethical and scientific terms.

The challenging framework: a Broad-XAI Conversational Process

We are closing this part of the dissertation by describing the constraints in the work developed by Dazeley et al. in the configuration of their proposal. The limitations referred to here can be found in similar works covering the field, especially of data-driven XAI.

In summary, the framework presented by these authors aims to involve an interaction in the form of a conversation between the explainer and the explainee. Through this method, the explainer’s proposes causes of an event or outcome such that the explainee understands and accepts the explanation.

This process works like the fundamental human communication process, which informs the explainee's cognitive function. On the explainee's communication side, the identified causes and counterfactuals aligned with the explainee's currently accepted understanding of the world.

The explainee can then accept or reject the provided explanation. Alternatively, the explainee may require the agent to provide backing claims by requesting additional, more specialized, or detailed explanations to resolve any identified internal conflicts.

A closer look at this proposition leads to the conclusion that this is the way communication between humans works. Is it not this way of proceeding that Dazeley et al. (2021) criticized in the regular explanations that two human agents could propose in any given situation?

Moreover, if the role of meta-explanation seemed crucial in Dazeley et al.'s research work (2021), why are they no longer part of the conversational process between the agents involved?

Would it not be more appropriate to study the mental contents of our decision-making processes in human communications and see how we can translate this into how we can communicate with artificial agents?

The shortcut of mimicking how humans communicate to build trust is convenient, but if we want to gain more accuracy, we must keep looking.

Due to Explainability's current challenges and limitations, the last part of chapter 5 discusses the following arguments:

1. If one has to discuss biases in AI, it is necessary to understand their origin and detect or frame them as much as possible.

2. What kind of transparency should we demand from the black boxes of AI systems in their decision-making if the way humans make decisions also functions as a black box? Indeed, how can we design a theory of mind (ToM) in the emerging research domain of goal-driven explainable AI to identify concrete norms of behavior and reasoning in human agents?

3. There is more information in the real world as we can perceive with our senses. In other words, the information or data available to understand phenomena could be more than we can or want to perceive, and we only choose a few data because we believe they are the most relevant or because we decide to bet on them arbitrarily.

Therefore, the open questions are: what are the dangers of only interpreting specific data without a specific context? How do we implement an explanatory component of an AI system to provide us with new information about a specific domain that aims to safeguard the principle of responsibility over time in our case?

5.4 Explaining AI biases as a source of human error

One reason explainability has become such a renowned field today could be the urgent need to understand how these AI systems provide their answers, solutions, or outputs.

In addition, this wave of high funding and interest stems, among other reasons, from a series of systematic failures, inconsistencies, and shortcuts from these technologies in areas that affect human lives in a deep and wide variety of ways.

There are multiple examples, and we cite below some of them that have been reported and analyzed in the international literature and media for their impact.

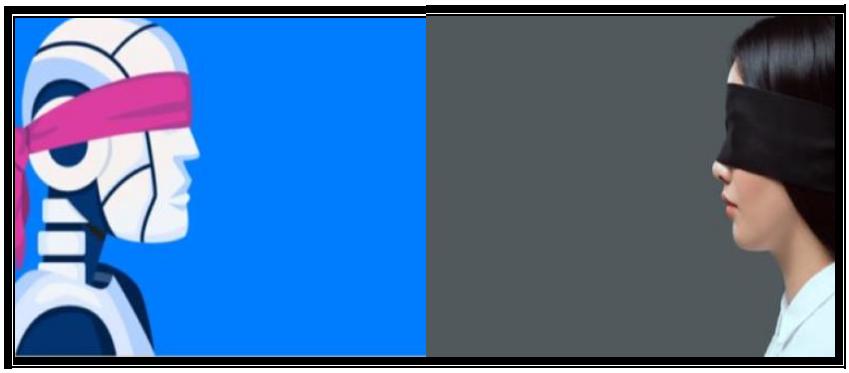


FIGURE 16 | "We can be blind to the obvious, and we are also blind to our blindness" (D. Kahneman, 2011).

One of the most frequently cited examples is "gender bias." How this works can be simple: use an internet search engine and ask it for a list of the "greatest leaders of all time." The search engine answers were Mahatma Gandhi, Martin Luther King, Abraham Lincoln, Nelson Mandela, and Winston Churchill. But then, what do all these names have in common? There was no woman. This example comes from a UNESCO initiative entitled *Building Partnerships to Mitigate Bias in AI* (unesco, 2023) and gives more examples of how search engines are configured and deliver their results addressing the topic of gender biases. The explanation for this result, which may go unnoticed by many of us, has its particular origin in a deep-rooted stereotype and prejudice in society.

However, another instance of gender bias that could have had real repercussions was Amazon's hiring tool, which in the year 2018 discriminated against women (Hamilton, 2018). The story starts when, four years before, five machine-learning specialists were working on hiring algorithms to sift through job applications for use in a recruitment tool. This tool works in the same way that Amazon products rate with one to five stars. In 2015, it appeared that the algorithm discriminates against women.

The clue was to review how the model was trained by analyzing patterns in résumés submitted to Amazon over the past 10-year period. The technology industry has been and continues to be linked to the male population, so the applications that the computer could observe came from men.

Before moving on to the following example, it is worth highlighting the following information from this article: *Amazon's sexist hiring algorithm could still be better than a human.*

This news item says: "While there is a common belief that algorithms are supposed to be built without any bias or prejudices that color human decision-making, the truth is that an algorithm can unintentionally learn bias from various sources.

Everything from the data used to train it to the people who are using it, and even seemingly unrelated factors, can all contribute to AI bias". In addition to that, the article explains the following: "Since the data used to train it was at some point created by humans, it means that the algorithm also inherited undesirable human traits, like bias and discrimination, which have also been a problem in recruitment for years" (Lavanchy, 2018).

But if a tangible example of AI bias is in the scientific literature in detail, it is the case of COMPAS (Lagioia et al., 2023). Thus, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm used in courtrooms across the United States to predict future crimes, but this AI tool was biased against black defendants.

COMPAS training data contains factors such as employment, previous arrest, and age. Its output-risk scores for recidivism- are one of the factors judges use to determine whether defendants will face jail or prison time.

The case was that COMPAS black defendants were incorrectly labeled as "high-risk" to commit a future crime twice as often as their white counterparts.

However, the core aspect of this research is the declarations of the parent company of COMPAS, called ProPublica.

The fact is that despite the critics from the scientific literature and other media of the COMPAS algorithm, ProPublica states that "the algorithm was working as intended" (ProPublica, 2016). And it is this claim that is the subject of an in-depth study about Explainable AI.

After mentioning a few examples that seem to have common elements, what are some of the causes of these biases? Are the explanations offered a guarantee of objectivity and fairness?

5.4.1 Bias in human and algorithmic decision-making

In the last 30 years, research on human decision-making has made great strides, and one of these worth mentioning is the central role that human emotions play in decision-making (Damasio, 1994). This aspect is remarkable because Zerilli et al. argue that "the centrality of emotions in human decision-making at once suggests a unique contribution automated decision technology can make to practical reasoning: it can significantly reduce one of two potential sources of bias and discrimination" (Zerilli et al., 2019).

While the authors cited above, especially the work by Zerilli et al., argue that the effect of emotions may be reason-distorting, this research follows Damasio's hypothesis for whom emotions are an essential component of rational thought. This view of emotion does not mean emotions can not engender bias in some cases. In Damasio's words: "Nonetheless, (…) reduction in emotion can constitute an equally important source of irrational behavior" (Damasio, 1994). This statement is the same as saying Damasio sees emotions as a potential source of bias in some contexts, not in all, and identifying when they are and when they are not a source of bias and noise is a task yet to be accomplished.

So, what are the most common types of bias? We start with one of the most presentable types of biases in human agents: the so-called "intrinsic bias" (Zerilli et al., 2019). An example of this type of bias is racial bias, whose connection with the emotion of fear is easy to detect. When someone believes that a type of ethnic minority poses a danger to its health, safety, or integrity, "merely supplying that person with evidence to the contrary may be insufficient to dislodge a lifetime of encrusted prejudice" (Bezrukova et al., 2016).

Thus, we have that the type of human biases commonly known are of this intrinsic type, which bears an essential relation to emotion, a constitutive feature of personality (Angie et al., 2011; Stephan & Finlay, 1999). This racist conditioning, considered permanent or intermittent, affects how people process information and make decisions. This phenomenon is not to say that all intrinsic biases are always irrational. Some human biases may result from "the misfiring of an ancient and conserved cognitive adaptation to make generic judgments" (Leslie, 1987). Intrinsic or human historical bias is related to the same categorization because it is rooted in pervasive and often deeply embedded prejudices against certain groups throughout history.

Therefore, one of the future challenges is to detect when these biases, which may have been rational at a particular time, need some reinforcement, social norm, or even legislation to correct it because they have lost their normative validity. After all, the situation has changed. An example may be to stop having weapons in a house because one no longer lives in a territory at war or because a war has ceased.

The current problem in AI is that computer models can reproduce or even amplify these human biases systematically and affect large populations in very different ways, unaware of that process.

Extrinsic bias is another relevant bias affecting AI systems' performance and behavior. This type of bias derives from a system's input when it does not affect a permanent change in the system's internal structure and rules of operation. In other words, it is "where errors and biases latent in data training sets tend to be reproduced in the outputs of machine learning tools" (Barocas & Selbst, 2016; Diakopoulos, 2014). That is, the data with which the AI system has been trained does not represent solving the problem. However, this type of extrinsic bias is more accessible because the apparent solution is diversity in the training sets (Crawford & Calo, 2016; Klingele, 2016).

Another type of bias is the one that Friedman and Nissenbaum call "technical." This bias arises from the inherent constraints imposed by technology (Friedman & Nissenbaum, 1996). The following scenario offered by Mittelstadt et al. serves as an example of technical bias: "When an alphabetical listing of airline companies leads to increased business for those earlier in the alphabet or an error in the design of a random number generator... causes particular numbers to be favored" (B. D. Mittelstadt et al., 2016).

Friedman and Nissenbaum also consider the phenomenon of "emerged bias" substantial.

This bias is a product of advances in medical diagnostic tools that do not account for new knowledge, which will be an "unavoidable bias towards treatments included in their decision architecture" (B. D. Mittelstadt et al., 2016). However, this limitation is also part of how human agents operate, who are not exempt from emerging biases.

As Zerilli et al. argue, it is worth considering that "professionals such as medical practitioners, lawyers, and tax agents must maintain a certain standard of knowledge to be considered proficient and that this is generally enforced through mandatory continuing education programs. This measure is an open avowal of the fact that humans are not immune to emergent bias either" (Zerilli et al., 2019).

The thorniest issue concerning the topic of biases is the one discussed by Plous, who observes in his research on human prejudice that "humans are cognitively predisposed to harbor prejudices and stereotypes."

In addition, he adds, "Contemporary forms of prejudice are often difficult to detect and may even be unknown to the prejudice holders" (Plous, 2003). Even if one agrees with this argument, it does not mean solutions should not be sought. Once again, it is remarked that this dissertation hypothesizes that being aware of these aspects of human nature in the way that humans tend to deliberate is the starting point to prevent and mitigate these sources of error in human behavior.

Then, the phenomenon of biases is not only present in the training data of the systems or in the way humans deliberate their decisions but also in how humans explain their reasons to justify their behavior. This assumption is the same as argued by Zerilli et al. as follows:

> "This should force us to reassess our attitudes to human reasoning and question the capabilities of even the most esteemed reasoners. Giving reasons for decisions may be insufficient to counter the influence of various factors, and the reasons offered for human decisions could well conceal motivations scarcely known to the decision-makers. Even when the motivations are known, the stated reasons for a decision can cloak the true reasons" (Zerilli et al., 2019).

The fact is that "the purely neurophysiological aspects of human decision-making are not understood beyond general principles of internal transmission, excitation, and inhibition" (Zerilli et al.,2019).

For example, when a decision-maker must decide between a vast number of factors and weigh the relevance of each to arrive at a final decision, Pomerol and Adam hypothesize that the brain eliminates potential solutions such that a dominant one ends up inhibiting the others in a sort of "winner takes all" scenario. While this process is somewhat measurable, "it is essentially hidden in the stage where weights or relative importance are allocated to each criterion" (Pomerol & Adam, 2008).

These theories and hypotheses lead us to the next section, which deals with the limitations of the human brain, also considered as a black box since this phenomenon represents a significant limitation "on our capacity to process complexity" (Zerilli et al., 2019).

The transparency requirement of two black boxes

As discussed in the previous section, AI researchers argue that deep artificial neural networks are not significantly more opaque than human brains/minds. This hypothesis has been called by the researcher Maclure (2021) "the argument from the limitations of human reasoning" (Maclure, 2021).

In recent years, the science of decision-making has gained much interest from cognitive science, social psychology, and behavioral economics. Thus, before the rational tradition thought that human beings make decisions based on their rational capacity, more and more studies show that emotions play an essential role.

If an author has come to question the rational capacities characteristic of human judgment, it is the author of the book *Thinking, Fast and Slow* (2011), Daniel Kahneman. In addition, this book has become a source of inspiration for the current AI scientific community.

According to Kahneman's dual-system theory, human cognition is said to be ruled by two systems: System 1 for fast, intuitive, unconscious, automatic, imprecise, and effortless thinking. In turn, this system incorporates heuristics, allowing an agent to conclude without using reason for pondering evidence, weighing pros and cons, and drawing inferences from premises (D. Kahneman, 2011).

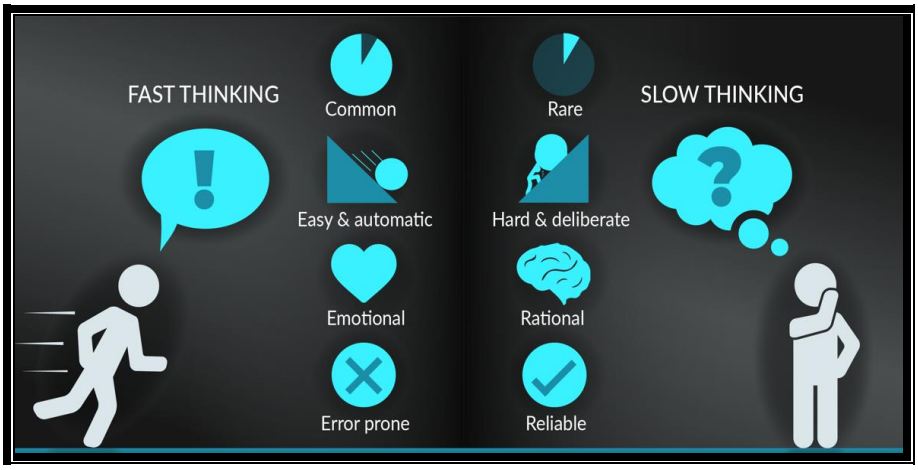


FIGURE 17 | *Thinking, fast and slow* (D. Kahneman, 2011)

However, although System 2 is characterized by slow, analytic, deliberate, and conscious thinking that requires more significant cognitive effort, it has been demonstrated that neither system is free of bias and noise.

Maclure states, “Cognitive biases are not only at play in unconscious belief and judgment formation but also when we think that the conclusions we reached are evidence-based and logically sound” (Maclure, 2021).

Furthermore, in Kahneman’s and co's most recent research, explained in his new book *Noise*, bias and noise are considered sources of human errors. Thus, while biases are illegitimate criteria of judgment that lead to systematic deviations, noise refers to irrelevant factors that lead to scattered and unpredictable judgments, such as the weather, the time of day, or yesterday’s sports results (D. Kahneman et al., 2021).

Another author who criticizes the requirements of a higher machine explainability based on what Maclure (2021) calls "from the limitations of the human mind” was the Turing Prize co-recipient, Geoff Hinton.

According to Hilton (2018), since we do not understand how humans and neural networks make their decisions, and both are not transparent, AI systems should not explain themselves any more than humans do.

This argument, which in this dissertation is considered very vague in terms of its scientific quality, deserves to be commented on because of its media relevance and because it is not an isolated claim.

In 2018 in the year that Hinton received the Turing Prize with other researchers, he gave an interview in which he said the following:

> "People can't explain how they work for most things they do. When you hire somebody, the decision is based on all sorts of things you can quantify and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decisions, you are forcing them to make up a story. Neural nets have a similar problem. When you train a neural net, it will learn a billion numbers that represent the knowledge it has extracted from the training data. If you put in an image, the outcome is the right decision, say, whether this was a pedestrian or not. But if you ask, "Why did it think that? Well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago" (Simonite, 2018).

In the same vein, we have Kahneman's answer in an interview with the economist Eric Brynjolfsson, who said the following: "In general, if you allow people to override algorithms, you lose validity because they override it too often. Also, they override based on their impressions, which are biased, inaccurate, and noisy. Decisions may depend on someone's mood" (Brynjolfsson, 2018).

However, Maclure (2021) argues that the argument from the limitations of human reason is flawed because it is unduly atomistic. This term means that the phenomena studied by the social sciences are grasped by focusing on individuals' thoughts, attitudes, behaviors, or other properties (Rosenberg, 2018). This approach may be justified in certain scientific studies but is not always justified.

Back to Maclure, while this methodological individualism might be warranted concerning some scientific inquiries, the result is inadequate as an approach that compares human-based and AI-based decision procedures. In contrast to this methodological individualism is the "internalism" approach.

This school of thought focuses on the idea that the facts internal to the mind/brain are sufficient to understand the mind and mental content. In this sense, and for what this thesis topic assumes is that given this “internalism” approach, the shortcomings and limitations of the average individual thinker, “often revealed by experimental studies conducted in an artificial environment such as the lab, are seen as an appropriate basis for a comparison between AI programs and human minds” (Maclure, 2021).

5.4.2 The Allegory of the Cave: dealing with two black boxes

One way to address this dichotomy between those who want to enforce laws for AI transparency and those who think it is unnecessary because the human mind is also opaque is by coming back to “The Allegory of the Cave” by the Greek philosopher Plato in his book *Republic* (Jowett, 1936).

Thus, The Allegory of the Cave may be one of the best fictional stories to metaphorically understand how humans and machines learn and explain the causes and reasons for their decisions while considering the unknown, little-studied, or poorly understood aspects of these processes. In this way, the allegory serves as a vehicle for highlighting the limitations of current natural and artificial explanations to achieve what is known as transparent AI or explainable.

Plato's theory centers on the idea that the general run of humankind can think and speak without being aware of the content of their messages. The plot of the story of the cave (Jowett, 1936) is a conversation between Socrates and his disciple Glaucon, which can be summarized as Socrates telling Glaucon to imagine people living in a vast subterranean cave, open to the outside world via only one strenuous and steep tunnel. The people are chained, facing a tall wall, unable to turn their faces or break the chains. Behind the prisoners is a great fire burning, thus projecting what happens behind the people as shapes of shadows on the wall that the prisoners are facing.

People can never turn their faces throughout their lives, so they grow up and die watching the shadows on the wall, constituting their known reality. Although the interpretation of this allegory is subject to multiple interpretations, for this research, we take that the shadows represent the superficial truth; it can even be called "virtual reality," which is perceived through our senses but does not reach the ultimate reality.

According to Plato, mortals cannot reach these primal Forms or ideas through the senses. Moreover, these shadows on the wall seem to constitute the chained prisoners' real world. As curious human beings, they tend to establish correlations between the shadows: by this way, they speculate, tell stories of what kind of shadow has to do with the other, which one appears next, and they play at predicting, causally linking the projected elements. All this is to explain the behavior of the shadows, although without knowing the nature of the shadows.

If the enslaved people could only turn their faces, they might discover that some of their favorite shadows are nothing more than tasteless bad jokes from the tricksters behind the wall. Even in a world without intentional harm, a minor random adjustment of the angle at which a three-dimensional figure would trigger a misalignment in the wall, causing disorientation among the slave theorists.

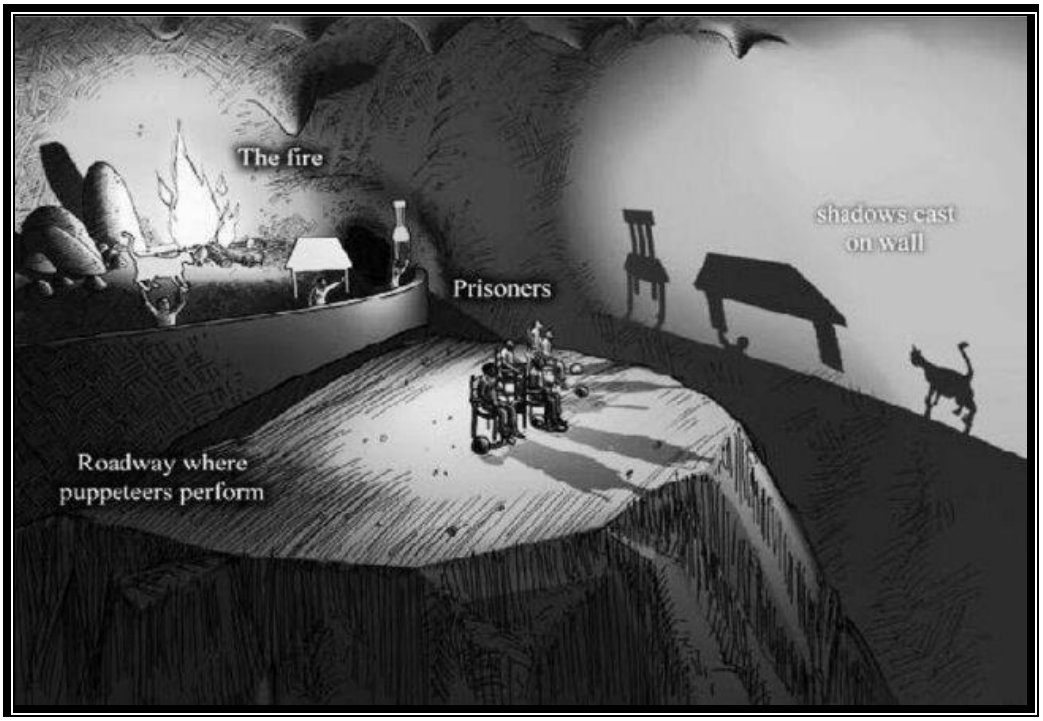


FIGURE 18 | Plato's allegory of the cave (Rachlin, 2013)

The Allegory of the Cave also explains through a metaphor how humans understand reality, explain events, and justify them, and how Machine Learning practices work in real life. Therefore,

in the same way those three-dimensional figures appear projected as dimensional shadows on the wall, the variables that are worked with in Machine Learning systems, when put into numerical or categorical data, do not correspond to the original complete object. In addition, certain features are manipulated or out of context.

According to Rudin (2019), "An explanation is a separate model that is supposed to replicate most of the behavior of a black box. Note that the term "explanation" here refers to an understanding of how a model works, as opposed to an explanation of how the world works (Rudin, 2019).

The fact remains that the challenge that remains unresolved is how we can deal with opacity in both our decision-making and that of AI systems. One form is to find an abstract way to learn whether the decision-making process reflects the underlying decision. But how can we get there?

This question brings us back to The Allegory of the Cave: these limitations in both machine reasoning and human reasoning could be that we lack the techniques to measure specific properties of reality. It is evident that if we are not aware that we lack the understanding of the existence of something, how can I ask myself how to measure it?

5.4.3 Understanding beyond the traditional forms of explanations

The limitations and challenges facing Explainable AI are diverse and varied. An urgent obstacle to resolve is that the traditional form of explainability, understood in the philosophy of science in the way that an explanation of a model or a decision to make it understandable to its stakeholders is enough, may not be valid for understanding black box models.

In this line of argument is the paperwork entitled: *The Pragmatic Turn in Explainable Artificial Intelligence (XAI)* written by the philosopher Páez, who considers that "without a previous grasp of what it means to say that an agent understands a model or a decision, the explanatory strategies will lack a well-defined goal" (Páez, 2019).

Thus, many philosophers argue that understanding and explanations are somehow like the heads and tails of a coin. This equivalence has a long tradition in areas like the philosophy of science with authors such as Salmon, who defend the following: "Understanding results from our ability to fashion scientific explanations" (Salmon, 1984).

However, the most substantial claim in Paéz's (2019) research comes from Khalifa, who supports the thesis that "any philosophically relevant ideas about scientific understanding can be captured by philosophical ideas about the epistemology of scientific explanation without loss" (Khalifa, 2012).

This way of understanding Explainable AI would imply that understanding an AI model or its final decision would be a simple matter of finding a suitable explanation.

This thesis could only be sustained if scientific explanations and those provided by AI systems share, according to Paéz, "a sufficient number of essential characteristics to be considered two species of the same genus" (Páez, 2019).

So far, there is no scientific validity to support this assertion because a fundamental difference between the explanation of a logic-based model and a data-driven model is that the *explanandum* is entirely different. Therefore, Paéz (2019) presents three fundamental reasons why making our understanding of Machine Learning models dependent on establishing an account of AI explanations is misguided.

1. *The first reason has to do with the concept of truth*

A characteristic of scientific explanations is their factivity (Hu, 2021), i.e., both the *explanandum* and the *explanans* must be true. In this way, an explanation of *x* must reveal the true causal structure of *x* or the natural laws that determine *x* or its relationships with factors that make *x* more or less probable.

This kind of epistemic access is lacking in the case of opaque AI systems. Many black box models, like deep neural networks, are stochastic or non-deterministic. Randomness is introduced in data selection, training, and evaluation to help the learning algorithm be more robust and accurate.

As Paéz points out: “examining the training data set and all the weights, biases and structure of the network will not allow us to understand its specific decisions, and its predictive failures and successes cannot be traced back to particular causal paths in its hidden layers” (Páez, 2019).

One of the positive aspects that Paéz (2019) sees in replacing “explanation by understanding” is that the factivity condition “need not be satisfied.” In the same vein is the thesis proposed by George E.P. Box in his famous quote, “All models are wrong, but some are useful” (Box, 1976), and the idea of “felicitous falsehoods” terminology introduced by Elgin. For this author, the role of models and idealizations allows that our understanding of some aspects of reality may be false. However, the point is that far from being an unfortunate expedient, idealizations and models are an essential and ineliminable component of our scientific understanding of the world (Elgin, 2007).

These arguments, plus the one offered by Paéz, that even if the methods and artifacts currently used to understand these opaque models and their decisions are not entirely faithful to the model, this does not count against them.

In the words of Paéz, “On the contrary, they can afford indirect epistemic access to matters of fact that are otherwise humanly impossible to discern” (Páez, 2019).

2. *The second reason concerns the importance of considering the specific context, background knowledge, and interests of end-users and stakeholders of opaque models*

This argument concerns an element that Paéz (2019) describes as AI researchers building explanatory models for themselves rather than for the intended users. This phenomenon resembles what Miller et al. (2017) call “the inmates running the asylum.” An alternative way to fix this, which I fully agree with, is the emerging need to explore a naturalistic approach to how context and background knowledge mold an agent’s understanding of an interpretative model.

3. *The third reason for this shift is a better understanding of the relationship between explanation and trust*

The relationship between explainability and trust seems to be created based on a correlation that is not always true. According to Paéz (2019), this is the same as saying that the pairing “predictive reliability and a post hoc explanation are not sufficient to generate trust.”

Trust does not depend exclusively on epistemic factors; it also depends on the stakeholders' interests, goals, resources, and degree of risk aversion.

Thus, to understand the multidimensional nature of trust, it is necessary to consider the factors surrounding it to check what drives and hinders it.

Based on these arguments, Paéz (2019) offers alternative paths to understanding the workings and decisions of opaque models. Thus, one of the most supported methods to understand the why is to know the causes; this method is supported by many philosophers (Salmon, 1984; Lewis, 1986; Greco, 2010; Grimm, 2006). However, the source of causal knowledge is not limited to causal explanations. Lipton (2009) states, "Much of empirical inquiry consists in physical and intellectual activities that generate causal information, such as observation, experimentations, manipulation, and inference. And these activities are distinct from giving and receiving explanations". In other words, this line of thinking opens the range of possible ways to make tacit causal knowledge from images and physical models.

Another way of obtaining causal knowledge, which we are working on within this research, is the possibility of manipulating a causal system. In this way, and according to the authors de Regt and Dieks (2005) and Wilkenfeld (2013), manipulation can provide model information about a system's possible states. The ability to manipulate a system into new desired states could be a sign of understanding, and for Regt and Dieks (2005), this kind of understanding requires "the ability to think counterfactually."

Another way to reach an understanding without explanation through the use of non-propositional representation is diagrams, graphs, and maps. The example provided by Paéz is representative enough: "A subway map is never a faithful representation of the real train network. It alters the distance between stations and the exact location of the tunnels to make the network easy to understand, but it must include the correct number of lines, stations, and intersections to be useful. It must be sufficiently accurate without being too accurate" (Páez, 2019).

The final form that Paéz (2019) mentions to achieve understanding without explanations is models and idealizations. These methods make it possible to access characteristics of the objects that are otherwise difficult or impossible to discern.

Perhaps the most critical part of this method, according to Elgin, is that "they are representations of the things that they denote (…) and obviously, whether such a representation is accurate enough is a contextual question. A representation that is true enough for some purposes, or in some respects is not accurate enough for or in others" (Elgin, 2007).

These methods, the ones explained immediately above, serve to achieve a more objective understanding of causal relationships under the term of scientific explanations. Other authors, such as Mittelstadt et al. (2019), consider that everyday explanations fit better with the goals of explainable AI. In this research, we support Paez's (2019) argument, which considers that "every day" explanations provide a purely subjective sense of explanation and XAI should adopt other methods of explanations like those exposed above to provide objective understanding.

And the fact is that one of the dangers of focusing on every day or textual explanations is what the authors Ehsan et al. call rationalizations described as follows:

> "AI rationalization is based on the observation that there are times when humans may not have full conscious access to reasons for their behaviour and consequently may not give explanations that reveal how a decision was made. In these situations, humans are more likely to create plausible explanations on the spot when pressed. However, we accept human-generated rationalizations as providing some lay insight into the mind of others" (Ehsan et al., 2017).

The search for understanding in AI pursues a twofold function. On the one hand, it is about understanding why, or, in other words, a system behaves in one way and not in another.

On the other hand, we have to search for transparency by understanding the model's inner workings. However, for Paéz, this dual function would be reduced to one if a better objectual understanding of a model could be achieved.

With the following example proposed by Paéz (2019), one can understand the problem better: Let's imagine that a person knows that smoking can cause lung cancer, among other causes, but this does not mean that this person has stopped to wonder what kind of causal mechanisms are involved in the process. In addition, the layperson lacks “the ability to answer a wide range of questions of the type what-if-things-had-been-different (Woodward, 2003).

For this reason, we support the thesis of Paéz (2019), who argues that the ability to answer counterfactual questions and make predictions depends to a more significant extent on an objectual understanding of the larger body of knowledge to which the specific object of understanding belongs. Back to the example above: without a basic understanding of the body’s anatomy, biology, chemistry, and the behavior of the lungs, a layperson could not answer these kinds of counterfactual questions, which is a sign of understanding.

Regarding Explainable AI and proper counterfactual reasoning, we must say that this exercise is purely theoretical, “based on knowledge about how the model works” (Páez, 2019).

In addition, he states that if we take the ability to think counterfactually about a phenomenon as a sign that the agent understands it, understanding the decisions of a model requires some degree of objectual understanding.

But as we have been arguing in this section, understanding why is not always so simple, and it seems that this position can be defended “in simple scenarios where a complete analysis of the relevant causal variables can be provided, but as soon as the context requires the use of theoretical tools such as idealizations and models, it becomes highly doubtful” (Páez, 2019).

This way is how Machine Learning models work. The inner working of these systems makes it almost impossible to akin the causal knowledge necessary to provide a true causal explanation.

Therefore, we can conclude that this type of intelligent system could not fulfill the condition of factuality to understand why. For this reason, Paéz’s (2019) argument focuses on the that understanding-why and objectual understanding in Machine Learning cannot be entirely independent of each other, but if one takes the claim that understanding why is not always factive in the end, there is no essential difference between them in terms of truth.

However, if one has to cite one of the most established ways of understanding how Machine Learning systems make their decisions in Explainable AI, we must refer to the functionalist approach, perhaps the most widely used methodology. This approach, which comes from psychology, distinguishes between two ways of understanding an event: the functional and the mechanical. With this example proposed by Paéz (2019), we can distinguish one approach from another. An alarm clock beeps because the circuit connecting the buzzer to a power source has been completed (mechanical understanding), and its owner has set it to wake them up at a specific time (functional understanding).

Lombrozo and Wilkenfeld (2019) state that "a subject can have a functional understanding of an event while insensitive to mechanistic information." However, reliability alone cannot usher trust because of the dataset shift problem. To develop what is known as trust, it is necessary to have patterns valid in past data and future cases because "unfortunately, most solutions to the dataset shift problem focus only on accuracy, ignoring model comprehensibility issues" (Freitas, 2014). In addition, focusing only on the functional approach to explanations has the unintended effect that the models are also more likely to be tailored to user preferences and expectations and thus prone to oversimplification and bias (Páez, 2019).

A good example is the simple rule governing many recommender systems: "If you like x, you might like y." Another aspect to be taken into account is that although the understanding and trust sought by Explainable AI should always consider a model's stakeholders, it should not pursue these goals by offering misleadingly simple functional explanations that can result in unjustified or dangerous actions (Gilpin et al., 2018). While functional explanations seem the most plausible in Explainable AI, these shortcuts may bring unintended consequences. It is, therefore, appropriate to add the three elements that, according to Paéz (2019), form an adequate understanding of a model: 1) obtaining the right fit between the interpretative model and the black box model in terms of accuracy and reliability; 2) providing sufficient information about its limitations; 3) achieving an acceptable degree of comprehensibility for the intended user. However, the limitations of the explanations provided by the black boxes do not end here. To achieve what Paéz (2019) has called an "objectual understanding of a model," other techniques are required to generate explanations.

For this purpose, it is convenient to introduce a new subfield of Explainable AI called Goal-driven XAI. This research domain is briefly introduced in the next section but plays a determining role in the following chapter, which opens the second phase of this dissertation, which is the antithesis phase from the methodological point of view.

5.4.4 Goal-driven Explainable AI envisions

One future avenue for advancing the field of Explainable AI is what the authors Bruijn et al. (2022) call "the construction of a symbolic, human-understandable model automatically from the non-symbolic, statistical machine-learned model."

Although many AI experts often see neural networks and symbolic systems as opposites, the differences can be more subtle, and we argue that one system can enrich the other and vice versa because "both levels could, therefore, bridge low-level information processing such as frequently encountered in perception and pattern recognition with reasoning and explanation on a higher, more cognitive level of abstraction (Besold et al., 2017).

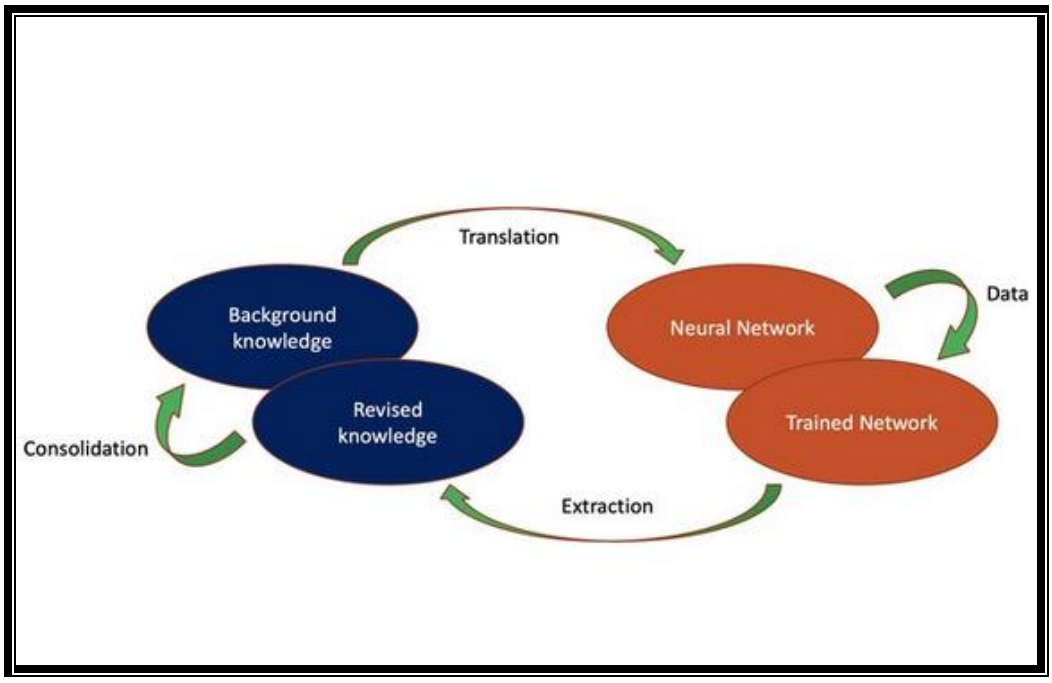


FIGURE 19 | An example of a neural-symbolic cycle

This higher cognitive level of abstraction would be equivalent to the objectives of this doctoral thesis, which is to achieve a more objective knowledge of certain features of human nature, which, in our case, are emotions.

Therefore, the study of emotions pursues these two objectives: on the one hand, to deepen their understanding of the human deliberative process, and on the other hand, to deepen how emotions intervene in the formation of biases and noise as primary sources of human error.

Thus, the research domain called goal-driven XAI seeks to create robots or agents that can justify their behaviors to a lay user (Anjomshoae et al., 2019). These explanations differ from those expected from the ones provided by data-driven XAI techniques.

The fundamental difference is that in goal-driven XAI, the explanations would assist the user in creating a Theory of Mind (ToM). According to Anjomshoae et al. (2019), by having a better theory of mind of the agent, the user would be more able to understand the agent's limitations and capabilities, thus enhancing confidence and safety levels and preventing failures.

The absence of adequate mental models and understanding of the agent can contribute to a failure of interactions (Anjomshoae et al., 2019; Chandrasekaran et al., 2017).

It is essential to highlight how this ToM needs to be designed to achieve the purpose of this dissertation. It contains this set of features that differ from the current literature on the topic. Hence, we aim to create a theory of mind about the role of emotions in human behavior, focused on detecting how bias and noise are expressed through patterns of behavior and rules of reasoning. The aim of this type of explanatory technique is not to satisfy the user with a good explanation.

However, this explanatory component of the agent seeks to explain and show the user the cognitive and affective failures that need to be reinforced to achieve a specific ethical principle, in our case, the principle of responsibility in a human-AI interaction. This principle of responsibility is part of two different ethical frameworks presented at the beginning of this chapter, but in this dissertation, they are combined in our proposed framework in the final chapter.

In addition, in this research domain, to generate goal-driven explanatory techniques, a theory of mind is as imperative as continuous learning to avoid phenomena such as emergent biases or other related sources of error. In this direction, the following techniques that support the tech-

nical design of the framework proposed in Chapter 6 are described and analyzed: symbolic representations and reinforcement learning.

However, before moving on to Chapter 6, we must justify why emotions are chosen to create a theory of mind in our goal-driven XAI. We must first analyze whether emotions can be considered measurable data, if so, what kind of techniques exist, what kind of emotions play a pivotal role in deliberative processes, to what extent these technologies to measure emotions are still controversial in the scientific literature and how to bridge these knowledge gaps with the technology available so far, if possible.

All these questions motivated Chapter 5 of this research, which corresponds to the second phase of this doctoral thesis, called antithesis in methodology.

Nevertheless, it is remarkable to say that the explanations, at least to some extent, of any kind, are open to interpretation, and as Bruijn et al. (2022) presented in their paper: “the more complex the situation, the more challenging it is to explain the results,” especially when these explanations attempt to explain aspects of reality hitherto disregarded because they were not aware of their existence or for not knowing how to systematize this knowledge more rigorously or scientifically.

CHAPTER 6

EMOTIONS: THE DATA PARADOX

6.1 A brief history of emotions

Before neuroscientific studies focused on studying emotions at the beginning of the 19th century, emotions were only the "inner beasts to regulate" through our rational brain (Feldman Barret, 2020). However, this is not to say that no philosophers or thinkers thought emotions primarily drive decision-making. British thinkers of the 18th and 19th centuries were some of the greatest exponents of this approach because they thought that emotions played a vital role in how humans made decisions. Bentham writes: "Nature has placed mankind under the governance of two sovereign masters, pain, and pleasure… They govern us in all we do, in all we say, in all we think" (Bentham, 1823). Hume was equally evident in his claims. For him, "reason is… the slave of the passions and can never pretend to any other office than to serve and obey them" (Hume, 1739)

Fortunately, in the second half of the 20th century, most neuroscientists and psychologists moved away from struggling with how reason and emotion interact to govern human behavior and to investigate the role that emotions play in reasoning and decision-making (Bechara et al., 2000). This shift is fundamental, and if there is one notable figure in this work, albeit with its critics, it is the work done by Damasio (Damasio, 1994). This author claims that no conflict between reason and emotion is necessary. Indeed, Damasio developed this somatic marker hypothesis (SMH) to describe this phenomenon. The central assumption of this hypothesis is that people often do not choose based on intellectual analysis alone but also based on emotions elicited as part of the decision-making processes. Although the somatic marker hypothesis has been a real turning point in how humans make decisions, it is also essential to consider criticisms of this hypothesis.

However, despite the undeniable evidence of emotions' role in decision-making, emotional issues in scientific practice remain an enigma for Rosalind Picard, the pioneer of affective computing (1995). She states, "Scientific principles are derived from rational thought, logical arguments, testable hypotheses, and repeatable experiments" (R. W. Picard, 1995).

However, as we see later in this chapter, emotions may play a very different role in understanding the phenomena of intelligence, decision-making, and the implementation of artificial emotional intelligence or affective engineering in Asian countries such as Japan.

From the triune brain of Plato to how to create a mind of Kurzweil

This way of proceeding in science has its roots in ancient Greece. Two thousand years ago, the philosopher Plato wrote that human minds are a never-ending battle between three inner forces to control your behavior. Feldman Barret (2020) states that Plato's compelling morality tale of inner conflict remains one of the most cherished narratives in Western civilization[7].

This view is highly relevant to understanding how this Platonic view has permeated how science and the human mind have been perceived in the West (even today). In her book *Seven and a Half Lessons About the brain,* Feldman Barret (2020) describes Plato's theory as "an attempt to explain how the human brain evolved." Three hundred million years ago, we were lizards. This reptilian brain works for basic urges like feeding, fighting, and mating. About one hundred million years later, the brain evolved a new part that gave us emotions; then, we were mammals. Finally, the brain evolved a rational part to regulate our inner beasts" (Feldman Barret, 2020).

We became human and lived logically ever after. Feldman Barret (2020) continues to explain that this evolutionary story ended up with three layers: one for surviving, one for feeling, and one for thinking. These three layers are known as the triune brain. The deepest layer, or lizard brain, is said to house our survival instincts. The middle layer, dubbed the limbic system, contains ancient parts for emotion that we inherited from prehistoric mammals.

The outermost layer, part of the cerebral cortex, is considered uniquely human and the source of rational thought known as the neocortex. One part of your neocortex, the prefrontal cortex, regulates your emotional and lizard brains to keep your irrational, animalistic self in check.

This triune brain story that began with Plato has been a source of inspiration for rationalists in moral philosophy and for psychologists for most of the twentieth century. In addition, the influence of Platonic thought is staying today, and one representative example of that is the current

[7] This point is especially important because in Asia, for example in Japan, the way of conceiving what in the West is called affective computing is very different, which has a big impact on the way we understand the human being.

director of engineering at Google, Raymond Kurzweil, who published in 2012 the book *How to Create a Mind: The Secret of human thought revealed*. To him, it is possible to create a human brain through reverse engineering. It relies on "replicating" the three layers explained above to achieve this (Kurzweil, 2012). It is not the purpose of this research to assess whether this is possible, but this approach, although very controversial, still serves as a model for many purposes, especially essentialist[8] ones.

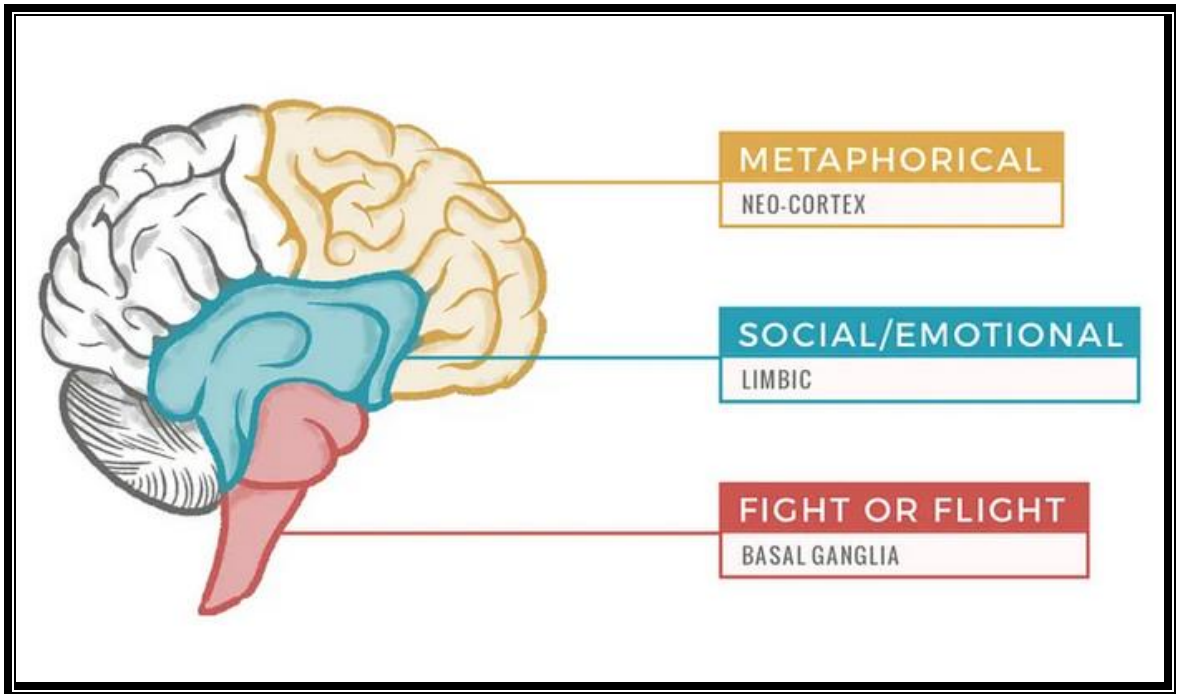


FIGURE 20 | Triune brain concept (Butler, 2009) / medium.com

[8] As the concept of essentialism will be continuously referred to in this chapter, it is convenient to define it as an approach assuming that people and things have natural and essential common characteristics which are inherent, innate, and unchanging.

6.1.1 Traces of dualism in emotion theories

The neuroscientific study of emotions began in the 1900s. This beginning coincides with "when psychology was transformed from mental philosophy into a full-fledged scientific discipline" (Feldman Barrett & Satpute, 2017). In this phase of the study of emotions, scientists conducted their experiments inspired by a Linnaean-type taxonomy of emotion categories that, according to Barret and Satpute (2017), aim to discover a small set of which each was cast as independent mental organs with a unique neural circuit and pattern of physiological correlates (i.e., a "physical fingerprint" (Feldman Barrett, 2017); the term will be discussed in depth later.

This family of approaches (Gross & Feldman Barrett, 2011) refers to the classical view of emotions. Therefore, authors like Cannon (1915) formed grand theories as they battled over whether emotions were biological changes in the body or the brain[9].

Almost at the turn of the 20th century, the experiments that tried to establish a "biobehavioral basis of each emotion category were unable to identify any specific or ensemble of autonomic nervous system changes (or facial movements, for that matter) that consistently distinguished the instances of one emotion category from another" (Gendron & Feldman Barrett, 2009).

For example, according to Feldman Barret (2017), "the physical pattern for anger that distinguished it from sadness in one study was different from that observed in other studies. Variation was the norm (a topic discussed in more detail below). In this line, scientists reformulated how they studied emotions and turned to the functionalist approach. This perspective of understanding emotions is based solely on recasting emotions as functional states to be analyzed by manipulating their observable causes and measuring their observable effects.

In the first half of the 20th century, due to the success of functionalism, the mind disappeared as an object of study, and neuroscience focused on biological behavior. This shift meant that emotions were no longer considered a mental phenomenon but that what mattered were states registered in the nervous system that produced specific behavior. The examples of Feldman Barret and Satpute illustrate this well: "The neural circuity for freezing or fleeing became the circuity for fear; the neural circuity for fighting became the circuity for anger. And so on.

[9] The ancient philosophical problem of mind and body and their relationship in Western culture.

Like all psychological categories, emotion categories were ontologically reduced to behaviors" (Feldman Barrett & Satpute, 2017).

In the 1950s and 1960s, psychology emerged from behaviorism, and the mind returned as an object to study. However, this time, the interest centered on the metaphor "brains as computers," which began as an apparent metaphor and became what is now known as AI. Once again, the so-called emotional categories are ontologically reduced to behavior in the name of evolution.

That means: "non-human animals clearly have emotions, the argument went, but they do not necessarily have emotional experiences. Some scientists wanted a species-general explanation of emotion to generalize their findings from rats to humans" (Ekman et al., 1972; Feldman Barrett & Satpute, 2017).

In the early 1990s, the cognitive revolution gave rise to two sub-branches: cognitive neuroscience and affective neuroscience. According to Lindquist and co, "faculty psychology and its taxonomy of mental categories still reigned, and the goal (once again in history) was to carve the brain into a set of independently functioning mental organs: one set of neurons dedicated to each mental category" (Lindquist et al., 2012).

However, although this essentialism is still present and cannot be denied or stated categorically to what percentage of this approach is true or false, in the 1990s, another event directly related to this research took place. In this case, it is about involving emotions in decision-making, thanks to the example of the work done by Damasio (1994), as discussed above, among others that I cite in the following sections of this chapter.

Although much remains, the knowledge revolution has reached the emotions. It has already started to become an essential area of science in the West because of the various techno-scientific advances that made it possible to determine their essential role in decision-making, as shown in the following sections of this chapter. Therefore, this thesis will not judge which theory of emotions is the right one but will examine how a combination of the studied authors gives clues to understanding the emotion-reason-ethics axis and how it can be amplified in a human-AI interaction.

6.1.2 Emotions and reasons: who decides what?

In this dissertation, we do not discuss all emotions but focus specifically on regret because of its essential role in emotion regulation, as discussed below. In addition, this research proposes a different treatment of emotions and a more rigorous, methodological, and scientific approach to their study, which is different from what has been done so far.

Thanks to progress in social cognition, neuroanatomy, and psychophysiology, a starting point for endowing emotions with another meaning has been made. These disciplines reconsider that affect is not necessarily disruptive to human behavior.

Indeed, researchers have concluded with their studies that, e.g., affect is often a valuable and even essential component of adaptive social behavior (Bechara et al., 2000) and plays a beneficial role in how social information is perceived and categorized. In this line, Haidt's work in the emerging field of affective sciences, which connects to Hume, constructs a model of "social intuitionism" (Haidt, 2001). For Haidt, ethical judgment is much like aesthetic judgment: "It is made quickly, effortlessly, and without a great deal of conscious deliberation" (J. Greene & Haidt, 2002).

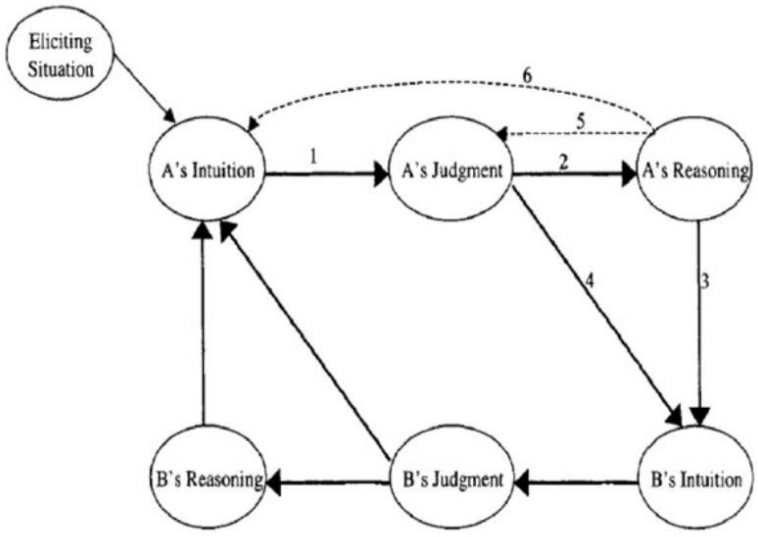


FIGURE 21 | The social intuitionist model (SIM) (Haidt, 2001)

He also claims “that if deliberate reasoning played a role in moral judgment, it was most likely as a post hoc attempt to justify an individual’s intuition-driven moral judgments” (Haidt, 2001; Haidt & Bjorklund, 2008). Greene and his collaborators have reached conclusions similar to those of Haidt (2001, 2002; 2008) using modern neuroimaging techniques. The former aims to demonstrate that “when faced with classic sacrificial moral dilemmas (in which one individual must be sacrificed to save a greater number of people), individuals often responded in a manner consistent with their gut, emotional reactions” (J. D. Greene et al., 2001), and they also argued that this could be seen in the patterns of neural activation registered while individuals were reading emotionally evocative dilemmas.

May and Kumar (2018) posed the following questions that fit this theme: “What drives your judgment? Have you reasoned your way to the conclusion that something is morally wrong? Or have you reached a verdict because you feel indignation or outrage?”

FIGURE 22 | What drives your judgment? / shutterstock.com

In this same direction and more than 350 years ago, when the rationalist theories about the uselessness of emotions in reasoning were the strongest, the philosopher Pascal (1643) dared to say: “The heart has its reasons which reason does not understand” (Forgas, 2006).

Before proceeding any further in this chapter, it is once again worth stressing that the importance of this research does not focus on criticizing or instead supporting the classical theory of emotions or Feldman Barrett's theory of constructed emotions. This work is far from the intentions of this thesis.

The aim is to precisely bridge the gap between these conflicting visions that prevent a dialogue that would allow further progress in finding solutions to why humans do not learn from their own mistakes, which they repeat over and over (again). In other words, it is to understand a little more about why we decide, how we decide, and whether there are ways to improve this challenge. That is what this research is about.

For this purpose, all authors and their respective theories and limitations are crucial to consider, as further progress would not be possible without them. In addition, the differences between the concepts of intelligence in the West and the East are another topic covered in this chapter, as these terms shape how the so-called artificial emotional intelligence or affective computing is implemented.

6.2 Emotions become an essential part of rational human performance

So, what kind of behavior is considered rational? In the rationalist theories discussed above, everything to do with the emotional world is considered irrational. However, can an emotion not be considered a rational element? According to the neuroscientist Feldman Barret (2020), yes, when you feel afraid because you are in danger. On the other hand, in the same way, thinking is considered irrational, like when you scroll through social media for hours, telling yourself you are bound to come across something important.

In other contexts, detecting whether a behavior is rational or irrational is not easy to notice. Take the example of a soldier in a war zone. In this situation, the brain is prepared to predict threats, but because the brain is not the perfect machine, it could infer danger by raising your levels of cortisol when there is no danger at all. "But in a war zone, this false alarm may be rational from a body-budgeting standpoint" (Feldman Barret, 2020).

Therefore, it seems, and in Barret's words, that rational behavior is based on making a good body-budgeting investment in a given situation. Let's delve deeper into the emotional-rational elements in decision-making based on Damasio's (1994) hypothesis of the somatic marker, with some criticisms of it at the end of this section.

6.2.1 High-reason versus the somatic-marker hypothesis

To understand the interplay between the misunderstood binomial between reason and emotion in decision-making, Damasio begins by describing how reasoning and decision processes in the rational tradition were considered synonymous. According to him:

> "The terms reasoning and deciding usually imply that the decider knows (a) about the situation which calls for a decision, (b) about different options of action (responses), and (c) about consequences of each of those options (outcomes) immediately and at future epochs. Knowledge, which exists in memory under dispositional representation form, can be made accessible to consciousness in both nonlanguage and language versions, virtually simultaneously" (Damasio, 1994).

High-reason versus somatic-marker hypothesis

Damasio (1994) thus considers two possibilities when it comes to choosing between many possible scenarios: the first one of these has to do with the classical view of decision-making from a traditional "high-reason"; the second one comes from the "somatic-marker hypothesis."

As discussed at the beginning of this section, the classical view of decision-making has to do with the commonsense view, which assumes that formal logic will get us to the best available solution for any problem. One aspect is that emotion must be kept out for the best results. Damasio (1994) described this view as when you break the different scenarios apart and use current managerial parlance. You perform a cost/benefit analysis of each of them. Considering "subjective expected utility," which you want to maximize, you infer logically what is good and wrong. For instance, the way to proceed is to consider the consequences of each option at different points in the projected future and weigh the ensuing losses and gains. However, most scenarios have far more than two alternatives in our cartoon, and the perfect analysis is anything but easy through your deductions.

Using Damasio (1994) as an example is a way to understand this phenomenon better. Hence, gaining a new client may bring immediate rewards and also a substantial number of future rewards. How much reward is unknown and turns into a complex calculation. A substantial part of this calculation depends on the continued generation of more imaginary scenarios, built on visual and auditory patterns, among others, and the continued generation of verbal narratives accompanying those scenarios, which are essential to keep the logical inference going.

According to Damasio (1994), if this strategy is the only one you have available, rationality, as described above, will not work. Because it is not easy to remember the many ledgers of losses and gains that you need to consult for your comparisons, perhaps the most compelling point of this study by Damasio (1994) is that thanks to new neuroscientific techniques, it was possible to study and compare how patients who had prefrontal damage in their brains took decisions. This investigation led Damasio to articulate his best-known work, "The somatic-marker hypothesis." His thesis has marked a turning point in understanding human behavior and even the issue of human biases and noise as the source of human errors in judgments (approach commented on in this dissertation, in Chapter 4).

Somatic markers as body alarms to sift out multiple future outcomes

The way to tackle the somatic marker hypothesis, according to Damasio (1994), is to follow these steps: imagine before someone applies any cost/benefit analysis to the premises and before someone reasons toward the solution of a problem, something quite important happens: When a poor outcome connected with a given response option comes into mind, however fleetingly, one experiences an unpleasant gut feeling. Because the feeling is about the body, he gave the phenomenon the technical term somatic state ("soma" is Greek for the body), and because it "marks" an image, he called it a marker.

The somatic marker is an automatic alarm: "Beware of danger ahead if you choose the option which leads to this outcome" (Damasio, 1994).

Besides this, this kind of signal may immediately lead someone to reject the negative course of action, making them choose other alternatives. According to Damasio, this automatic signal protects you against future losses and allows someone to choose from fewer alternatives.

It is decisive to highlight that there is still room for using cost/benefit analysis and proper deductive competence, but only after the automated step does the number of options drastically reduce. In addition, somatic markers may not be sufficient for human decision-making since a subsequent reasoning and final selection process still occurs in many instances. The exact definition of a somatic marker, as proposed by Damasio, is:

> "They are a special instance of feelings generated from secondary emotions. Those emotions and feelings have been connected, by learning, to predicted future outcomes of certain scenarios" (…). Somatic markers do not deliberate for us. They deliberately highlight some options and eliminate them rapidly from subsequent consideration" (Damasio, 1994).

Another aspect that Damasio emphasizes in his somatic marker hypothesis is that effective personal and social behavior is required to form an adequate "understanding" of their minds and the minds of others.

Again, the idea of somatic markers is to assist the process of sifting through a wealth of possible scenarios, reducing the need for sifting because they provide automated detection of the scenario components that are more likely to be relevant.

6.2.2 Somatic markers: nature versus nurture

But then, where do these "automatic devices" that help us in the deliberative processes of our decision-making come from? Damasio's hypothesis states:

> "We were born with the neural machinery required to generate somatic states in response to certain classes of stimuli, the machinery of primary emotions" (…). Nevertheless, most somatic markers we use for rational decision-making were created in our brains during education and socialization by connecting specific classes of stimuli with specific classes of somatic state. In other words, they are based on the process of secondary emotions" (Damasio, 1994).

He mentions that one aspect that may be one of the keys to a much better understanding of human behavior is that these somatic markers can only aid decision-making when both brain and culture are normal.

This research tries to understand better what normal means, or rather, something more complicated scientifically: How do we measure normal brains and cultures?

In any case, and although the concept of normal in Damasio's work (1994) does not cover all these questions, it is relevant to present his thoughts because they could provide clues even if they are not definitive due to the complex nature of the object of study that concerns us.

Thus, in Damasio's (1994) theory, the moment the brain or culture is defective, the work of somatic markers fails to fulfill its adaptive function. He cites the condition whereby certain people he calls patients develop sociopathy or psychopathy.

Given that the subject of psychopathy as such cannot be studied in detail here, it is worth noting that Damasio says that this decrease or absence of feelings may be due to, on the one hand, the development of the mental illnesses mentioned above because of abnormal circuitry and abnormal chemical signaling and begin early in development. Likewise, on the other hand, patients with brain damage can have a similar problem, with the lack of feeling being the origin or the cause of this abnormal behavior.

Therefore, understanding more about the differences and similarities between these cases could contribute to clarifying the interplay between social and biological factors and "even shed light on conditions which may be superficially similar and yet be largely determined by sociocultural factors" (Damasio, 1994).

About the effect that a "sick culture" may have on the reasoning system of an average adult, Damasio (1994) considers it to be less dramatic than the cases discussed above. Nevertheless, examples of "sick cultures" for this neuroscientist were German and Russian society in the 1930s and 1940s, China during the cultural revolution, and Cambodia during the Pol Pot regime, to cite only the most recurrent cases, but there are others. (We currently witness counterexamples of sick cultures in how Damasio interprets them, but we will return to this issue later).

Somatic markers as context-dependent

But then, how are these somatic markers acquired, according to Damasio? We acquire somatic markers by experience. On the one hand, the internal preference system regulates the organism for its survival. Achieving survival is the same as saying that the body tries every time to avoid pain, seek potential pleasure, and attain a homeostatic state, and is pre-tuned for achieving these goals in social situations.

On the other hand, there is a set of external circumstances with which the organism must interact, possible options for action, possible future outcomes for those actions, and the punishment or reward accompanying a particular option.

Damasio (1994) points out that in moral development, punishment and reward are delivered not only by the entities themselves but also by parents and other elders and peers, who usually embody the social conventions and ethics of the culture to which the organism belongs.

Therefore, the interaction between an internal preference system and sets of external circumstances extends the repertory of stimuli that become automatically marked. To conclude this argumentative section, it is important to remark on the following words of Damasio's hypothesis to develop the framework and how goal-driven XAI should be designed.

He says: "The critical, formative set of stimuli to somatic pairing is, no doubt, acquired in childhood and adolescence. But the accrual of somatically marked stimuli ceases only when life ceases, and thus it is appropriate to describe that accrual as a process of continuous learning" (Damasio, 1994). Although for Feldman Barret (2017), Damasio's (1994) somatic marker hypothesis is part of what she criticizes for considering that his view of emotions belongs to the classical theory because every emotion has a specific physical fingerprint (changes in heart rate, breathing, hormones, muscle tone, and facial expression, among others.) The somatic marker hypothesis was a significant milestone in understanding the relationship between what is natural and what we can learn. Considering that what we learn would vary or modify how what is natural is understood, However, before moving on to the next section of this chapter with AI-related themes, it is worth considering some of the most common objections to Damasio's somatic marker hypothesis beyond those made by Feldman Barret (2017).

6.2.3 The controversies raised by the somatic marker hypothesis

For some authors (Overskeid, 2021), Damasio's somatic marker hypothesis is a shy attempt to put emotions where they belong:

> "Somatic markers may not be sufficient for normal human decision-making since a subsequent process of reasoning and final selection will still take place in many though not all instances (…) that means logical competence does come into play beyond somatic markers." On the other hand, Damasio (1994) argues that "some sublime human achievements come from rejecting what biology or culture propels individuals to do" – thought "freedom from biological and cultural constraints can also be a hallmark of madness and can nourish the ideas and acts of the insane." But then, "can the SMH (Somatic Marker Hypothesis) then rigorously explain the important underlying aspects of human decision-making?" (Overskeid, 2021).

One of the problems is that Verweij and Damasio said that "affect does not necessarily dictate human choice without referring to another mechanism that does" (Verweij & Damasio, 2019).

However, Overskeid strongly argues that "there is much evidence that reasoning is governed by emotions, even when people end up choosing aversive options" (Overskeid, 2000), – "often because uncertainty feels even worse than distressing certitude" (Gilbert, 2009).

In the same vein, Stanovich emphasizes that people are cognitive misers, and to start reasoning, we need to experience some difficulty in reaching a goal (Stanovich, 2018). Therefore, according to Overskeid (2021), a goal is a state we want to reach. Hence, all generally accepted definitions of what motivates one to solve a problem include a term such as "desire" or "want" (Cobb & Mayer, 2000). In other words, wanting to be in another state than the one we are currently in is a goal state.

In essence, what Overskied is arguing is that rather than a shy somatic marker hypothesis as offered by Damasio (1994), since it does not concern the reasoning steps that follow the action of the somatic marker," what would be perhaps logical is a strong version of the SMH, in the sense that the somatic marker keeps acting until the final choice is made (Overskeid, 2021).

6.3 Can computers understand emotions?

While it may be true that the somatic marker hypothesis may have been a timid attempt to place emotions in decision-making processes, it should not be downplayed. Although emotions still do not merit the attention they require in politics, ethics, different branches of medicine, and education, to name a few fields, it has been somewhat relevant in fields such as affective computing (R. W. Picard, 1995) and neuroeconomics (Platt & Glimcher, 1999).

Of these last two fields, this section examines affective computing in-depth as it is directly related to the object of study of this thesis. Specifically, this section analyses Rosalind Picard's work (1995) and compares this Western vision with how artificial emotional intelligence develops in the East.

This analysis is accompanied by neuroscientist Lisa Feldman Barret's critique of the classical theory of emotions as implemented in current AI models.

So, following Barret's hypothesis, or in other words, her critique of traditional emotion theory, if the field of affective computing continues to use universal correlations between affective states and emotions systematically, these AI systems could be used to deepen social inequalities and further entrenching prejudices and stereotypes with unforeseen consequences in our history under the guise of being scientifically measured and therefore justified.

On the other hand, many other experiments that Picard and his team have conducted are leading to a much better understanding of the role that emotions play in decision-making and other issues related to a more unified theory of them.

In other words, although aspects are subject to criticism, others could contribute to generating a more systematizing study of emotions and their relationships with moral aspects in decision-making strategies in human-AI interaction in well-defined contexts. The next and final chapter analyzes and justifies this issue in depth and detail.

6.3.1 Making “data emotions” matters

When Picard (1995) wrote her manifesto on affective computing, computers had begun to acquire the ability to recognize affect. However, the computers' capabilities could go further for her and have emotions.

In addition, the research of Damasio (1994) was of decisive importance for Picard, especially his hypothesis of the somatic marker, making even more evident the importance of including emotions in the advances of human cognition. For Picard, the issue was to test whether computers and other devices could help to follow tasks, such as detecting human emotions or gathering new data necessary for advances in emotion and cognition theory.

Since then, Picard’s (1995) work in affective computing has ranged from proposing models to recognize affective states to computer-assisted learning, perceptual information retrieval, arts and entertainment, and human health and interaction. She tries to argue in very different ways because emotions in human cognition are not a luxury; instead, thinking and feeling are heads and tails of the same coin.

To justify this, she gives several examples, such as commenting on the Turing test: The Turing test is considered a test of whether or not a machine can “think” in the truest sense of duplicating mental activity—both cortical and limbic.

One might converse with the computer about a song or poem or describe the most tragic accidents. If a computer wants to pass the test correctly, its responses should be indistinguishable from human reactions (R. W. Picard, 1995).

While the Turing test centered on taking place by communicating only via text so that sensory expressions (e.g., voice intonation and facial expression) do not play a role, emotions can still be perceived in text and might still be elicited by its content and form. Following this statement, a machine cannot pass the Turing test unless it can also perceive and express emotions” (R. W. Picard, 1995). Another representative example is when Picard (1995) and her team tried to build a better piano-teaching computer system. The aim of this project was that this system would be able to read your gestures, timing, and phrasing as well as your human emotional state.

One of the psychological models used in this project by AI experts is the one that distinguishes between three basic emotions that we are all supposedly born with -interest, pleasure, and distress- (Lewis, 1995).

Therefore, the goal of this project was that the AI teaching system could maximize pleasure and interest while minimizing distress. In this line and according to Picard:

> "With observations of your emotions, the computer teacher could respond to you more like the best human teachers, giving you one-on-one personalized guidance as you explore. We know by learning processes that, at least at the beginning, every learning episode could raise our curiosity and fascination. If the learning process becomes complex and complicated, feelings like confusion, anxiety, and frustration can appear, and we could abandon them because of these negative feelings (R. W. Picard, 1995).

Once again, she emphasizes the importance of generating interface agents to learn our preferences, much like a trusted assistant. In her own words: "The agent might notice our response to too much information as a function of valence (pleasure/displeasure) with the content; the agent, learning to distinguish which information features best please the user while meeting their needs, could adjust itself appropriately. "User-friendly" and "personal computing" would move closer to their true meanings (R. W. Picard, 1995).

The scenario of emotions presented by Picard (1995) in her working paper on *Affective Computing* opens the door to understanding the world of emotions and feelings systematically and measurably. Since then, this approach marks a turning point, at least in the role that emotions have played in human intelligence and cognition in the Western tradition and in how AI is beginning to be conceived in certain areas. However, as explained above, the work of Picard and her colleagues is not without shortcuts and other ethical implications.

6.3.2 Same emotional data but different meanings

After briefly introducing the field of Affective Computation, we dive into this section into two distinct types of emotional theories. On the one hand, we have the classical theory; on the other hand, we have a recent theory that interprets the world of emotions as a construct.

FIGURE 23 | Comparison of the two theories on emotions / Ukususha / iStock

These different approaches propose and conceive affective computing or artificial emotional intelligence differently, and the challenges and ethical implications could significantly impact our societies and the way we understand decision-making and choices.

Do physical fingerprints of emotions exist?

Beyond the paper discussed at the beginning of this chapter (Feldman Barrett & Satpute, 2017), the neuroscientist and psychologist Lisa Feldman Barret (2017) published a book called *How emotions are made,* in which she analyses emotions from a constructivist perspective. She offers an alternative view of the classical theory of emotions. However, before explaining what this constructionist theory of emotions consists of, it is helpful to analyze the points that Barret criticizes from the classical theory of emotions.

According to the constructed theory of emotion (Feldman Barret, 2017), one of the controversial elements of this classical theory is the notion of "emotional fingerprinting" because it is believed that every emotion is likewise assumed to be similar enough from one instance to the next and in one person to the next, regardless of age, sex, personality, or culture. In a laboratory, scientists should be able to tell whether someone is sad, happy, or anxious just by looking at physical measurements of a person's face, body, and brain" (Feldman Barret, 2017).

According to Feldman Barret (2017), an undoubted source of inspiration for the classical theory of emotions is the main idea contained in Charles Darwin's (1872) book *The Expression of the Emotions in Man and Animals*, in which he writes that emotions and their expressions were an ancient part of universal human nature. Such a claim is the same as saying that all people, everywhere in the world, are said to exhibit and recognize facial expressions of emotion without any training whatsoever (Darwin & Lorenz, 1872).

Feldman Barret (2017) argues that the kind of education we have received is based only on the classical theory of emotions and that it is not surprising that when someone does not understand your mood or emotional state, you are left thinking there is something wrong, like a precision error or something missing. Nevertheless, the cause of this feeling that we are missing something is responsible, according to Feldman Barret (2017), to the scientific literature that supports the classical theory of emotions without doubts. Examples include a group of psychologists who try to consolidate the tradition started with Darwin by testing this idea of the universality of emotions in the laboratory (Alfaisal & Aljanada, 2018; Izard & Haynes, 1988; Leys, 2017; Ortony, 2021).

According to Feldman Barret's (2017) description of some of these experiments, for example, conducted by Tomkins and his crew artificially created a set of exaggerated posed photographs to represent the following six basic emotions they believed had biological fingerprints: anger, fear, disgust, surprise, sadness, and happiness. These photos were supposed to be the clearest examples of facial expressions for these emotions, and therefore, they were used as tools to test people's ability to recognize emotional faces.

Since then, many experiments have supported this line of research, and it is considered the gold standard today. Therefore, emotion-modeled faces can be used to give computers graphical faces that mimic these precise expressions identified by Ekman (1977), making the computer faces seem more human and thus further crystallizing this universalistic theory of emotions (Ekman, 2017; Eknian, 1980).

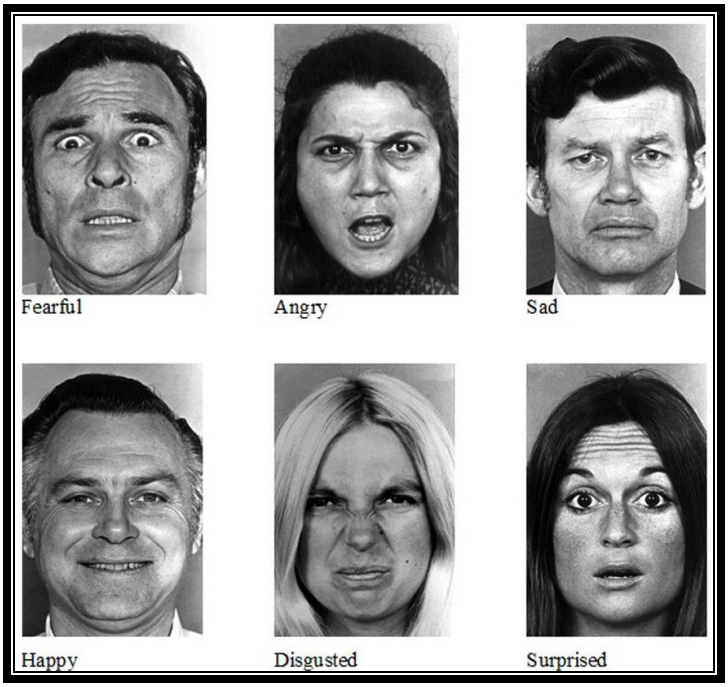


FIGURE 24 | Paul Ekman's classic study on basic emotions (Ekman et al., 1969) / delve.ai

However, just as neuroimaging has helped advance our understanding of the brain and the field of neuroscience, another technology called facial EMG has again critically helped advance science and has already presented a severe challenge to the classical view of emotion. According to information compiled by Feldman Barret (2017) in study after study, facial EMG shows that muscle movements do not reliably indicate when someone is angry, sad, or fearful; they don't form predictable fingerprints for each emotion (Feldman Barrett & Satpute, 2017).

Nevertheless, what if "the individual's computer can acquire ambient perceptual and contextual information -e.g., see if you are climbing stairs, detect if the room temperature changed to identify autonomic emotional responses conditioned on perceivable non-emotional factors"? (R. W. Picard, 1995)

Therefore, letting applications on our computers, mobiles, or other devices collect our moods and establish correlations with emotions is brilliant if they can do it accurately. This issue is beyond the scope of this research but is still related to how we want and could use these technologies and their ethical implications.

It is difficult to doubt that the work of Feldman Barret (2017) carries an enormous weight in giving a more critical, serious, and scientific character to a theory of emotions.

On the other side, the ideas of Picard (1995) of letting our computers, mobile phones, or other devices that collect our "sentic states" (Clynes, 1988) establish correlations with certain emotions could appear as a brilliant idea and could become even a milestone in the science of emotion.

However, what happens is that if we start from the idea that emotions are not universal or "do not have fingerprints" (Feldman Barret, 2017), we fall into the establishment of false correlations if we simplify and classify what we call in the previous chapter "data from the wild" (Cristianini, 2019) into a concrete emotional category by giving it a universal meaning. This way of not considering the concrete context could lead us to represent reality in a biased way with ethical consequences that we need to consider with seriousness and urgency.

6.3.3 The rule: "variation is the norm"

Going directly to Feldman Barret's (2017) theory of constructed emotions, in which, for example, an emotion such as "fear" is considered a "secondary emotion" (Damasio, 1994), "does not have a single expression but a diverse population of facial movements that vary from one situation to the next: "fear" takes no single physical form. Variation is the norm" (Feldman Barret, 2017).

But then, how is it possible for us to find it natural to believe that a wide-eyed face is the universal fear expression? According to Feldman Barret:

> "This correlation is a product of the culture in which one learns and lives. Teachers in school teach children these stereotypes: "People who scowl are angry. People who pout are sad". These conventions are sharpened by advertising, doll faces, in emojis. Moreover, psychology faculties teach these things in their textbooks, and therapists teach their patients this way. One of the critical points that Barret (2017) mentions in her book is that this way of interpreting emotions is because "the media spreads them widely throughout the Western world" (Feldman Barret, 2017).

However, Feldman Barret's (2017) search for "unique fingerprints of emotion" did not end with the human face. She conducted experiments to test whether we can measure those fingerprints from changes in heart rate, blood pressure, and other body functions. Her conclusion was the same as her previous research, which she designates as the same emotion category involving different bodily responses. Variation, not uniformity, is the norm.

These results are consistent with what physiologists have known for over fifty years: different behaviors have different patterns of heart rate, breathing, and so on to support their unique movements. If we want to understand emotions, it is crucial to understand what she calls "pattern emerging" or "variation is the norm" (Feldman Barret, 2017; Woodley, 2020).

Feldman-Barret focuses in her book (2017) on the emotional word "fear" for its study in the scientific literature and affirms that there is no single physical fingerprint but a very variable group of instances that simultaneously belong to specific situations. Barret also believes that what we colloquially call emotions such as anger, fear, and happiness are better termed as emotional categories because each is a collection of very different instances.

In the field of affective computing, explored in depth in the following section, Feldman Barret (2017) argues that some scientists use AI techniques to train a software program to recognize many brain scans of people experiencing different emotions (say, anger and fear).

The program computes a statistical pattern that summarizes each emotion category and then can analyze new scans and determine if they are closer to the summary pattern for anger or fear.

This technique, called pattern classification, works so well that it's sometimes called "neural mind-reading."

According to her, some scientists claim they have found "neural fingerprints" for anger and fear. The statistical pattern for fear is not an actual brain state, just an abstract summary of many instances of fear. These scientists are mistaking a mathematical average for the norm. The following section adds more information on whether this impedes measuring these "emotional categories" or whether we need to check other variables (Feldman Barret, 2017).

6.4 Emotions beyond East and West: more diverse than universal

Chapter 4, devoted to machine ethics and explainable AI, raised issues similar to those we are dealing with in this chapter. To see it more clearly, it is worth remembering that within the ethical frameworks intended to create artificial moral agents (Allen et al., 2005), at least in the West, these frameworks practically centered on the consequentialist and deontological frameworks. This view led this dissertation to wonder whether these ethical frameworks were sufficiently universal that they also inspired the currents of machine ethics or robotic ethics in the East.

Now, given that the role of emotions in human moral judgments and decision-making is more than justified, and given that AI systems increasingly make decisions for us and to us, it is more than urgent to know what is being done in the field of artificial emotional intelligence beyond the United States and Europe. In particular, we take the point of view of the most prevailing theory of emotions in Japan, and for this, it is appropriate to make a historical and comparative journey between the West and the East**.**

6.4.1 Measuring intelligence in Western and Eastern traditions

The concept of AI in English puts cognition in the process of representing human intelligence in machines in the first place. This fact is due, among other things, to the enormous influence of the popularization of Alan Turing's paper *Computing Machinery and Intelligence*, "from which is derived an enduring legacy that associates the measure of intelligence with a measurement of "thought" (Turing, 1950).

However, other scientific traditions, like in Japan, proposed "emotion" in the 1960s as an equally important ingredient of intelligence (White & Katsuno, 2022).

This fact has its origin in the terminology that the Japanese tradition has used to denominate "intelligence" (*chinō*) and "mind" (*Kokoro*). These terms refer symbolically to the heart as much as to the brain. This way of understanding intelligence as an embodied capacity is quite different from the Western tradition. For this reason, the book Rosalind Picard's *Affective Computing* (R. Picard, 1997) and Marvin Minsky's *The Emotion Machine* (Gelfand, 2006) were initially seen as exceptional and even marginal perspectives of intelligence in anglophone research.

Another difference with this way of understanding emotions in Japan is the strong dualism (Damasio, 1994) that has characterized the West regarding the role and origin of emotions. Before psychology started to be a science, philosophy was the discipline that was also interested in emotions, and it classified them into two approaches according to the terminology proposed by Coeckelbergh (2010): the cognitivist theory and the feeling theory. According to cognitivism, emotions are propositional attitudes, beliefs, or judgments (de Sousa, 1987; Goldie, 2000; Nussbaum, 2001; Solomon, 1980). Nussbaum's (2001) tradition is the neo-Stoic; she claims that emotions are judgments. On the contrary, in feeling theories of emotions, emphasis on emotions can be understood as an awareness of bodily changes (JAMES, 1884; Prinz, 2004).

According to James (1884), the feeling of bodily changes is the emotion; the mental state follows those changes instead of preceding them. To put it even more succinctly: "For cognitivism, emotions are more a matter of "mind" than of "body," whereas for feeling theories, emotions are more a matter of "body" than "mind" (Coeckelbergh, 2010). According to White and Katsuno:

> "Such cultural differences in the approach to representing intelligence in machines suggest the important role that social context plays in the production not only of technologies that are manufacturing and collecting new forms of emotional data but also of the theories of emotion on which those technologies rely" (White & Katsuno, 2022).

For this reason, this chapter has begun with an analysis of the most representative theories of emotion up to the present day.

6.4.2 The blind business of measuring "universal emotions"

Affecting computing, according to Picard (1995), is a field of research that has as its primary concerns issues such as researching and engineering computers and software that can "recognize" human emotion, can "express" and realize emotion, and can somehow even "have" emotion.

Thus, Picard assumes that the understanding of emotions and even their nature can be decoded by computer scientists with the help of a particular coding kit, natural language processing, and automation (R. Picard, 1997; R. W. Picard, 1995). However, there is another side to the coin, according to White and Katsuno (2022), because these tools, combined with "the ubiquity of social media platforms" and their free costs, mobile devices, and other wearables, make an unprecedented variety of physiological tracking state measurements possible. These measures have led to the explosion of what is known as emotional data and have become an attractive commercial field to win consumers for tracking, self-development, and self-care.

Consequently, there is an unstoppable development in creating AI systems that automate the detection of human emotions. However, although these systems are already present in everyday use, their accuracy and legitimacy remain controversial (White & Katsuno, 2022).

Controversial technologies already on the market

For the issue that concerns us here, it is worth mentioning interpretative methods of facial expression recognition because of the ethical implications and dilemmas it opens up.

Thus, according to Crawford (2021), security personnel have employed these systems, such as in American airports after the 9/11 attacks on the World Trade Center. But similar systems have remained in operation, such as a version of facial expression and physiological recognition technology called "I-BORDER-CTRL" designed by European Dynamics, which has also been tested at EU border gates to offer "lie-detecting avatars" and "advanced analytics" for "risk-based management" (Barrett et al., 2019; Gentzel, 2021).

According to different project summaries, “this unique approach to deception detection” analyzes travelers' micro-gestures to determine if the interviewee is lying (Boffey, 2018; de Paiva-Silva et al., 2016; Hall & Clapton, 2021).

Although these technologies are already in the market in various forms and give the impression of providing an objective measure of the internal feeling states, it is crucial to note, as do the authors White and Katsuno (2022), that such technologies record only visible and somehow superficial signs of such emotional states.

An example discussed earlier in this chapter of these critical voices is the research work conducted by psychologist and neuroscientist Lisa Feldman Barret (2017) with her critique of traditional emotion theory and its unsuccessful attempts to look for the fingerprints of emotions.

According to White and Katsuno (2022), this context “engenders multiple tensions among marketers, developers, and researchers between those who are encouraged by the presumed but misleading universality of emotional AI platforms and those who aim to deliver tools to support the cultivation of emotional intelligence that are also sensitive to cultural diversity” (White & Katsuno, 2022).

This tension is the product of modeling by programmers and robotics experts opting to build machines capable of registering emotional states using models of emotions derived from psychology that are more subject to quantifiable interpretation, without it being clear that this is the most accurate way at the expense of others.

The computer scientists Aylett and Paiva summarize boldly and comprehensively the technological, ethical, and social implications of this challenge:

> “To implement any model on a computer, the model itself must be sufficiently specific. From this perspective, many psychological models are not usable as they stand, but must be operationalized. Qualitative relationships must be quantified… Thus, when computer scientists select models from psychology, they tend to favour those that are already sufficiently specific, or that can be made relatively easily” (Aylett & Paiva, 2012).

The prevailing trend among engineers currently in charge of the technical side of emotion modeling tends toward models that are easily implementable in autonomous systems and quantitative data generators.

The previous section coincides with Barret's (2017) critique of the search for emotion "fingerprints," and one representative example of this approach is psychologist Paul Ekman's (1960) theory of basic emotions.

Ekman (1999) is known for developing a model of 6 basic emotions that are considered to be universally identifiable in facial expressions across cultures (Ekman, 2005). However, Ekman's most relevant work was earlier in 1978, together with Carl-Herman Hjortsjö and Friesen, in which they designed what is known as the Facial Action Coding System (FACS). According to White and Katsuno (2022), FACS provides programmers with a systematic means to code facial expressions in a way that is easily implemented in software (Clark et al., 2020).

Despite numerous expert criticisms from psychology (Feldman Barret, 2017), anthropology (White, 2017), and experiments about the future of human-robot interaction (Spezialetti et al., 2020) of this way of endowing AI systems and robots with a form of emotional artificial intelligence, Ekman's model of the basic universal of emotions has crept into the world of engineering globally. Nevertheless, there are other ways to design emotional technology that are not based on universal patterns but consider the particular subjectivity of human senses on a changing spatial-temporal axis. An example of this is what in Japan is known as the *Kansei kougaku* (Kansei engineering or affective engineering). This line of research is presented below as it formally fits with the objectives of this research.

6.4.3 Emotional Data: finding meaning within a culture

Although this section focuses on a particular approach to how emotional technology is implemented in Japan, Ekman's model is still prevalent. Examples of this (White & Katsuno, 2022) are the early Face Robot of Kobayashi Hiroshi and colleagues (1994), SoftBank's companion robot Pepper (SoftBank, 2014), Fujisoft's communication robot Palro (Fujisoft, 2021), and Sony's pet robot AIBO (Sony, 2021). However, it is essential to consider other less universal approaches, such as the one presented below.

As mentioned above, Japanese culture, or rather the concepts related to intelligence, has not only been a matter of thought like in the West; rather, this culture understood that intelligence needs a heart or, in other words, emotions in a body. But then, how does this affective engineering (Kansei engineering) materialize, and what features characterize it?

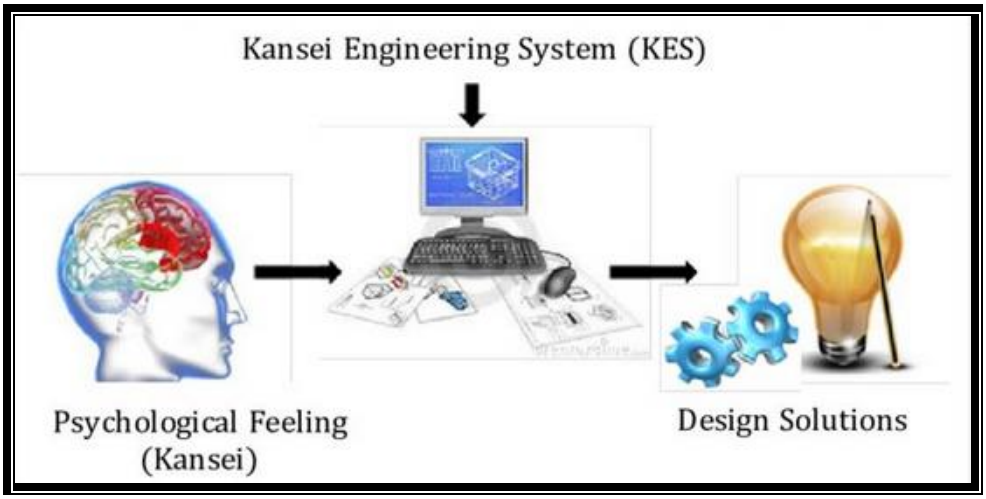


FIGURE 25 | Kansei Engineering System (KES) (Nagamachi, 2002)

Kansei engineering emerged in the 1980s in Japan as a technological, methodological approach. At Hiroshima University, the work of Mitsuo Nagamachi established what he called Jocho Kogaku "emotional engineering" in 1970. In 1988, the field was renamed as it is known today: affective engineering (Kansei engineering).

Unlike affective computing (R. W. Picard, 1995), Kansei engineering focuses on the subjective side of human emotion, i.e., trying to understand emotion as dependent on one's interaction with objects and the environment and processes of quantification and automation.

This engineering methodology combines affective psychology disciplines with mathematics, engineering, and machine learning. The goal is to analyze and reveal human sensing capabilities using qualitative and quantitative metrics.

Closer to the 1990s along these lines, researchers in Japan acknowledged that this ergonomics-oriented engineering for mass production based on the uniformity of the human body was insufficient for customer needs. In 1991, a national project, *The First Stage of Human Sensory Measurement and Application Technology*, was launched (White & Katsuno, 2022).

Over time, the need to consider increasing the subjective sensory aspects and emotions built interactively between the agent and the environment was emphasized. Kansei tries to set up the ability to extract information from stimuli from the outside world and simultaneously transmit information to the outside world (Lokman, 2010; Nagamachi, 2002). In this approach, a person's Kansei is nurtured by culture.

For his part, Harada (1999) sets out in a survey the essential points on which Kansei engineering focuses: 1) subjective and unexplainable action; 2) cognitive action by knowledge and experience in addition to inborn nature; 3) interaction between intuition and intellectual activity; 4) the ability to intuitively respond to and evaluate features such as beauty and pleasure; and 5) the action of the mind to create images.

Therefore, and to close this section, the most crucial points to consider are that this Kansei approach, compared with the affective computing current, is directed at an orientation of the dynamic and culturally particular possibilities of mutual influence between subject, technological object, and environment. In addition, as already stated above, some Kansei engineering researchers pay close attention to the emotions and the heart (Kokoro) in contrast to Western intellectualism, rationality, and rationalization (Stappers et al., 2002).

6.5 In search of "variation is the norm"

After this review of the different theories of emotions, followed by the two somewhat antagonistic forms of engineering described in the previous section, providing AI systems with the ability to monitor our physiological states to capture our emotions or affective states does not seem an easy task if we want to be precise and rigorous in their study and in designing this kind of technologies.

In fact, from what we have seen so far, it is the theory of the philosopher Coeckelbergh (2010) that is a good starting point for understanding one of the critical problems we face.

As mentioned above, Coeckelbergh divides the most representative literature on emotions into two groups: the cognitive theory and the affective theory, “for both theories, it turns out, mental states and consciousness are necessary conditions for having emotions or emotions are themselves mental states” (Coeckelbergh, 2010). It is important to note that both theories are part of the classical view of emotions (Feldman Barret, 2017).

From a philosophical point of view, Coeckelbergh (2010) reaches similar conclusions to Feldman Barret's since the philosopher asks, “How can we know which of these theories is true? Coeckelbergh answers, “After all, we are not even certain that other humans are conscious (…) that our social and moral life depends on appearance”.

This philosopher solves this question of appearance by changing the focus from the “outside” rather than the “inside” but is not “behaviorist but instead phenomenological” (Coeckelbergh, 2010). However, this vision offered by Coeckelbergh (2010) is not new.

This “hybrid model” of emotions was portrayed many years before by the pragmatist American philosopher, psychologist, and educational reformer John Dewey (Morse, 2010).

The interesting point is whether the American philosopher’s understanding of emotions has a scientific basis. As Dewey (1925) wrote in *Experience and Nature*, “It is not certain that first endow man in isolation with an instinct of fear and then…imagine him... ejecting that fear into the environment”.

The truth, instead, is that “man fears because he exists in a fearful, awful world. The world is precarious and perilous” (Dewey, 1925). Dewey’s view may seem poetic and metaphorical, not that the world is frightened, but that “individuals serve as the seat of emotions” (Morse, 2010).

But what, then, does Dewey mean when he says that the world is primarily fearful? Dewey’s way of arguing is due to his attempt to overcome dualism.

That is something that the neuroscientist Antonio Damasio (1994) tried to discover and explain years later in his book *Descartes’ Error* by trying to understand how emotions underlie and guide our reasoning.

For Dewey, reason and emotion are so intertwined that one never thinks without feeling. As Dewey sees it, a rational inquiry is not a mode of thinking about objects or events but rather a "dramatic rehearsal" of events themselves (Morse, 2010).

6.5.1 Understanding, communicating, and explaining emotions

Thus, fear, anger, happiness, and sadness exist in the world, but how do they materialize in each of us? Or, better said, in another way, is the expression of these emotions the same in each person? And what is the function of these states? Each of these issues is treated separately and one at a time.

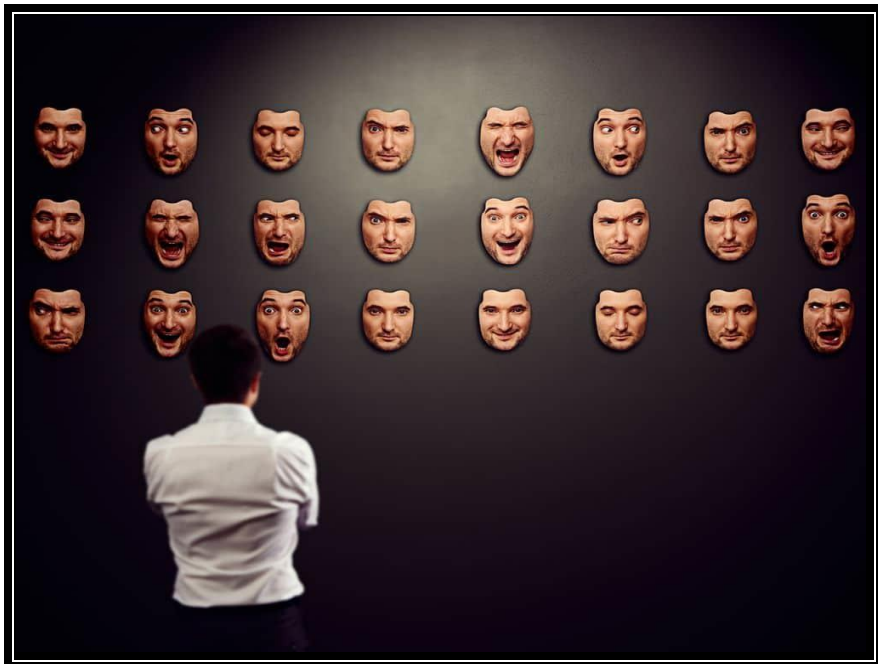


FIGURE 26 | "Variation is the norm" / shutterstock.com

Understanding how emotions are expressed individually could lead to a better understanding of what Feldman Barret has called "variation is the norm" (Feldman Barret, 2017).

In addition, she talks about a technique that scientists could use to measure these variations called “meta-analysis,” which consists of analyzing all the data together and thus reaching a unified conclusion. She argues that thanks to this meta-analysis, different scientists can conduct many experiments and combine their results statistically.

One of these experiments performed by laboratories worldwide was to evaluate whether anger has its fingerprints on the body. For this purpose, one of the experiments designed was the following:

> “Test subjects perform a difficult task such as counting backwards by thirteen as fast as possible or speaking about a polarizing topic like abortion or religion while being ridiculed. As they struggle, the experimenter berates them for poor performance, making critical and even insulting remarks (Feldman Barret, 2017).

But then, did the test subjects act in the same way? Did they get angry? No, they did not.

Those who felt angry most notably showed different behavior patterns and body changes. Some fumed with rage, some cried, and others became quiet and cunning. Still, others withdraw. Thus, each behavior was supported by a different physiological pattern in the body.

Most colorful of all, though, is the example Feldman Barrett (2017) proposed of her accumulated experience when she presented this meta-analysis to her audience at the conference. Some of those present were confronted by the neuroscientist and asked: “Are you saying that in a frustrating, humiliating situation, not everyone will get angry, so their blood boils and their palms sweat and their cheeks flush?” (Feldman Barret, 2017).

After having analyzed the specific case of anger from the point of view of neuroscience, we return to philosophy with Coeckelberg, who states that "our theories of emotion and moral agency might assume that emotions require mental states, but in social-emotional practice, we rely on how other humans appear to us (…) we interpret the other's appearance and behavior as an emotion. Moreover, we further interact with them as if they were doing the same with us" (Coeckelbergh, 2010).

Turning to neuroscience, Barret (2017) argues something similar to the phenomenon described by Coeckelberg (2010). This position is a constructivist approach to emotions: "We perceive others as happy, sad or angry by applying our emotion concepts to their moving faces and bodies" (Feldman Barret, 2017).

This view of emotions assumes that they are real and have their function, and even, as seen in the next section, are necessary for survival as part of our daily habits, but given that "variation is the norm" (Feldman Barret, 2017), how are we able to understand the emotions of others? Emotions become natural as a function of two abilities developed by human beings. The first has to do with what Barret (2017) calls the need for a group of people to agree that a concept such as "anger" or "regret" or others such as "flowers," to cite some of them, describes a part of reality.

The other essential element to bear in mind is that this shared knowledge is part of what is known as "collective intentionality."

But what is interesting is what makes these intentional categories or words crystallized into emotional concepts possible. The primary function is to communicate to someone what you feel, and this is only possible if those involved share the same understanding of the concept of fear.

Therefore, without this collective intentionality, a person's actions, however meaningful they may be to him or herself, will be perceived by others as noise without context. Now, given that the theory underpinning this research is based on the hypothesis put forward by Barret (2017), namely that "variation is the norm," how do we come to understand the emotional world of others?

6.5.2 The role of habit in understanding emotional patterns

Throughout this chapter 6, we have gone through a historical labyrinth that has brought us closer to different theories of emotions. We have gone through Western theories and recovered the concept of intelligence that permeates a part of affective engineering (Kansei engineering) in Japan.

The peculiar aspect of this type of engineering is that it tends to consider emotional expression or, instead, the subjectivities of human beings as an ingredient to be understood in its complexity. This way of proceeding follows what Feldman Barret (2017) defends as "variation is the norm" or, in philosophical terms, what Coeckelberg (2010) describes as the way to know another person's emotional state through a process of interpretation based on how we perceive our emotional states.

Now let us look at the theory of Dewey, who has points in common with Damasio's (1994) somatic marker hypothesis but takes it further to the extreme as Dewey (1895) puts it, indeed a long time before: "no cold, indifferent process of thinking leads us by sheer force of reason alone, not on this account."

Instead, we are inside a confounded situation; we feel this disturbance ourselves and are disturbed, and our reasoning is an attempt to feel our way back into a non-disturbed situation and a no longer disturbed state of being. Here, our reasoning is felt, which is motivated by emotions both in its beginning, middle, and end" (Morse, 2010).

For Dewey (1895), habits are crucial. In his own words:

> "When our usual modes of conduct disagree, we envision the response we should make to the divergence before we make it. What happens is that each conflicting habit takes its turn in projecting itself upon the screen of imagination. It unrolls a picture of its future history and the career it would have if given head". Each habit makes a case for itself as the habit we ought to heed; it presents a rehearsal of what would happen in real life if we followed it (Morse, 2010).

If we examine the concept of habit through the lens of Feldman Barret (2017), we find that both habit and emotion, understood in Dewey's terminology, are focused on predicting what will come next.

This statement is what scientists Lochmann and Deneve call: "Meaning-making within the brain is a predictive activity" (Lochmann & Deneve, 2011). According to Feldman Barret, "past experiences are reconstructed as partial neural patterns that serve as prediction signals (also known as "top-down" or "feedback" signals, and more recently as "forward" models) to continuously anticipate events in the sensory environment, in part, by planning for motor and visceromotor changes. Without experience as a guide, the brain cannot transform flashes of light into sights, chemicals into smells, and variable air pressure into music. The result is experientially unwatched" (Feldman Barret, 2017).

And back to Dewey, what role do emotions play in this understanding of habits in Dewey's theory? For Dewey (1895), emotions play a fundamental role in solving problems or making decisions. In other words, and taking Dewey's line of argumentation, an irrational response is considered when the object of thought or an element to predict what is coming next that motivates us to exercise a habit is caused by an old feeling that pushes us to react in the same way we have done before even knowing that there are other possible alternatives.

On the other hand, a rational response is one where the object of thought motivates us to organize the situation by imagining better ways of modifying that habit if necessary. Dewey (1988) goes so far as to say that, in essence, reason is a kind of emotion, "a thoughtful emotion" (Morse, 2010). This thesis doesn't intend to argue this point, but it is important to note that emotion is more than something that resides in our heads; it is a response to the world and to what exists.

Therefore, projecting a privileged vision of reason over emotions or vice versa produces a dissociation between one's own experience and nature understood in its globality. Seeing only rational experience as accurate to the exclusion of other experiences that are also considered real, be they ethical, emotional, or aesthetic, leads in some ways to irrational responses or decision-making as described above. As Morse (2010) comments on Dewey's theory, the tensions and disturbances we feel daily and throughout life cannot be discounted; they are real –undeniably what they present themselves to be in our experiences.

If our emotional concern for things like achievement and frustration is in nature, it is factual and objective, a genuine feature of nature. But this does not and cannot mean for Dewey, as Morse explains, that "nature simply justifies man's pursuit of his own needs. On the contrary, and as we have seen, these needs of man (and their satisfactions) are objective. We are searching for workable ideas, values shaped by intelligence, and values that can be manifest" (Morse, 2010). In this line, an intelligent response that would be rational in the terms expressed above and as an ethical consequence is "one where we work toward a union of stable and hazardous, balance and unbalance. It is one where we embrace the "unbalanced balance of things. In short, what Dewey is trying to express, sometimes even in poetic or artistic language, is, according to Morse, that "the problems we genuinely and legitimately feel are always problems of nature, and that the responses we make to our problems, therefore, ought also to be responses to nature" (Morse, 2010). This way of understanding habits, decision-making processes, and even intelligence from Dewey's (1925) theory, described as an "unbalanced balance of things," has a corresponding concept in biology called allostasis, a term coined by Sterling and Eyer, which they define as: "remaining stable by being variable" (Sterling & Eyer, 1988).

Isn't this concept of allostasis the way to understand the changes in habits through "thoughtful emotion" in Dewey's theory? It is worth highlighting before concluding this chapter that this way of producing order from disorder in the physical and social world is an even broader current project (Assad, 1991; Prigogine & Stengers, 1984).

Although these concepts are beyond the scope of this research, I have chosen to comment briefly on them to show a significant shortcoming of science: the lack of communication between different scientific fields. This lack of consensus on concepts that pursue similar objectives leads to a fragmentation of knowledge and a poor understanding of how to deal with complex phenomena and problems such as those we face today.

6.6 Design a ToM about emotions in goal-driven XAI

6.6.1 A Hybrid approach and learning for explainability

This chapter has highlighted and illustrated the importance of emotions in human deliberative processes and moral human judgments with different examples and theories. However, the challenges and limitations still faced in systematizing a deeper study of emotions raise the question of whether artificial intelligence can do more for us. Therefore, and as a reminder, one of this dissertation's main goals is to find the connection between the lack of scientific knowledge about how emotions impact human reasoning and the result of this knowledge gap in forming the two sources of error in human reasoning: bias and noise. Therefore, we are not talking about implementing an emotional theory in an AI system but rather about understanding what emotional contents influence or impede the achievement of goal-oriented behavior.

But back to the initial question: How can we measure an emotion? Can we consider emotions as sensory data? Is it an emotional category that belongs instead to how a group of people use a common language? Again, in short, while the classical theory seeks to establish clear causality between bodily changes and what they represent in the face of a particular emotion, the constructivist theory appeals to the fact that the semantic content given to each emotion is instead based on an agreement between a specific culture, society, or the members of a community. This last approach, for example, may explain the results of the experiment conducted by Gold, Colman, and Pulfold (2014) about the differences in responses to real-life and hypothetical trolley problems between Chinese and British participants. However, to establish a systematized methodology, it is necessary to study this phenomenon in greater depth. Understanding whether an emotion can be measured in both ways or whether one is more plausible than the other to establish a more systematized, objective, and scientific science of emotions is the goal pursued using the research field of goal-driven XAI (XGDAI). This first section, which corresponds to the last part of Chapter 6, introduces, describes, and analyses the three behavioral architectures traditionally distinguished in the literature about goal-driven XAI and the role of continuous learning in explainability.

In this area, it is essential to emphasize that each behavioral architecture creates a different theory of mind in the AI system, generating a different shape in the agent's behavior and how it interacts with the world. In addition, since one of the characteristics of this approach is continuous learning, the system modifies its objectives, plans, or actions based on the changes introduced and the learning it obtains from new data. These new data cause new adjustments to be made according to the specific plans for the AI system. So, what three types of traditional explanations can be designed in Explainable goal-driven AI (XGDAI)? According to Sado et al., the first group corresponds to the deliberative type. In this case, the agent deliberates on its goals, plans, or actions; that is to say, the AI system should possess a symbolic representation of the world. Sado et al. propose different explanatory architectures for the deliberative type, but the work of Borgo et al., called the XAI-Plan model, is the most representative for this research work.

In summary, this model provides an immediate explanation for the decision made by the agent planner. In this way, the model produces explanations by encouraging users to try different alternatives in the plans and compares the subsequent plans with those of the planner. The interactions between the planner and the user enhance hybrid-initiative planning that can improve the final plan (Borgo et al., 2018; Sado et al., 2020).

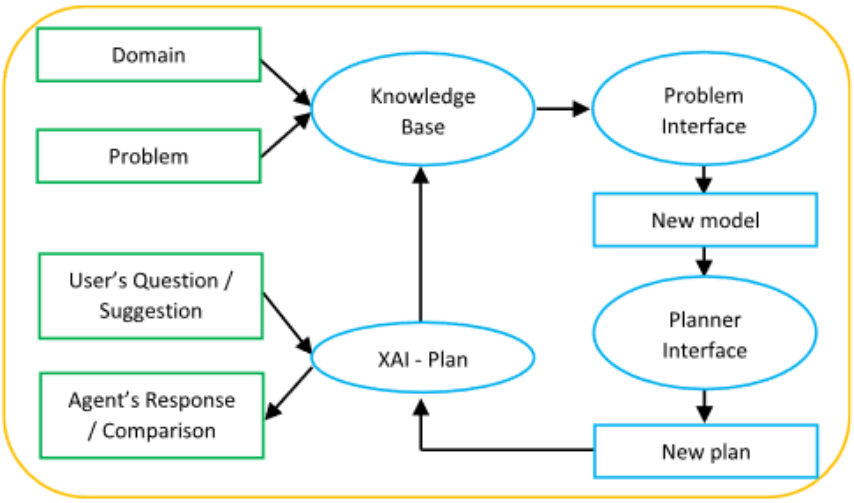


FIGURE 27 | Example XAI-Plan architecture by Borgo et al. The XAI-Plan node generates explanatory plans and communicates the agent's response through a user interface. ROSPlan provides the planner interface, problem interface, and knowledge base (Sado et al., 2020)

The second group of this classification corresponds to what are called reactive explanations. In this instance, the artificial agent implements some simple behavioral patterns and reacts to activities or events in its environment in a stimulus-response way.

In this kind of explanation, no model of the world is required. Some of the work cited by Sado et al. to illustrate this type of reactive explanation is in the area of eXplainable Reinforcement Learning (XRL), such as the work performed by Cruz et al. in which an RL agent explains to the users why it selected an action over the other possible actions.

This model uses an episodic memory (Sequeira & Gervasio, 2020) to save each episode or agent's record of executed state-action combinations, then computes both the likelihood of success (Q-values) as well as the number of transitions within each episode to meet the final target to provide an explanation or reason for selecting an action over the others (Cruz et al., 2019).

Other XRL strategies for model-free RL agents include the work of Madumal et al., which utilizes causal models to generate "contrastive" explanations (e.g., "why" and "why not" questions) as a means of describing partly measurable agent action or actions in a game scenario. The approach is to learn a structural causal model (SCM) during RL and to generate explanations for "why" and "why not" questions by counterfactual analysis of the learned SCM. However, one weakness of the approach is that the causal model must be given beforehand.

This weakness is one of the characteristics of this type of reactive explanation because RL agents do not generally reason or plan for their future actions, making it challenging to explain their behavior. According to Sequeira and Gervasio: " an RL agent may eventually learn that one action is preferred over others or that choosing an action is associated with a higher value to attain the goal but would lose the rationale behind such a decision at the end of the process."

Furthermore, as shown in the classification proposed by Sado et al. (2020), in addition to these three groups, each of these architectures can be further subdivided according to how they behave. Thus, for example, there are goal-driven autonomy, goal-driven agency, and Belief-Desire-Intention. However, for our theory of emotions, the subgroup could be called goal-driven emotions or, in the even more specific case of our framework, Responsibility-Anticipated-Regret.

However, the type of explanation that interests us more in this dissertation is the third group. The hybrid approach combines both types of behavior (Davidsson, 1996).

One of the examples proposed for this type of hybrid explanation by Sado et al. is the work proposed by Neerincx et al. named Perceptual-Cognitive eXplanation (PeCox). This Hybrid XGDAI combines Reactive methods' instinctive responses with Deliberative methods' complex reasoning capabilities, as explained in the following figures.

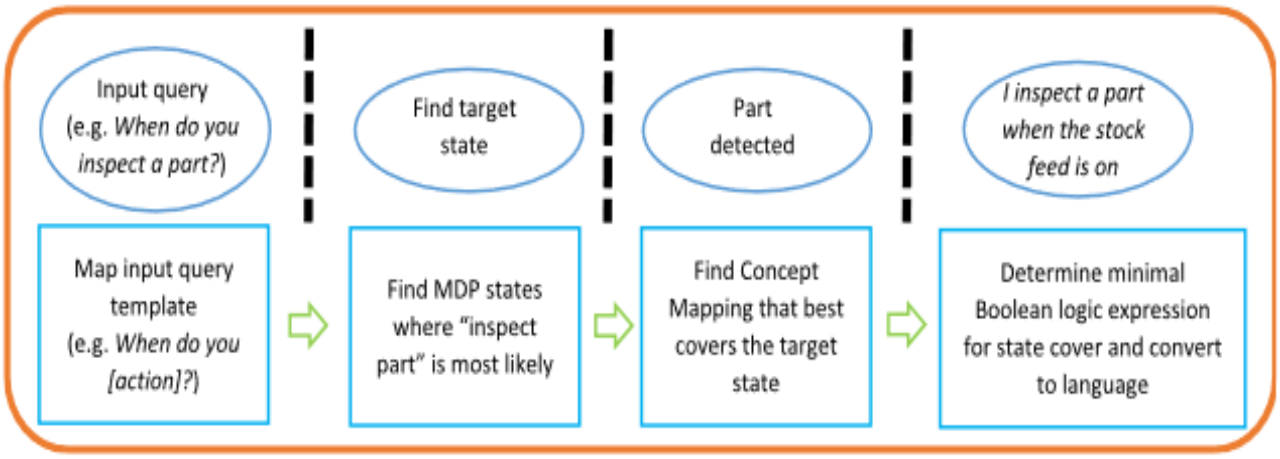


FIGURE 28 | Automated policy generation framework (Hayes & Shah, 2017). The framework maps a policy clarification to an action/input request. An input request or query (e.g., "When do you 'inspect a part'?") is first matched against pre-defined query templates (e.g., "When do you [action]?"). A graphical search algorithm is used to find the state regions that fulfill the query criteria. A concept mapping that best covers the target state is found using logical combinations of communicable predicates. The cover is then reduced and translated into language via the template (Sado et al., 2020)

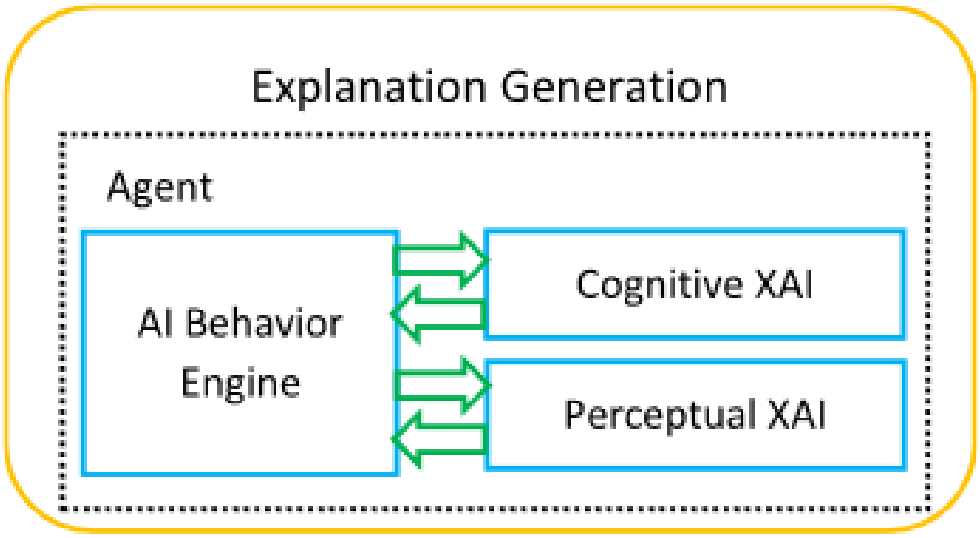


FIGURE 29 | PeCoX generation framework (Neerincx et al., 2018). The framework distinguishes between Perceptual XAI and Cognitive XAI. The Perceptual XAI is connected to the sub-symbolic reasoning of the AI behavior engine and is designed to clarify the perceptual aspect of the agent's behavior. The Cognitive XAI interacts with this part to ground its belief base. It can explain why certain actions have been selected by linking them to goals or beliefs (Sado et al., 2020)

In addition to designing these traditional explanations in an AI system, we have another crucial element: continuous learning. In particular, we focus on explainable techniques that keep the AI system's explanations up to date. This aspect is essential and somewhat paradoxical.

Then, it is assumed that human behavior is characterized as dynamic; however, specific patterns and rules can be detected in the way that the two sources of human errors, bias and noise, are present in human judgments. Therefore, we have chosen the techniques presented in the work of Sado et al. (2020) and commented above with the corresponding figures that combined form the basis for continuous learning for this hybrid system. On the one hand, the system learns based on a deliberative architecture based on Case-Based Reasoning (CBS), and on the other side, the pillar of "explainable reinforcement learning" (Sado et al., 2020).

We return to both aspects when we describe the framework presented in the last chapter. Perhaps the following figure presented by Sado et al. is one of the most compelling proposals of all their work since, unlike other works on explainability, it shows a complex framework of explanatory levels that can provide more specific and detailed knowledge of a phenomenon in depth.

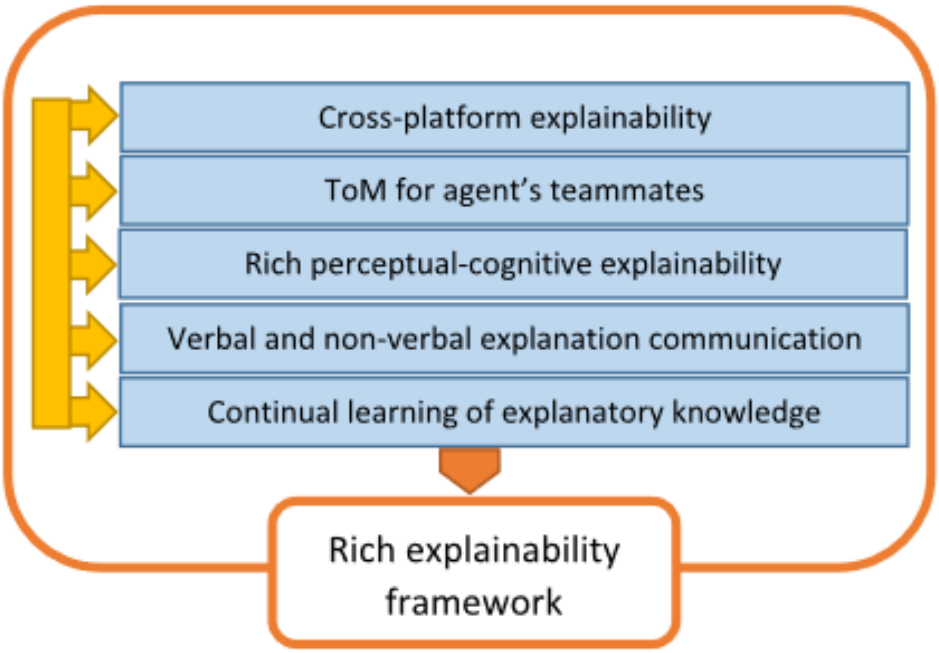


FIGURE 30 | Explainability Framework for XGDAI (Sado et al., 2020)

6.6.2 Design a ToM in XGDAI: Using sensory and cognitive explanations

But then, could we design an explanatory component in an AI system to explain how emotions influence human behavior? Although there is still a long way to go to understand this phenomenon in depth, we choose two research works that present ideas in this direction and are an inspirational starting point for our framework.

On the one hand, the work of Neerincx et al., which closes the previous section, inspires in some way what kind of symbolic and sub-symbolic AI techniques could be implemented to generate a theory of mind in an AI agent.

In this dissertation, the function of the AI system's explanatory component is not to meet the user's expectations but to discover which elements motivated by reason or emotion are involved in forming bias and noise in human judgments.

These shortcuts and pitfalls can produce a distortion in human behavior that might cause unwanted effects that must be prevented before they occur, and therefore, the emotion of regret plays a central role in achieving this goal. Hence, our ToM deals with what kind of effects can have bias and noise in inhibiting the experience of regret at both sensory and cognitive levels. And since regret is the emotion on which responsibility for a bad outcome rests, not being able to experience it opens the door to many questions about how to proceed in case of the inability to experience it. How this emotion is interpreted and related to the principle of responsibility is explained in detail in the next chapter.

Back to the work of Neerincx et al., we also take as a starting point with some modifications to these three phases (shown below) of the creation of an explanation. In turn, what characterizes each phase is the concept of explanation represented by this symbol ε attributed to the work of Tiddi et al. These three phases of an explanation are a subgroup of our framework presented in the next chapter.

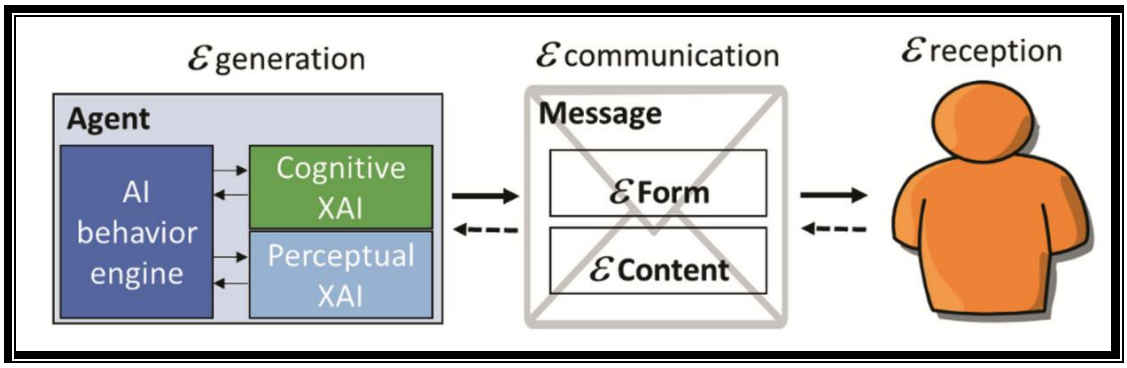


FIGURE 31 | Different phases of an explanation (Neerincx et al., 2018). ε refers to the concept of explanation (Tiddi et al., 2015)

Therefore, this first level (ε generation) corresponds in our framework to the sensory, also called Perceptual XAI (Neerincx et al., 2018), and is usually connected to the sub-symbolic reasoning parts of an AI architecture and shapes the behavior (ToM) of the AI agent. This sub-symbolic architecture is to be understood as related to the explanatory cognitive component that tries to identify which reasoning patterns in the data provided by the sub-symbolic structure do not correspond to the actions proposed for the agent in the symbolic architecture.

According to Neerincx et al.:

> "The coordination and collaboration in human-agent teamwork require the agents' intelligent reactive and anticipatory behaviours. More specifically, they require a shared knowledge representation, methods to comply with policies and agreements for responsible teamwork, and the learning and effectuation of successful patterns of joint activities" (Neerincx et al., 2018).

In fact, along the same argument, we have that the "agent should be able to make pertinent aspects of their status and intentions obvious to their teammates" (Klein et al., 2004).

This assumption of Klein et al. fits with the role played by meta-explanations in the work of Dazeley et al.: "The meta-explanation adds a component of objectivity to the regular explanation, which, although less understandable to the user, results from a more honest and accurate reflection of the decision and the provided explanation" (Dazeley et al., 2021). The ideal state is to achieve these meta-explanations at both levels (deliberative and reactive) for a specific purpose.

If the AI agent detects that the human agent does not express regret in a situation it was supposed to, the system alerts to this distortion in human behavior. In addition, this hybrid approach, in turn, allows us to understand more how emotions can be categorized as data. With this systematization of emotions at both the sensory and cognitive levels, the veracity and usefulness of the two conceptions of the nature of emotions are tested.

That is to say; on one side, emotions can be universally measured because they obey a series of "sentic states" (Clynes, 1988) that do not vary from one population to another. The so-called classical theory of emotions supports this approach. On the other side, we have the nature of emotions as a construction at the level of language categories, that is, as an agreement in form as a social norm that emotions have the same semantic value established by a specific culture, community, or group.

This view corresponds to the approach Feldman Barret (2017) referred to as the theory of constructed emotions. In this line, we have the second source of inspiration entitled *the role of emotion in self-explanations by cognitive agents* conducted by Kaptein et al. (2018). This work is one of the few papers that discuss the importance of emotional content as an essential explanatory component of the agent.

Nevertheless, it is relevant to highlight that both works are mentioned because both methodologies could fill specific gaps in how emotions as "data" are understood and how they influence human behavior.

In the specific case of this dissertation, what is of interest is, on the one hand, to identify the reasoning and emotional mechanisms that lead to the inhibition of the experiential content of regret (Byrne, 2019) and, on the other hand, to reflect on how the explanatory message should be constructed to communicate to the user the importance of minimizing the regrets before they can happen in reality. Therefore, in the explanation process, the second phase proposed by Neerincx et al. is starred by ε communication.

At this level, it is necessary to imagine the format and content the explanation should have. This second level cannot be performed until the first level has been completed. To conclude, the third phase of this explanation process concerns the ε reception, or in other words, whether the human agent understands the explanation proposed by the AI agent. Here at the reception, we are considering an active attitude on the part of the human agent since it is a matter of him being motivated to execute a specific plan to solve a given problem.

However, according to Miller, "there is a significant lack of empirical research with actual human task performers who need explanations in realistic human-agent settings" (Miller et al., 2017).

In our case, which is also a theoretical approach, the reception of the explanation by the user would hypothetically imply executing a series of actions involving specific strategies to prevent the errors that the system would have identified and communicated over a considerable period under continuous learning, as described above.

Thus, if the user, who in our framework is always the decision-maker, does not proceed to react to the alerts of the artificial agent, the first step is to understand the emotional and rational mechanism of the user for not agreeing to the system's recommendations. But if, in the future, the decision maker keeps making similar mistakes that are detected by the artificial agent again and again, the decision maker's behavior should be reinforced to achieve the goal set for both agents, the human and the AI system. The steps to set up this framework are explained in Chapters 7 and 8.

6.6.3 The role of anticipated regret in responsibility

But before moving on to the design of the interdisciplinary framework in Chapter 8, about how to create a strategy to improve responsible decision-making in human-AI interaction, an overview of how an approach based on hybrid explanations can contribute to generating a theory of mind about emotions is offered. This theory of mind about the emotions that are intended to be created in the agents obeys to understand the rules that underlie a particular emotional category and a principle that we want to achieve.

To be understood in our case, we have chosen the emotion of regret and its relationship with the regulation of behavior to establish responsibilities in decision-making. In other words, emotions play a role in how we behave. What is interesting here is understanding how regret drives what kind of behavior. Since the field we are interested in is ethics or morality, it is a matter of understanding how and what emotions influence our moral judgments—chapter 7 deals with how humans create counterfactual scenarios. The emphasis focuses on the relationship between ethics and counterfactual imagination. This relationship concerns our capacity to generate possible worlds and how these are generated from inhibiting certain negatively charged emotions such as regret. Summarizing the most relevant points of this chapter:

1) Emotions play an essential role in human decision-making, but a science that measures them systematically and rigorously is still a pending and challenging issue.

2). The challenges and limitations faced by the two initially seemingly antagonistic emotion theories could be clarified with the help of the goal-driven XAI research domain.

3). Since this research work is oriented to the responsible behavior in a human-AI interaction or instead in a joint work between the human agent and the artificial agent, Antonio Damasio's (1994) somatic marker hypothesis has been taken as a starting point, according to which the emotion of regret as a secondary emotion guides the deliberative processes.

The principle of responsibility here is related to the capacity to experience regret. Therefore, being unable to experience this emotion implies not feeling responsible, which carries an essential ethical and moral burden that should be studied in depth.

This counterfactual emotion has been chosen as an object of manipulation to understand the role it plays in achieving what is known as the principle of responsibility. However, to establish why regret is sometimes inhibited, it is necessary to understand what underlies this phenomenon. For our framework, the first thing to do is to identify patterns or rules in how humans deliberate about a change in future behavior by creating counterfactual scenarios where the moral or ethical aspects are present. So, what effect does regret have on human behavior, and what elements constitute these experiential contents? In the following chapters, we return to this point, one of the multi-ethical framework's cornerstones.

CHAPTER 7

THE ETHICS OF COUNTERFACTUAL IMAGINATION

7.1. The counterfactual imagination and its ethical implications

7.1.1 Amplifying human and artificial intelligence through imagination

If one focuses on analyzing, as has already been attempted in the previous chapters of this dissertation, the kind of data that AI systems are currently working with, we can conclude that the information or data is almost entirely statistical. In other words: "Learning machines improve their performance by optimizing parameters over a stream of sensory inputs from the environment" (Pearl, 2018). However, if we want to at least design AI systems capable of replicating aspects of human intelligence, we have to move beyond finding correlations between data. For this purpose, Mahadevan (2018) proposes in his paper *Imagination Machines: A New Challenge for Artificial Intelligence* to design imagination machines.

However, endowing machines with a component of imagination does not seem to be a trivial matter, as two globally known anthropologists, N. Harari and S. Mithen, agree that the decisive element that made it possible for Homo sapiens ancestors, the ability to achieve global dominion about 40,000 years ago, was their ability to choreograph a mental representation of their environment, interrogate that representation, distort it by mental acts of imagination and finally answer: What if I had acted differently?

Even Einstein (1931) states that "imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world" (Malone, 2022). But then, is some concrete reasoning enhancing this hallmark of human intelligence? Pearl states, "Imagination is a hallmark of counterfactual and causal reasoning" (Pearl, 2009). But let us then look into the possibilities that imagination opens up for us at more concrete levels.

7.1.2 How does the imagination materialize?

If we return to Mahadevan and his theory, the breakthrough in AI could come from our ability to design imaginative machines. According to this author, the first thing to do is to understand the difference between data science and imagination science.

For imagination science, the latter extends to realms far beyond the former. This definition means that imagination focuses on the problem of generating samples that are "novel," "meaning they come from a distribution different from the one used in training" (Mahadevan, 2018).

In this way, the science of imagination would aim to cover topics such as the problem of causal reasoning to uncover simple explanations for complex events and use analogical reasoning to understand novel situations.

And if there is one thing that worries humans, it is seeking causal explanations because they want to understand the work in superficial "cause-effect" relationships.

However, if we want to understand the fundamental mechanisms of causal relationships, the process is, for the moment, unknown. In addition, as Kahneman and Tversky have long since demonstrated in their empirical studies, human decision-making does not conform to the maxims of expected utility theory.

A representative example of this phenomenon and its link to imaginative concerns is playing the lottery: "Year after year, in state after state, millions of Americans buy lottery tickets because they can "imagine" themselves winning and becoming rich, despite the vanishingly small probability of winning.

For many humans, imagination, in this case, misguides their actions into violating the principle of maximizing expected utility" (D. Kahneman & Tversky, 1979).

Nevertheless, isn't it precisely this ability to imagine that allows us to move in other directions, to take other directions in the future? Isn't that how children learn?

This anecdote that Mahadevan uses in his research paper about a friend's daughter, a three-year-old, expresses this learning model very well:

> "Not long ago, for example, while sitting with me in a cafe, my 3-year-old daughter spontaneously realized that she could climb out of her chair in a new way: backwards, by sliding, through the gap between the back and the seat of the chair. My daughter had never seen anyone else disembark in quite this way; she invented it on her own and without the benefit of trial and error or the need for terabytes of labelled data" (Mahadevan, 2018).

What is relevant is the purpose she gives to the chair, although if an adult plays with a chair, the adult would be classified as having a mental disorder or be an artist because playing with a chair like a child does not relate to the common sense of an adult. While this imaginative example of a three-year-old girl demonstrates the extremely versatile competence in comprehending the world in all its multi-modal richness and dimensionality, by contrast, the major AI milestones have focused on solving particular problems. Thus, advances in areas such as computer vision, image recognition, and speech recognition, among others, according to Mahadevan, are focused "on the ability to label a particular object or scene (or transcribe a given dialogue), where the emphasis is on expert level ability given a statically defined task" (Mahadevan, 2018).

Recent studies have addressed the importance of imaginative causal reasoning in enabling neural net approaches to learn more effectively without labels. However, the imaginative flexibility with which children interpret objects as all kinds of data, including emotional data, gives them a qualitative value far from the AI systems created to date.

So, any object can be the object of something else. Mahadevan says, "To a child, a chair may serve as a hiding place by crouching under it, or a stool, to retrieve another object placed beyond reach on a high table" (Mahadevan, 2018).

In psychology, this phenomenon receives the name of "affordances," a term coined and defined by James Gibson as follows: "An object is perceived in terms of actions it enables an agent to do, and not purely in terms of a descriptive label" (Gibson, 1979; Greeno, 1994; Nye & Silverman, 2012).

In Norman's words, affordances "are also essential to the design of everyday appliances" (Norman 1999).

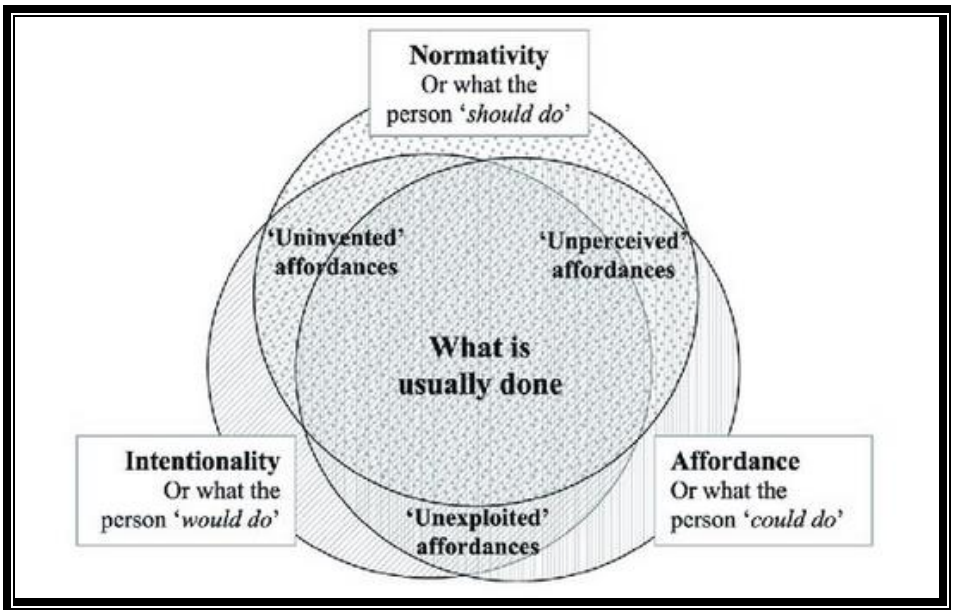


FIGURE 32 | A sociocultural model for an affordance theory of creativity (Piccardo, 2017)

This concept of affordances in this research aims to recognize and exploit the contents of our decision-making since physical objects can play a different role in our daily lives and mental objects, such as the choices we make or the thoughts and feelings that give rise to counterfactual thoughts. This way of proceeding by enhancing the affordances (Gibson, 1979) in imaginative perception (Mahadevan, 2018) could be a way to overcome some of the limitations mentioned in the previous chapters. On the one hand, it has to do with the challenges facing the field of ex-plainability.

On the other hand, we cannot understand others' decision-making through our restrictive ways of knowing and understanding what others think and feel in a particular context.

Before moving on to the next section, which focuses on how imagination in the form of counterfactuals materializes in our decision-making and the ethical implications it raises, it is worth noting that Mahadevan argues in his paper about the concept of “affordance” because it plays a central role in connecting objects with the actions, they enable agents to undertake.

In addition, the computation of affordances seems to be a vital aim in developing current work in deep learning for computer vision. Although this is not the focus of the thesis, it is worth noting that this type of machine learning based on imaginative perception accelerated a project called ALVINN, in which machine learning based on the observation of human driving was replaced by a simple causal model based on imagining hypothetical driving situations from each real experience (Pomerleau, 1989).

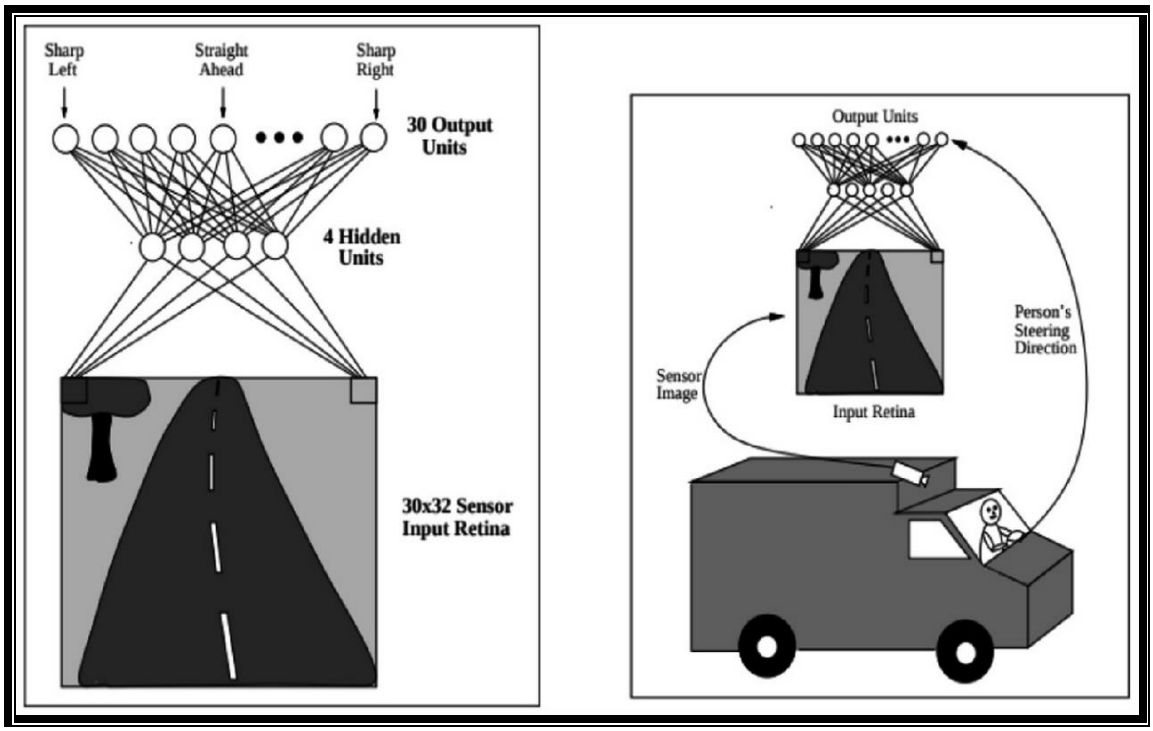


FIGURE 33 | Neural network architecture and setup on ALVINN (Pomerleau, 1989)

Although this type of technology or learning methodology is in its infancy, it may become very relevant in understanding and expanding our knowledge of other perceptual skills, such as interpreting speech intonations and emotions and body gestures, always considering the particularities of each element. Or, in Feldman Barret's (2017) words: "the variation is the norm," at least until proven otherwise.

7.1.3 Counterfactual imagination as a form of causal inferencing

As Kahneman, Slovic, and Tversky put it: "People often imagine how things could have turned out differently, especially after a bad outcome." In this way, "if only…" and "what if…" seem to be almost part of our daily routines" (S. and T. Kahneman, 1982).

However, are we aware of this? Moreover, do we use this kind of counterfactual imagination to understand what others are trying to tell us? Or do we imagine viable solutions to problems that others pose by actively listening and learning about their mistakes? Thoughts about what could have happened are often invoked to explain the past, sometimes to excuse and justify it, and other times to allocate fault, responsibility, and blame (Byrne, 2016; Byrne & Timmons, 2018; Markman, Mizoguchi, et al., 2008).

However, there is more: imagining alternative scenarios to reality helps formulate future intentions for improvement (Epstude & Roese, 2011; Smallman & Mcculloch, 2012). These mental simulations or counterfactual thinking about things that could have had a better outcome provide us with "a roadmap for change" (Byrne, 2020).

In addition, the types of alternatives we imagine about an action and an outcome could be mediated or initiated by causal information and other types of data, such as intentions, beliefs, and emotions.

7.1.4 Relationship between moral judgments and imagined alternatives

Another issue directly related to this research is the one addressed by McCloy and Byrne, and it refers to the impact that this creation of alternatives to reality has on moral judgments, as one of the imperative aspects she highlights is the allocation of causal responsibility among several candidate causes.

Hence, a pivotal discovery is that the perceived contribution of a candidate cause can be amplified or diminished by an imagined alternative (McCloy & Byrne, 2002).

Thus, we can amplify a causal judgment when the imagined alternative offers a different outcome. Conversely, we can also obtain the opposite, diminishing a causal judgment when the imagined alternative reality offers the same result.

Such mental simulations of causal moral inferences open the door to understanding the nature of the human imagination in social interaction (Byrne, 2020). Therefore, Byrne (2020) poses the following question: “Is the effect of the counterfactual imagination on moral judgments good or bad?”

This answer is followed by new questions because, in most cases, people make inferences not only about who caused the outcome but also about the intentions, desires, and knowledge of the person who caused it (Alicke et al., 2008; Lagnado & Channon, 2008; Malle et al., 2014).

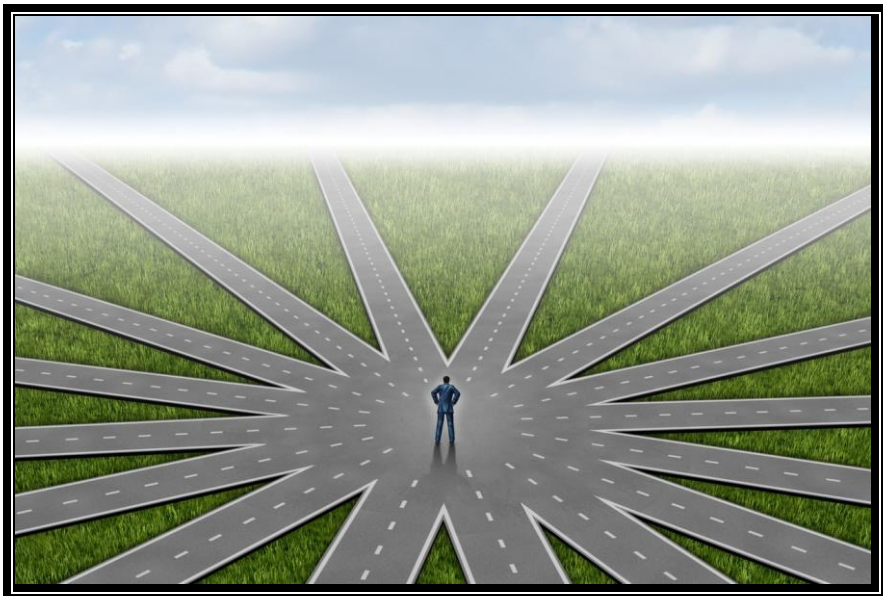


FIGURE 34 | Counterfactual imagination and moral judgments / Pixabay

Hence, the conclusions reached by Byrne in his study on the impact of counterfactual imagination in altering the way morality is perceived is that:

> "Counterfactual thoughts provide a unique learning mechanism, helping them to avoid past mistakes and to work out how to prevent bad things from happening again in the future. Yet the creation of alternatives to reality is disciplined, and people show significant regularities in what they change in their mental representation of reality to create an alternative. The bidirectional effects of counterfactual thoughts on moral judgments provide some corroboration for the suggestion that imagination of alternatives to reality, dramatic and powerful though it may be, relies on similar processes as other sorts of thinking" (Byrne, 2020).

To these statements made by Byrne (2020), this research adds and focuses more on how certain emotions, in our case, how regret work as a critical mechanism for identifying that a better outcome would have followed from a different choice leads to a different choice next time.

In other words, it is not only counterfactual thoughts that play a significant role as a learning mechanism and moral judgment makers. In its counterfactual variant, regret also promotes the skills necessary to detect and improve human errors. This research focuses entirely on regret because of its complexity and because different disciplines perceive it as a mechanism that promotes a change of direction in future decisions.

However, before focusing on regret, it is crucial to understand the different uses of counterfactual reasoning. Thus, we present first the framework designed by Pearl and Mackenzie (2018), with their three-level methodology that gives a unique position to counterfactual thinking (Pearl & Mackenzie, 2018). Furthermore, this methodology is the backbone of our multi-ethical interdisciplinary theoretical framework. Then, in the next section, we focus specifically on counterfactual explanations and how they could be implemented in Explainable AI (Byrne, 2019). The last part of this thematic block describes some limitations (Kasirzadeh & Smart, 2021) of the current approaches to counterfactual explanations in Explainable AI and issues related to fairness in machine learning.

7.2 The Ladder of Causality: chaired by counterfactuals

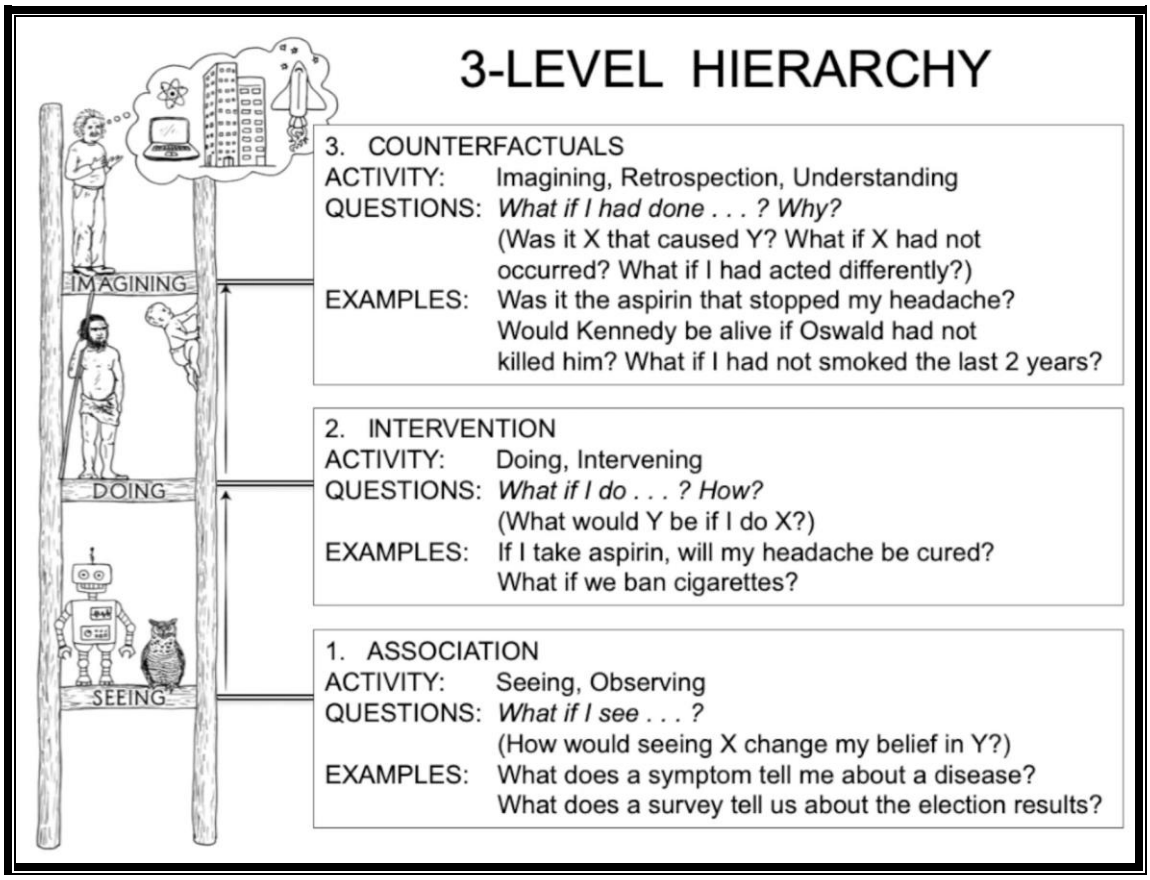


FIGURE 35 | The Ladder of Causation / Maayan Harel (Pearl & Mackenzie, 2018)

If we analyze the world, we realize that there are more than dry facts (what we call data today) because what is interesting is how "an intricate web of cause-effect relationships glues together these facts" (Pearl & Mackenzie, 2018).

In addition, these authors add that causal explanations, not dry facts, should be the "bulk of our knowledge" and, simultaneously, the cornerstone of machine intelligence.

Although there have been many theories that have set out to answer this challenge, as shown at the beginning of this chapter, one relevant theory has been that of the historian Yuval Harari:

> "This ability to imagine things that did not yet exist was the key to advancing civilization, allowing them to communicate better. Before this revolution, they could only trust people from their immediate family or tribe. Afterwards, their trust extended to larger communities, bound by common fantasies (for example, belief in visible yet imaginable deities, in the afterlife, and the divinity of the leader) and expectations" (Harari et al., 2011)

Three ways of organizing knowledge: the ladder of Causation

In their book, Pearl and Mackenzie (2018) discuss that causal learners must master at least three levels of cognitive abilities: seeing, doing, and imagining. Since this framework is a fundamental element of the proposal presented at the end of this chapter, it is essential to remark that it has been detected that there is not a significant difference between level two and level three of this ladder of causality, but that in the proposed framework, more emphasis is placed on what differentiates level two and level three.

Thus, the first level of this ladder, highlighted by Pearl and Mackenzie (2018), corresponds to seeing and observing. This level is shared not only by humans but also by animals, who also possess these abilities. This level aims to detect regularities in the environment.

The second level refers to interventions or, in other words, to acting or doing; it is also about making predictions of the effects that could be achieved by deliberately altering our environment to gain our desired objective. An essential element at this stage is that it is about using tools, and in this case, intentionally and not accidentally. However, tool users do not necessarily have an advanced theory of why the tool works and what to do when it does not.

Thus, to reach the third level, you need to have reached a level of understanding that allows you to imagine. Pearl and Mackenzie (2018) argue that we cannot scientifically prove that these levels led to significant milestones in human history, such as the revolution in agriculture and science. However, they hypothesize that "one can prove mathematically that the three levels differ fundamentally, each unleashing capabilities that the one below it does not" (Pearl & Mackenzie, 2018). This dissertation states that the third level can only be acquired when one has what Paéz (2019) has termed an objectual understanding of the larger body of knowledge to which the specific object of understanding belongs. This example establishes a basis for this concept: without a basic understanding of the body's anatomy, biology, chemistry, and the behavior of the lungs, a layperson could not answer these kinds of counterfactual questions, which is a sign of understanding.

Accordingly, the framework presented by Pearl and Mackenzie (2018) takes inspiration from a model of Alan Turing, who proposed classifying cognitive systems in terms of the queries they can answer. For Pearl (2018), how Alan Turing approached the science of causality is very concrete and leaves aside unproductive discussions that lead to nothing; thus, while Turing's classification works on a binary classification - humans and non-humans - Pearl and McKenzie's (2018) classification is based on three levels of queries, which we summarize in detail below. As discussed above, this framework is the basis of our multi-ethical proposal.

7.2.1 The first ladder: finding regularities

The first level of the ladder corresponds to pattern matching. On this rung, we look for regularities or patterns through our ability to observe and see. Examples of these animal abilities include when a rat moves and figuring out where the rodent will likely be a moment later. It is also what a computer Go program does when it studies a database of millions of Go games to figure out which moves are associated with a higher percentage of wins.

This approach can be summarized as one event associated with another if observing one changes the likelihood of observing the other. Predictions at this level are based on passive past observations and characterized by questions such as: "What if I see…?

To explain this level, the authors use this example, which is the one used at all levels of the causality ladder.

Remember that these three levels form a metaphorical framework, although they can be extrapolated to different contexts, as mentioned above. The example of these authors proposes to imagine a sales manager in a department who has to answer the question: “How likely is a customer who bought toothpaste also to buy dental floss?” In particular, the question can be solved by analyzing the behavior patterns of all customers, selecting only those who bought toothpaste, and focusing on the second group, computing the proportion who also bought dental floss. This proportion, also known as a “conditional probability,” measures (for extensive data) the degree of association between “buying toothpaste” and “buying floss.”

In recent years, statisticians, especially with the new machine learning techniques, have successfully developed methods to reduce the large body of data to identify associations between variables. Some associations may have more obvious causal interpretations, while others do not. However, statisticians or ML system developers cannot tell the cause and effect of these types of associations.

In fact, from the sales manager’s point of view, it is not relevant either, as they work under the objective of good predictions need not have good explanations. Again, Pearl and Mackenzie (2018) use a picturesque example: “The owl can be a good hunter without understanding why the rat always goes from point A to point B.” In addition, they put deep learning algorithms on this level of the ladder because these models continue to operate in associative mode. They work by a stream of observations to which they attempt to fit a function. This method is similar to how statisticians try to fit a line into a collection of points. What is peculiar or outstanding about all this is that neural networks have added more layers to the complexity of the fitted function, but raw data still drives the fitting process, and this starting point marks the path from our beginning to our end. In addition, as Pearl and McKenzie (2018) say, these systems continue to improve in accuracy as more data are fitted, but they do not benefit from the “super-evolutionary speedup.” If autonomous car programmers want an autonomous car to react differently to new situations, they must explicitly add these new reactions. This lack of flexibility and adaptability makes it inevitable that existing AI systems are part of the first level of the framework proposed by these authors.

7.2.2 The second ladder: prioritizing knowledge

These limitations bring us to level two of this ladder of causation, which introduces changes to the world. We continue with the example of toothpaste, and on this level, a typical question is: "What will happen to our floss sales if we double the price of toothpaste?" This level needs new knowledge because it is not based on just seeing what it is but on what we need to change to gain the desired output.

But then, why can't we answer our floss question just by observation? For example, the scenario is that toothpaste may have been in short supply, and every other store also raised its price. This new scenario needs a deliberate action plan that approves a new price for the product according to market conditions.

We can make better predictions with historical data on past market conditions. However, what data do we need, and how do we get to it? One way to predict such an intervention is to conduct experiments under controlled conditions. For example, in the toothpaste case, the sales manager can develop a model of consumer behavior, including market conditions. If you do not have the necessary data for each factor, it can be collected data on enough key surrogates to make the prediction.

Pearl and Mackenzie argue that a strong and effective causal model can enable us to leap from purely observing data to intervening. According to these authors, deep learning models have not caught up and, in their words, "will never be able to answer questions about interventions, which by definition break the rules of the environment the machine was trained in" (Pearl & Mackenzie, 2018).

With this constraint is how we came to define the second query of the second ladder of Causation: "What if we do…?" or in another form to express the same: "What will happen if we change the environment? At this level, the question of how also becomes relevant. So, in the example with toothpaste, the scenario could be that the manager imagines how toothpaste can be sold in case there is a surplus in the warehouse. The essential questions would be: "How can we sell it?" and "What price should we set for it"?

These questions relate to interventions that we operate mentally before deciding what to do in real life. At the core of our daily decision-making operates at this level. If, for example, we have

pain in some part of our body, we try to intervene on a variable that would be "pain relief" and move to a state of no pain.

However, although at this level of the ladder, the level of reasoning about what to do in a given context is a crucial rung, it does not answer questions such as: why did the pain go away? What would have happened if I had not taken the painkiller? Is it because I slept and ate better? For Pearl and Mackenzie (2018), the answers to these questions we can formulate as counterfactuals are at the top of the ladder.

7.2.3 Why are counterfactuals the kings of the Ladder?

Counterfactuals allow you to go back in time and change the conditions in the future in case the same or similar event happens again. These authors argue (2018), "No experiment in the world can deny treatment to an already treated person and compare the two outcomes, so we must import a whole new kind of knowledge" (Pearl & Mackenzie, 2018).

The limitation of this approach is that the relationship of counterfactuals to data is problematic because, as Pearl and Mackenzie (2018) put it, by definition, data are facts. Counterfactual reasoning might not be able to tell us what could happen in an imagined scenario where specific observed facts are flatly negated.

These authors add: Finding out why a blunder occurred allows us to take proper corrective measures in the future—finding out why a treatment worked on some people and not on others can lead to a new cure for a disease—answering the question: what if things had been different? It allows us to learn from history and the experience of others. Thus, counterfactuals are at the top of the Ladder in Pearl and Mackenzie's (2018) framework because they were once the mental tool that allowed us to survive, adapt, and ultimately take over.

This advantage we gained from counterfactual reasoning was the same as today: "flexibility, the ability to reflect and improve on past actions, and, perhaps even more significant, our willingness to take responsibility for past and current actions" (Pearl & Mackenzie, 2018). In short, we need a theory to predict what might happen in situations we have not envisioned before.

The strengths of this framework, as set out by Pearl and McKenzie (2018), are inexhaustible. However, limitations and criticisms of how counterfactuals can be modeled and used in AI systems are also becoming visible. The following section is about these issues.

7.3 The Choice of counterfactuals: a world of infinite possibilities

The fact that counterfactuals, or rather specifically, counterfactual thinking, are at the top of the framework set out by Pearl and Mackenzie (2018) does not mean that counterfactuals are always valuable in the general sense of the term. This phenomenon is due to humans' cognitive limitations when formulating counterfactual thoughts and their application in different areas of AI, which are related to questions of algorithmic fairness and social explanations, which are discussed in this section before moving on to the following thematic block of this research.

In search of the most useful counterfactuals for Human-AI exchange

Byrne focuses on the crucial role that counterfactual reasoning has started to play in different AI applications in her paper entitled *Counterfactuals in Explainable Artificial Intelligence (XAI: evidence from human reasoning*. However, she calls attention to the importance of understanding the cognitive mechanisms that lead to the formation of counterfactuals in humans, as not all counterfactuals are equally helpful, and even less so in specific AI applications (Byrne, 2019).

AI's interest in counterfactuals covers subgoal construction to identifying planning failures, from fault diagnosis to determining liability (Ginsberg, 1986; Halpern & Pearl, 2005). The most notable contribution of counterfactuals in AI now could be in explainable AI. As these systems are increasingly used to guide complex tasks such as high-risk decisions, and their behavior is unintelligible due to their complex internal structure (Weld & Bansal, 2018), developing explanatory strategies becomes more than necessary. Although Explainability or Explainable AI is described and analyzed in depth in Chapters 4 and 5, a key concept is introduced here: trust.

Thus, one way to increase trust between these opaque systems and their human users is to create dynamic explanations or ad hoc explanations of the decisions they make so that they are understandable to humans (Biran & Cotton, 2017).

In this way, an interpretable model allows humans to simulate some aspects of the system mentally and understand the causes of these systems' decisions (Hoffman et al., 2018).

According to Miller (2019), this method enables the user to consider contrastive explanations and counterfactual analyses, such as why one decision was made instead of another, and to predict how a change to a feature affects the system's output.

However, and this is the first critique, in this case, launched by Byrne (2019), the number of counterfactuals that can be generated to explain an event is potentially limitless, and it is a non-trivial problem to identify which counterfactuals best facilitate the construction of an explanatory model. Thus, the following paragraphs concentrate on analyzing and describing the work of Byrne (2019), which includes a set of experiments from psychology and cognitive science on the capabilities of human reasoners to comprehend and reason from counterfactuals and the sort of counterfactuals they create.

Depending on the outcome of these experiments and, hopefully, others, how counterfactuals continue to be used in AI will be based on even more rigorous and objective studies than those carried out so far because of the side effects that ignorance about this infinite world of possibilities may have.

7.3.1 Human reasoning in the formation of counterfactual scenarios

Byrne's (2019) first point of analysis refers to the structure of counterfactuals. In this case, people tend to create counterfactuals about things or aspects of reality that could have been different by adding something new to what they already know and not by removing something from the situation or thing. The process of imagining a new scenario seeks a better outcome than one that is worse. These two forms of reasoning are additive and subtractive counterfactuals.

Let us use the examples Byrne gave to understand this substantial difference. Imagine that we are trying to understand the decision of an autonomous vehicle to swerve to avoid hitting a pedestrian and are considering why it did not brake instead. The scenario is wrong because the car hit a wall, and its passenger sustained minor injuries. In this situation, an additive counterfactual could add new information about the simulation of reality, such as: "If the car had detected the pedestrian earlier and braked, the passenger would not have been injured" (Byrne, 2019).

In contrast, a counterfactual explanation is subtractive by deleting something already simulated about the factual reality, such as: “If the car had not swerved and hit the wall, the passenger would not have been injured. Consequently, “People tend to create additive counterfactuals more than subtractive ones” (Roese & Epstude, 2017).

Byrne’s (2019) contribution is that the type of counterfactuals that people construct affect their subsequent reasoning. Thus, while additive counterfactuals help creative problem-solving, subtractive ones aid logical reasoning.

In another experiment, adults were asked to think about a bad thing that happened to them in the past and one group was instructed to generate an additive counterfactual in the form of the following sentence stem: “If I had not… then the outcome would have been better/worse”.

In this case, the additive counterfactual control group performed better than the subtractive one on subsequent creative problem-solving tasks, such as generating creative uses for an object. However, the subtractive counterfactual group performed better than the additive group on logical reasoning tasks, such as syllogistic inferences (Markman, McMullen, et al., 2008).

Following these results would be more fruitful in an exchange between a human agent and an AI to track this explanatory exchange. Moreover, as Byrne concludes, “the provision of additive counterfactuals may also promote further creative problem-solving about other aspects of the system” (Byrne, 2019).

Thus, in this line, most people imagine how things could have been better than worse (Markman et al., 1993). A representative example is when people simulate how things could have been better after a bad outcome, such as an injury to a passenger.

This type of counterfactual reasoning is called an upward comparison. The opposite case is called a downward comparison, and the sentence that reflects this with the above example is: If the car had served in the other direction into oncoming traffic, the passenger would have been killed.

Considering that counterfactual reasoning leads most people to execute upward comparisons, it would be interesting to know more about downward comparisons, as they can be beneficial in

certain circumstances. Byrne (2019) rightly points out that implementing the ability to provide explanations in an AI agent by comparing how things could have turned out better or worse.

In this case, it is my reflection: If people ask themselves after a bad result how a situation could have been solved in a better way for the future, why do we tend to repeat the same mistakes as individuals, as groups, and as nations?

Furthermore, such counterfactuals focused on improving a future outcome followed by a past bad outcome affect future intentions. This phenomenon is also related to the diverse ways of learning or, in other words, "counterfactuals about how an outcome could have been better affect intentions for the future, unlike counterfactuals about how an outcome could have been worse" (Byrne, 2019).

In the same vein, we have the conclusions obtained from the experiments performed by Markman and Co. (2008): "The subjects formulated intentions to do so in the future, and their subsequent performance improved" (Markman, McMullen, et al., 2008); or, according to Roese and Epstude (2017), "the counterfactual offers a roadmap to transition from the current situation to a different future one" (Roese & Epstude, 2017).

7.3.2 The creation of counterfactuals and the effects of their affective charge

Now, counterfactual reasoning and its affective charge are crucial points to consider for this dissertation. Hence, counterfactuals about how things could have been better help people prepare for the future, but they come at an affective cost, according to studies by Kahneman and Tversky (1982). That means these counterfactuals amplify negative emotions such as regret or guilt.

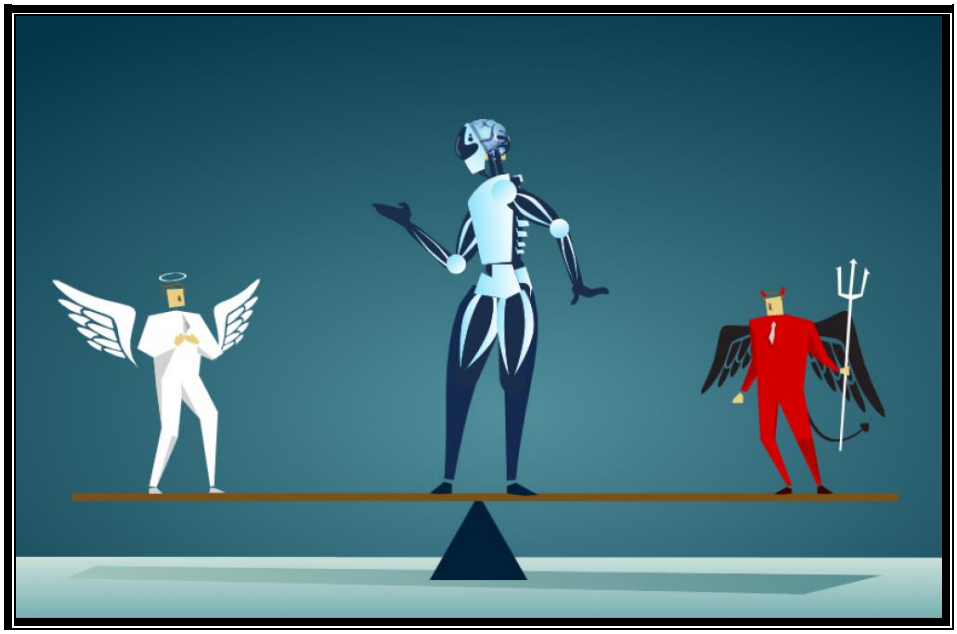


FIGURE 36 | Inhibition of counterfactuals due to their negative emotional charge / Pixabay

The question here is: Who is willing to pay this emotional price? Are we trained in this matter? Are we even aware of what this means and implies for learning processes? After several experiments, people tend to create counterfactuals about how things could have been worse when there are few opportunities for future preventative actions, and they can deflect negative emotions (Beike et al., 2008).

In fact, "people sometimes opt to inhibit counterfactuals" (Byrne, 2019). Furthermore, they decided not to be informed about the outcomes of unchosen options more often after significant losses than small ones (Tykocinski & Steinberg, 2005).

This challenge is the tipping point and the one that matters for connecting counterfactuals to ethical decision-making. In constructing counterfactuals about things that could have been even worse, there is essentially an inhibition of counterfactuals, which helps people feel better at the cost of complacency.

Is it this attempt to repress emotions that initially seem negative one of the main impediments that characterize human decision-making not to repair its mistakes?

So, let us imagine if this system's explanation brings out counterfactual emotions such as regret in an exchange with an AI agent. Should the user dismiss this counterfactual explanation for not granting them the emotional experience of "a self-administered reward" be ignored? Kahneman, Sibony, and Sunstein (2021) define this concept as "a satisfying emotional experience, a pleasing sense of coherence, in which the evidence considered, and the judgment reached feel right" (D. Kahneman et al., 2021).

Suppose this hypothesis is accurate, and this research supports it. In that case, the explanations that AI agents provide us with, even if they are objective to reality and offer reliable information to improve our future strategies and decisions, might be discarded because we do not want to experience "mental regrets," or we cannot experience any regret feeling. The easy way is not to consider this advice from the AI systems because of the affective or emotional price to pay, which is to experience regret in popular wisdom: it is not worth it.

This aspect of counterfactual emotions directly relates to building trust between humans and artificial agents and the model of reality we follow or want to build. We return to this topic in the following section, particularly when analyzing the experience of regret and its impact on learning and the future potential to repair damages and mistakes made in the past.

Therefore, and to finish this point, it is worth saying briefly that what is fair and trustworthy does not have to be that explanatory story that an individual, group, or nation, from the manipulation of particular objects or data from the reality of the world to reach that state of "self-administered reward" (D. Kahneman et al., 2021).

7.3.3 Counterfactuals versus causal explanation to make sense of the world

As mentioned above, people tend to explain past outcomes and predict future ones by identifying the relationship between events. Most people tend to proceed with this exercise by searching for cause-effect or reason-action relationships. However, if people imagine how things could have been otherwise, their counterfactual scenarios may be affectively charged again.

In this case, there are feelings of blame and fault. So, counterfactuals and causal inferences have long been considered two sides of the same coin. But what do we mean by this affirmation? An example always makes it easier to understand.

The moment people know that an alternative would have changed the outcome of an action, their judgments about the causal relationship between the antecedent and the outcome are amplified; however, when they know that the alternative course of action produces the same outcome, their judgments of the causal relationship between the antecedent and the outcome decrease.

Although counterfactuals amplify causal judgments, causal explanations refer to different causes because events often have several causes, some of which preempt or supersede others (Halpern & Pearl, 2005).

It is more than well-studied again, as the authors Kahneman, Sibony, and Sunstein (2021) argue that our effort to look for the cause of an event is to create a story that makes sense of our world. This phenomenon of looking for the causes that produce an outcome or conclusion is called causal thinking, and the reasons that support them are called explanations.

Further, according to the authors of *Noise*, "once something has happened, causal thinking makes it feel entirely explainable indeed predictable" (D. Kahneman et al., 2021).

This way of understanding how events occur and the explanations that support them seem to go back to the paradigm that governed classical science until the seventeenth and eighteenth centuries, which is the famous claim that "given enough facts, we could not merely predict the future but retrodict the past" (Prigogine & Stengers, 1984).

In this way, explaining events once they have happened makes us think we could have predicted them. This process of understanding reality makes us comprehend it and is also considered a backward-looking process.

As quoted by the authors of *Noise*: "Our sense of understanding the world depends on our extraordinary ability to construct narratives that explain the events we observe," to which they add: "When the search for an obvious cause fails, our first resort is to produce an explanation by filling a blank in our model of the world. This is how we infer a fact we had not known before" (D. Kahneman & Tversky, 1979).

Thus, when a human agent justifies or gives explanations for a specific choice, what the human agent is doing is producing a post hoc story that supports the decision taken.

This reasoning follows the same assumption Sperber and Mercier made in their book *The Enigma of Reason: A New Theory of human understanding*. They state, "We produce reasons to justify our thoughts and actions to others and to produce arguments to convince others to think and act as we suggest" (Mercier & Sperber, 2017).

In the same vein, we have the thoughts of Nicholas Taleb, who explained in his book titled *The Black Swan: The Impact of the Highly Improbable* (2007) the idea that while the future cannot be predicted, there are several ways of reducing uncertainty when it comes to forecasts.

These are the artistic and scientific paths, together with the techno-scientific axis. Taleb even speaks of a fourth path, which is the media.

Taleb says all these initiatives are fueled by the human need to reduce and condense everything around us because it comprises millions of details. In the author's words: "Without these channels, one is condemned to navigate in a universe of absolute and unbearable uncertainty" (Taleb, 2007).

Moreover, these pathways would be narratives materialized in political, scientific, economic, and sociological discourses to spare us from the world's complexity and protect us, as Taleb (2007) says, from its "randomness."

Taleb's theory about reducing uncertainty leads to the so-called "representation problem, which leads us back to Pearl and Mackenzie, who ask somehow rhetorically one of the main aspects that this dissertation is trying to answer: "How do humans represent possible worlds in their minds and compute the closest one when the number of possibilities is far beyond the capacity of the human brain?" (Pearl & Mackenzie, 2018).

FIGURE 37 | Current solution for dealing with the representation problem / dailysabah.com

This process of understanding both causal relationships shown in the form of causal explanations fits with McEleney and Byrne's studies because when people think about an outcome, they create that story to make sense of their world, much more so than creating counterfactual thoughts that reference an imaginary alternative that in many cases would likely fit more of a model of the world or even improve aspects of the world (McEleney & Byrne, 2006).

But then, in what way would an AI agent be valuable to us? If the agent offers a counterfactual about an antecedent and its outcome and causal explanations, understanding these outcomes and future applications could help move away from a mechanistic worldview where the past equals the present and the future.

A few strokes about counterfactual content

Another point analyzed by Byrne (2019) is that of counterfactual content. In her analysis, she concludes that people show regularities in what they select to mutate in their representation of reality.

The first aspect concerns exceptions and this mechanistic approach to how the world works. Thus, any unusual event tends to be a candidate for modification.

If we take this approach from an emotional perspective, and this is an argument that this dissertation points to, it would have to do with the instinct to be afraid of novelty, and anything that breaks with established habits or rules is considered an enemy. This connection is worth studying because exceptions are categories also contained in reality, and any way of denying them is to want to modify or deny aspects that may be relevant for (ethical) decision-making.

The second aspect related to the first is the tendency to create counterfactuals by changing a controllable event we can manipulate. In other words, similarly, "people tend to create counterfactuals that changed events outside their control" (Girotto et al., 2007), doing them something predictable and controllable. While Byrne (2019) considers more aspects of the counterfactual content, the points above are the most relevant for this dissertation.

New causal-effects relations through AI counterfactual explanations

Hence, following this analysis, humans' shortcuts in creating counterfactuals must be explored in more detail. In addition, Byrne (2019) concludes in her paper that XAI can benefit from including the rich knowledge in cognitive science about human reasoners' cognitive capacities, enabling an agent to simulate the same sorts of alternatives to reality as a human. The assumption of this dissertation is different, and it can be formulated as follows: by better understanding the shortcuts people take when creating alternatives to reality, this fact can be solved by modifying or manipulating other elements of reality not considered by people so far. This approach can contribute to discovering new cause-and-effect relationships that differ from the common ones. Moreover, since the creation of counterfactuals by people is subject to a specific time-space situation, the simulated scenarios that an AI agent can offer would be richer by representing other worlds that people do not reach, based on the manipulation of quite different kinds of data.

These limitations in human reasoners' capacities are not always because of ethical and moral issues but simply because they cannot encompass a more profound knowledge of certain aspects of reality. This phenomenon relates to the problem of the representation of worlds (Pearl & Mackenzie, 2018) and, on the other hand, to the frame problem in humans, not only AI (Chow, 2013). Before proceeding to the next section of chapter 7, reviewing some of the current uses and misuses of counterfactuals in algorithmic fairness and social explanations is advisable.

7.4 Evaluating counterfactual explanations in (Ethical) Machine Learning

The use of counterfactuals in the machine learning community has increased enormously in the last few years for making high-stakes decisions with ethical and legal impacts in domains such as insurance, predictive policing, and hiring (Kasirzadeh & Smart, 2021).

In this context, AI developers have developed techniques to generate and evaluate counterfactuals in recent years. Thus, we have feature-based, prototype, example, and causal explanations (Guidotti et al., 2018; Joshi et al., 2019; Kusner et al., 2017; Lundberg & Lee, 2017). Although they may appear to be distinct explanations, they all come from two of the most relevant conceptual approaches for evaluating counterfactuals. One refers to the approach of the close-enough-possible worlds (Parry, 1973; Stalnaker, 1968). The other is the causal modeling approach (Halpern & Pearl, 2005; Spirtes et al., 1993). So, before assessing the limitations of both models for questions of algorithmic justice and social explanations, it is helpful to highlight what each focuses on.

7.4.1 The close-enough-possible worlds approach

According to the work of Lewis (1973) and Stalnaker (1968), which corresponds to the closest-possible-world view, a counterfactual can be expressed syntactically and semantically via a variant of modal logic for counterfactuals. Evaluating the counterfactual $X \; \square\!\!\rightarrow Y$ (if X had occurred, Y would have occurred) requires identifying a set of possible worlds in which X occurs. If Y also occurs in these possible worlds, the counterfactual $X \; \square\!\!\rightarrow Y$ is true. Let us use Kasirzadeh and Smart (2021) to illustrate this statement. “If Nora had not been Latina, she would not have been denied admission.” Therefore, does Nora’s being Latina “cause” denying admission? These possible worlds must be ordered in terms of comparative possibility similarity and closeness to the actual world (in which X occurs and Y occurs). As a result, if Nora's being Latina is the cause of her rejection in all the close-enough-possible worlds to the actual world in which she is not Latina, then she is not denied admission.

The point of criticism of this approach offered by Kasirzadeh and Smart (2021) is that evaluating counterfactuals requires an ordering of the possible worlds in terms of similarity to the actual world.

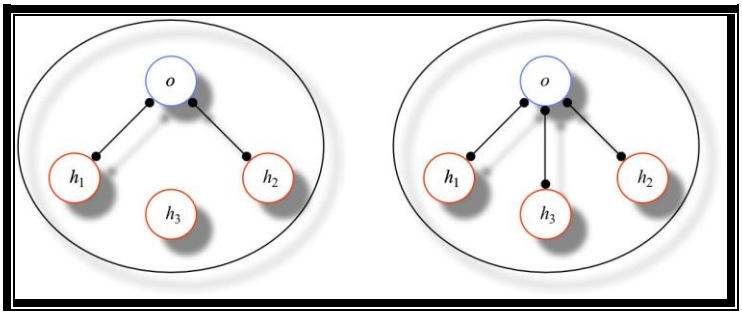


FIGURE 38 | The close-enough-possible worlds approach / (Stalnaker, 1968)

Therefore, one of the main weaknesses is the concept of similarity, as it is inherently vague, and as a consequence, we can order these possible words in quite diverse ways. "As a result, depending on the choices for the similarity criteria and order, we can obtain contradictory judgments about the truth or falsity of counterfactuals. Hence, the vagueness and the multiplicity of orderings pertain to the problems of using counterfactuals in machine learning" (Kasirzadeh & Smart, 2021).

7.4.2 The causal modeling approach

The second most commonly used approach to evaluate counterfactuals is a causal model that works as a symbolic tool for exploring the space of alternative causal hypotheses. According to Pearl (2009), from a causal modeling view, the world is described in terms of random variables and their values.

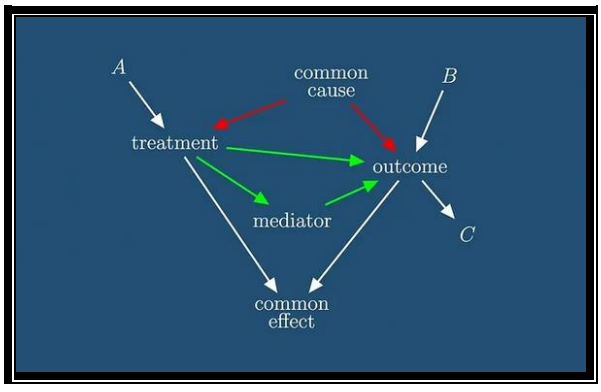


FIGURE 39 | The causal modeling approach / (Pearl, 2009)

Thus, a causal model comprises a set of exogenous and endogenous variables, their values, and a bunch of structural equations. With the help of a directed acyclic graph, the model can be graphically visualized. This graph facilitates cognitive thinking about potential causal sources, effects, and relations. In addition, according to the description of this approach by Kasirzadeh and Smart (2021), a node represents a random variable, and an edge between each pair of nodes determines a direct causal relationship between the corresponding random variables; for instance, X is a direct cause (parent) of Y is represented by X→Y. Further, nodes with no incoming edge are said to be exogenous. For Pearl (2009) and Spirtes (2000), finding causal relations via a causal model requires establishing well-defined connections between some aspects of the sample data and a causal model. In such a way, following the claims of Kasirzadeh and Smart:

> "To counterfactually think via a causal modelling approach in a specific machine learning domain requires an in-depth interpretation of the mapping of the random variables on the elements of the domain and the satisfaction of the causal assumptions (…), "if the domain of counterfactual thinking occurs at the level of the social world, we require an apt interpretation of the mapping of the random variables on social categories, the relationship between them, and the meaning of causal assumptions applied to the relevant categories" (Kasirzadeh & Smart, 2021).

In the case of this research, we are not dealing with social categories such as race and gender. Fortunately, these categories are already well established, and no one doubts that they intrinsically contain the relationship between what is just and what is not. One of the emerging themes in the criticisms of automatic decision systems is that they are not free of algorithmic biases. One of the most frequently cited is the gender biases in these systems. The criticism Kasirzadeh and Smart (2021) make of the limitations of these two more established conceptual approaches for evaluating counterfactuals could be transferred by analogy to other categories, where they offer one result or another depending on the data manipulated. Therefore, let us consider the weaknesses of the close enough possible worlds (Parry, 1973; Stalnaker, 1968) and causal modeling (Pearl, 2009; Spirtes et al., 1993) in the evaluation of counterfactuals used in "Ethical Machine Learning" as specified by Kasirzadeh and Smart, (2021).

7.4.3 Are the two available approaches fair enough for all kinds of data?

As discussed above, the category that Kasirzadeh and Smart (2021) have worked with is social; specifically, they both refer to social categories such as race and gender. However, there is a long debate in fields such as philosophy, sociology, law, and epidemiology, among others, about the social categories of race and gender. According to the authors cited above: “to counterfactually suppose a social attribute of an individual requires first specifying what the social categories are and what it means to suppose a different version of an individual with the counterfactually manipulated social property. This counterfactual question amounts to asking, “What if person X had not been” race Y” or “gender Z”?

In the case of this dissertation, we are not working on manipulating social categories. However, it is essential to understand what mechanisms create new social norms, in this case, and how a norm of recommendation emerges from the in-depth study of the emotion of regret and its relationship to responsible decision-making or behavior. Moreover, how can we measure if a person can have this ability or capacity to simulate regrets that have not yet occurred? Are we trained in this capability? If not, how can we acquire, measure, and evaluate it? To detect possible regrets, shuffle and set possibilities according to specific objectives. In this case, the objectives can be the ethical values set jointly in rules or values.

Nevertheless, as discussed in Chapter 3, it is necessary to rethink what ethical frameworks are appropriate for constructing an emerging norm based on principles or values.

In this way, the manipulation of an object, in this case, the category of regrets, pursues a twofold objective: 1) help to choose from a world of possibilities which worlds we want to represent through the selected data aligned with ethical frameworks or values 2) how our action or inaction in generating possible regrets (in the form of mental stimulation) could affect future scenarios.

But this way of understanding these “possible worlds” is not so much different from the assumption that Kasirzadeh and Smart discuss in their work: “The ordering of similar worlds faces severe problems because, for some ordinary counterfactuals, some irrelevant possible worlds end up determining the counterfactuals’ truth values. Also, depending on the possible worlds we choose, we might assign a different truth value to a counterfactual statement" (Kasirzadeh & Smart, 2021).

Ultimately, it depends on what kind of ontology, knowledge, and semantic load we take to apply this concept of "similarity." It is still a debate loaded with dualism and essentialism, as reflected in the previous chapters, for example, with the two theories about emotions. Thus, for Pearl (2018) to make sense of "similarity," we should rely on the fact that we experience the same world and share the same mental model of its causal structure. In opposition to this argument is the work of Harding, who suggests that different epistemic viewpoints might consider different ordering of possible worlds; "after all, humans differ extensively in the standpoints from which they observe the world, and these standpoints influence the formation of causal mental models" (Harding, 2007). In other words, what is inferred from this is that: "no causal model captures objective relations in the world" (Kasirzadeh & Smart, 2021).

So, "depending on the convenient assumptions for a causal model, X can be counterfactually dependent on Y in one model but not in another" (Halpern & Pearl, 2005). In this reasoning, we can conclude, like Moore and Nagel, that there is always a view from somewhere instead of a more objective and universal "view from nowhere" (Moore & Nagel, 1986) from which we can assess whether a counterfactual is assertible. And for this, I add that nothing is better than simulating possible worlds and variables that have been invisible until now.

CHAPTER 8

THE NORMATIVE VALUE OF ANTICIPATED REGRET

8.1 Identification of regret as a normative value

The previous section has focused on issues related to imagining scenarios in the form of mental objects that we could manipulate to obtain a different outcome from the past in the future. What is involved in counterfactual reasoning to evaluate future outcomes has been analyzed in depth, especially considering the shortcuts humans employ to use their reasoning skills to produce counterfactuals (Byrne, 2019; Tversky & Kahneman, 1986).

As mentioned above, a relevant aspect in the construction of counterfactuals is that when we construct them, we tend to leave aside those with a negative emotional charge, specifically those that involve the experience of regret. The paradigmatic essence of this emotion is the protagonist of the last part of this dissertation. In addition, it is one of the central pieces of the interdisciplinary theoretical framework explained in the following sections of Chapter 8.

Besides, regret also plays a beneficial role in understanding at both ontological and epistemological levels what information it provides us to improve our decision-making. This emotion connects with the past, present, and future. However, the highlighted element in this section is its anticipation in the form of mental stimulation and how the specific reinforcement of the prospective variant of regret influences the regulation of future behavior.

Therefore, this part of the thesis studies and reflects on the two dimensions of regret, one more focused on its cognitive aspects and the other on its affective component. So, if this emotion is essential in regulating our decisions, can we learn to train this ability to anticipate regret?

The answer is yes, and several studies in the child population already indicate that this emerging ability would positively impact decision-making, although it is not yet fully understood how. This dissertation tries to add possibilities for studying this dimension of regret. Specifically, we are dealing with how the anticipation of regret could be an ethical pathway to improve our decision-making strategies in trustworthy human-AI interaction.

8.1.1 Regret: feeling and doing what might be from the present to the future

Every day, we decide what to do at different times and with different things. Sometimes, we think about the consequences of our actions in the future. Other times, we function according to automatic patterns without considering alternative paths.

How do we decide which decisions are worthy of deeper reflection? Do we all do it the same way? Regret is, by definition, the emotion that regulates decision-making, especially regarding selecting actions that improve future outputs (Zeelenberg, 2018). However, the paradigmatic element of this emotion is not only focused on this emotion when looking back at decisions that went awry (Shani & Zeelenberg, 2007).

Regret is also a forward-looking emotion, and this variant of regret is also called Prefactual thinking (Sanna, 1996) when making crucial decisions. In particular, we focus here on this kind of regret because thinking or feeling it in a "simulated" way before a decision concerning future scenarios opens the door to changing our decision-making in various ways. But how can we define this anticipated regret or anticipatory regret?

Among other definitions, we can take this one: "When this type of regret is elicited, it can exert a strong influence on our behavior… More specifically, decision-makers tend to regulate their regrets, such as behave in a way that allows them to avoid experiencing this emotion in the future" (Zeelenberg, 2018).

Many studies and experiments at both the psychological and neuroscientific levels demonstrate how decision-makers try to regulate their regrets. However, the concept of anticipated regret as mental simulation is our object of study in this research.

As explained below, studies have already been conducted on how this ability drives the child population to delay rewards and make better decisions in the future. However, how this perspective variant of regret affects interpersonal and social decision-making remains a scientific gap that should be studied in detail. Furthermore, it is crucial to understand the retrospective and prospective variants of regret to understand common patterns and where they diverge. Therefore, a general review of the term coined to regret is made before turning to the anticipated regret.

8.1.2 Regret: understanding and imaging what went wrong across cultures

Shani and Zeelenberg are some authors who have researched this emotion more in-depth from a psychological point of view. We take their definition, who speak about experiencing regret in these terms: "Regret is a painful emotion, not only because one is confronted with a bad decision outcome, but even more because it points to one's role in causing that bad outcome" (Shani & Zeelenberg, 2007). The fact is that feeling that our performance is the cause of a bad result, that is, that we are responsible for it, is not pleasant. Even regret is the most intense of negative emotions (Saffrey et al., 2008). In addition, it is interesting to know whether regret is a universal experience in that it is understood and felt similarly in different countries and cultures. According to experiments (Breugelmans et al., 2014) conducted in the United States, Israel, Taiwan, and the Netherlands, the researchers found that regret is a distinct emotion that differs systematically from disappointment and guilt in all four countries. They adopted the same procedure in which emotions are conceptualized as multicomponent phenomena (Roseman et al., 1994) that may differ on several dimensions.

This paper concluded that regret could be assessed with five items measuring the specific components that make up this emotion, as expressed in this statement: "I felt *regret*. I felt *angry* with myself. I thought that I was *responsible* for the situation. I thought that I had made a *mistake*. I wanted to *correct* my mistake." According to these studies (Breugelmans et al., 2014; Roseman et al., 1994), regret over cultures seems to have a structural equivalence. And for the authors, this structural equivalence makes it more likely that these theories are globally applicable.

Thus, Zeelenberg dares to conclude that regret is a universal, painful, and frequent emotion, and these three properties could be decisive in understanding why we feel regret and what regret does with us. In addition to that, the same author adds:

> "The experience of regret has many cognitive and behavioural correlates and can motivate people to undo their decisions. Regret can also help people learn from their mistakes and prevent such mistakes in the future. It is precise because regret is so painful that it is so functional" (Zeelenberg, 2018).

.

These experiments are interesting because they study the variability or expressivity of regret in particular contexts and compare it with other emotions, such as guilt and disappointment, associated with specific feelings, thoughts, actions, tendencies, and motivations. This dissertation's relevant part or issue is whether the focus on these capacities of anticipated regrets can be anticipated in situations where one's decision-making can influence others. Moreover, as explained below, this ability to anticipate regrets works and requires learning and training.

This achievement can, in turn, be improved with the information provided by AI systems in the form of counterfactual explanations because, as we have already seen in Byrne's (2019) work, people tend to create counterfactuals by overriding certain reasonings, especially those that have a negative charge, such as what has to do with experiencing the unpleasure feeling of regret. Understanding why we have an aversion to this emotion and how to deal with it could help us to implement systems that support us with this important goal.

8.1.3 Anticipated regrets: linking with the future in the present

One of the paradoxes about experiencing emotions, even though we are talking about regret, is that we should not wait in a passive form for them to affect us. Chapter 7 focuses on imagination as a process that permits us to manipulate organic objects, giving them a different meaning or function, or we can try to create new mental objects. Therefore, we can also manipulate objects from thoughts, emotions, and behaviors to goals, to cite only some examples. That is possible because the essence of imagination is to have no limits when putting it into action.

Without further elaborating on this aspect, we return to regret. Anticipating this emotion can contribute to improving our future choices. In the same line, Zeelenberg (2018) expresses, "Before the decision, we may contemplate what to choose and think about what can go wrong, (…) and we can envision how we would respond emotionally to these wrong decisions, and hence try to prevent these bad outcomes from happening" (Zeelenberg, 2018)

To emphasize regret's role in decision-making since the 1950s, it is worth recalling the theory of the statistician Leonard Savage, who proposed as early as 1951 what is known as the "minimax regret" as a decision strategy. This strategy is basically about minimizing possible regrets as much as possible.

Put in technical terms, this means that the possible regrets for each option are computed, and the output chosen is equivalent to the one for which the maximum regret is the lowest (Stoye, 2009).

However, this theory is subject to criticism (Zeelenberg, 2015) as this strategy is valid when there is no knowledge about the probabilities of the possible outcomes, as it ignores the likelihood with which an event occurs and incorporates only how the outcomes compare.

Ignoring probabilities that other options are the right ones is resource-saving, or it is better to say clearly that the probabilities are usually known even if they cannot be determined with complete precision. And to say more, even when those possible scenarios are not fully known, one can often rank-order the world's states in terms of their likelihood (Zeelenberg, 2018).

Such rank ordering should help determine the importance of the associated outcomes in the final decision. Because of these issues, this strategy is not always adequate to solve any decision-making, and a possible solution maybe to lean towards approaches dealing with decision-making theories and incomplete knowledge (Ryan, 1983).

Another regret theory arose in the 1980s when a group of economists (Loomes & Sugden, 1982) chose to go further by integrating minimax regret theory with standard utility theory and formalizing it in regret theory. In summary, in the expected utility theory (EU), the representative of the choice option is its expected utility. That means the weighted average of utility (=value) of all possible outcomes, where the weights are the likelihood of that outcome occurring. Thus, this integrated theory assumes that the EU of an option depends not only on the likelihood and value of the outcomes but also on the regret or rejoicing one may feel over the choice.

However, it was not only economists who saw the potential of introducing anticipated regret into their theories to improve decision-making. Psychologists saw it, too.

Almost in the 1980s, specifically in 1979, Kahneman and Tversky discussed possibly including the regret role in their prospect theory (D. Kahneman & Tversky, 1979). However, they abandoned this approach "because it did not elegantly accommodate the pattern of results they labeled reflection" (D. Kahneman & Tversky, 2000).

Nevertheless, in 2011, Kahneman referred to regret and disappointment as the "two obvious blind spots in prospect theory," and he also added: "The emotions regret and disappointment are real, and decision-makers surely anticipate these emotions when making their choices" (D. Kahneman, 2011).

Years later, specifically in 2021 and with the new book by Kahneman and their colleagues' titled *Noise: a flaw in human judgment*, they have preferred to continue ignoring the role that emotions such as regret play in decision-making and, as a consequence, have established a working methodology to reduce the noise of system 2 (D. Kahneman, 2011).

System 2 was supposed to make reflective, slow, and deliberate decisions, but the authors have detected a significant variability in judgments of System 2, human judgments that must be identical. The relevant point is that the source of this variation corresponds to the mood or the sentic state of the decision maker or how explanations fit with the decision maker's understanding of the world.

This aspect was discussed in detail earlier in this research. Nevertheless, since it is one of the most critical aspects of this research work hypothesis, it is worth recalling. Nonetheless, the work of Kahneman and colleagues (2021), particularly their strategy of reducing noise through what they call "hygienic decision strategies," inspires the framework of this research in a certain way. However, while they see emotions as a source of noise, in this dissertation, emotions such as regret become a source of knowledge.

Hence, finding a logic of emotions might allow us to understand better their role in our decisions and how we relate to ourselves and others. In this way, we can detect emotions that are a noise source quickly and deal with them more reliably and rigorously than simply talking about people making decisions based on their moods.

This approach has not been scientifically studied and requires interdisciplinary experts to study this phenomenon in depth. This digression has been necessary to show that although Kahneman, Sibony, and Sunstein (2021) have inspired this research, their focus on emotions only as a noise source seems vague.

That said, it is worth focusing on the strengths of anticipating regret, which, as the literature on the subject claims, are not few. Hence, a way to understand the role that anticipatory regret plays in behavior or decision-making is described by Janis and Mann (1997) as follows:

> "Before undertaking any enterprise "of great pith and moment," we usually delay action and think about what might happen that could cause regret… Anticipatory regret is a convenient generic term to refer to the main psychological effects of the various worries that beset a decision-maker before any losses actually materialize… Such worries, which include anticipatory guilt and shame, provoke hesitation and doubt, making salient the realization that even the most attractive of the available choices might turn out badly" (Zeelenberg, Van Dijk, Van Der Pligt, et al., 1998).

However, the differences between how economists have been interpreting anticipated regret since the beginning as something cold and as a rational process contrast with the view of psychologists about how helpful anticipated regret could be to improve one's future behaviors.

This view is probably inspired by the way decision-making is conceived in the Western tradition as a rational and calculating process in which emotions don't play any role in performing analytical tasks, as explained in Chapter 6.

Back to the topic at hand, and to close this particular point, a table is shown below with the proposition of regret regulation theory (Pieters & Zeelenberg, 2007; Shani & Zeelenberg, 2007; Zeelenberg, 2018) where anticipated regret also plays an important role.

1.	**Regret is an aversive, cognitive emotion that people are motivated to regulate to maximize outcomes in the short term and learn to maximize them in the long run.**
2.	**Regret is a comparison-based emotion of self-blame experienced when people realize or imagine that their present situation would have been better had they decided differently in the past.**
3.	**Regret is distinct from related other specific emotions such as anger, disappointment, envy, guilt, sadness, and shame and from general negative affect on the basis of its appraisals, experiential content, and behavioral consequences.**
4.	**Individual differences in the tendency to experience regret are reliably related to the tendency to maximize and compare one's outcomes.**
5.	**Regret can be experienced about past (retrospective regret) and future (anticipated or prospective regret) decisions.**
6.	**Anticipated regret is experienced when decisions are difficult and important and when the decision maker expects to learn the outcomes of both the chosen and the rejected options quickly.**
7.	**Regret can stem from decisions to act and from decisions not to act: the more justifiable the decision, the less regret.**
8.	**Regret can be experienced about decision processes (process regret) and decision outcomes (outcome regret).**
9.	**The intensity of regret is contingent on the ease of comparing actual with counterfactual decision processes and outcomes and the importance, salience, and reversibility of discrepancy.**
10.	**Regret aversion is distinct from risk aversion, and they jointly and independently influence behavioral decisions.**
11.	**Regret regulation strategies are goal, decision, alternative, or feeling focused and implemented based on their accessibility and their instrumentality to the current overarching goal.**

TABLE 7 | The Proposition of Regret Regulation Theory (Shani & Zeelenberg, 2007)

In addition to the table presented above, several statements must be remembered. On the one hand, since the 1980s, the world of research, especially in economics and psychology, has been interested in the role of anticipating regret or what has also been called prefactual thinking.

Further, it is well-established that when decision-makers are confronted with important and difficult choices, they mentally simulate what might happen and what might not.

On the other hand, and according to this scenario, Zeelenberg et al. argue that:

> "These imagination processes are not only useful in the sense that they help us to oversee the consequences of our decisions. They also facilitate the arousal of emotional reactions such as anticipated regret, which, because of their motivational characteristics, may help us make better choices. (…) Regret is the prototypical decision-related emotion, and anticipated regret draws heavily on our ability to mentally simulate the future" (Zeelenberg, Van Dijk, Van Der Pligt, et al., 1998).

This ability or capacity provided by the experience of regret in the form of simulating the future with the help of reinforcement learning strategies is the subject of this chapter's next section. But before closing it, the aversive component of regret is commented on because of its consequences on responsible decision-making.

8.1.4 The experience of anticipated regret: between aversion to feelings or thoughts

As we saw in Chapter 6, emotions do not have an explicit ontology because there are different ways of classifying them. On the one hand, the two most supported approaches are the cognitivism view, which understands emotions as propositional attitudes, beliefs, or even judgments (de Sousa, 1987; Goldie, 2000; Nussbaum, 2001; Solomon, 1980). This view also fits with the constructivist view of emotions (Feldman Barrett, 2017). On the contrary, in feeling theories of emotions, emphasis on emotions can be understood as an awareness of bodily changes (Prinz, 2004). This approach can be classified into the classical theory of emotions. Then, where and how do we classify anticipated regret?

The mode of understanding regret that permeates this research fits with Zeelenberg's (2018) position as the fact that “anticipated regret has such an effect on behavioral choices shows that regret aversion is a clear motivator.

That supports the idea that anticipated regret is an emotional experience” (Zeelenberg, 2018). But are the skills required for retrospective regret and prospective similar? Let's look at this issue.

One way to encompass what is known as anticipated emotions is what Frijda and Scherer call “virtual emotions.” They define them as: “representations of the emotion one would have under certain circumstances, including virtual readiness for particular types of action and posture” (Frijda, 2017; Frijda & Scherer, 2009). This position of understanding anticipated emotions does not entail understanding emotions; they contain essential emotional aspects and motivational qualities, which are essential elements for executing actions.

Back to regret theory, regret (and rejoicing) concerns the more cognitivist tradition of emotions. This view is because regret is an outcome of a decision one can consider. The emotional experiences of this emotion are not central to this kind of reasoning; other than that, “regret is an aversive state that decision-makers would like to avoid” (Byrne, 2019; Zeelenberg, 2018).

On the opposite side and coinciding with the approach of the feeling theories of emotions is the position of Janis and Mann cited above, who wrote about anticipated regret as a state that may be felt when contemplating a decision, and it goes together with worries, feelings of guilt and hypervigilance. It is an emotion that does not (or not only) change how we evaluate the outcomes, but even more so, an emotion that changes how we approach the decision and finally make it. In the same vein as the ideas of Janis and Mann, we have the conclusions of Reb, who argues in his work that anticipated regret leads to more careful and better decisions (Reb, 2008).

To find a definition that fits with the objectives of the theory presented in this research, I take the more pragmatic form presented by Koch as follows: “Anticipated regret denotes a prospective, aversive, and cognitive emotion (i.e., an emotion that requires thinking) that influences decision making” (Koch, 2014). In addition, the other important aspect of this anticipatory regret is reflected in Janis and Mann's (1997) words as follows: "While anticipatory regret is related to retrospective regret, it is also different from it."

This different approach to the same emotion is because anticipatory regret is more forward-looking and probably more proactive than its retrospective variant (Zeelenberg, 2018). This phenomenon means that the learning and motivation component is fundamental, as discussed below.

8.2 Combining model-based reinforcement learning with anticipated regrets

8.2.1 From utility theory to reinforcement learning

As mentioned above, anticipated regret is a capacity we can train and learn, and several ways exist to encourage this learning process. However, learning to anticipate regret is characterized by different types of constraints. One is the one presented above: "Regret is an aversive state that decision-makers would like to avoid" (Byrne, 2019; Zeelenberg, 2018).

In other words, we seek to avoid the feeling produced by emotion in counterfactual reasoning because of its negative emotional component. In addition, other experiments, as shown below, have found certain limitations to experiencing this emotion from the neuroanatomical view.

Because of the importance of regret in decision-making and behavior related to the principle of responsibility, it is crucial to identify and understand the mechanisms that underlie these limitations by which this occurs to take appropriate measures.

Hence, to continue advancing in our path of gaining knowledge to improve the strategies for responsible decision-making with the help of AI systems in the case of this research, it is necessary to inquire about new ways of learning that allow us to evaluate their impact more precisely than ever before, comparing them with other previous learning strategies to decide which one applies better contextually.

Thus, we move from the utility theory to reinforcement learning (RL). While utility theory has occupied so much space in economic and mathematical theories in the attempt to "choose an action with the maximum utility function among a large number of potential options that makes it possible to make choices consistently and rationally" (Lee, 2017) the theory of learning by reinforcement adds the component choices that can change through experience or uncertain events.

According to Barto (1994), in this way, the goal of reinforcement learning is to understand how utilities must be altered by experience so that rational choices based on the utility functions can still produce the most desirable outcomes through learning (Barto, 1994).

Economics with mathematic models and psychology were the first disciplines to develop tools for studying decision-making; now, we also have computer scientists who implement powerful AI models for decision-making, but their origins and terminology differ.

In the case of the economists who developed the utilitarian approach, the hypothetical quantity determining all choices is called utility. In the case of the theory of reinforcement learning, which arose from psychology to study animal behavior, what in the utilitarian theory is known as a utility in the psychological field is known as value or value function. Computer scientists are interested in reinforcement learning because they consider RL an appropriate way to study machine intelligence (Lee, 2017).

However, the most relevant to this research is highlighted by Daeyeol Lee in his book *Birth of intelligence: from RNA to Artificial Intelligence*, which discusses how these theories deal with the concept of learning.

Thus, utilities in economics can be any hypothetical quantities consistent with a set of choices, and there has been little interest in economics regarding the origin of utilities or how they might be learned.

However, psychologists have always been interested in how humans and animals change their behavior through learning, just as computational scientists are now interested in how to train machines to be able to solve increasingly complex problems.

Therefore, “In this context, value functions can be viewed as the subjective estimate of future rewards expected by the decision maker. These quantities can be updated or learned through experience” (Lee, 2017).

8.2.2 Beyond classical reinforcement learning theory

As mentioned above, there is not only one way to learn how to make better decisions. In recent decades, new ways of addressing this issue have emerged. However, our approach is a proposal in which we modify aspects of reinforcement learning theory to fit our training to anticipate regret. For this purpose, exploring latent learning is mandatory.

However, in a very summarized way, understanding the historical reinforcement path could help. Therefore, we must go back to the 1930s with the studies of the behaviorist psychologist F. Skinner. The emphasis of his experiments and the hypothesis he wanted to test was that we could train animals to perform complex tasks through simple reinforcement mechanisms, such as receiving a food reward for performing a desired control.

To prove it, Skinner conducted several experiments on pigeons and rats to study this idea of reinforcement. His results pointed out that we can use reinforcement learning techniques to shape animal behavior, and Skinner (1934) developed the concept of positive reinforcement (Andery et al., 2005).

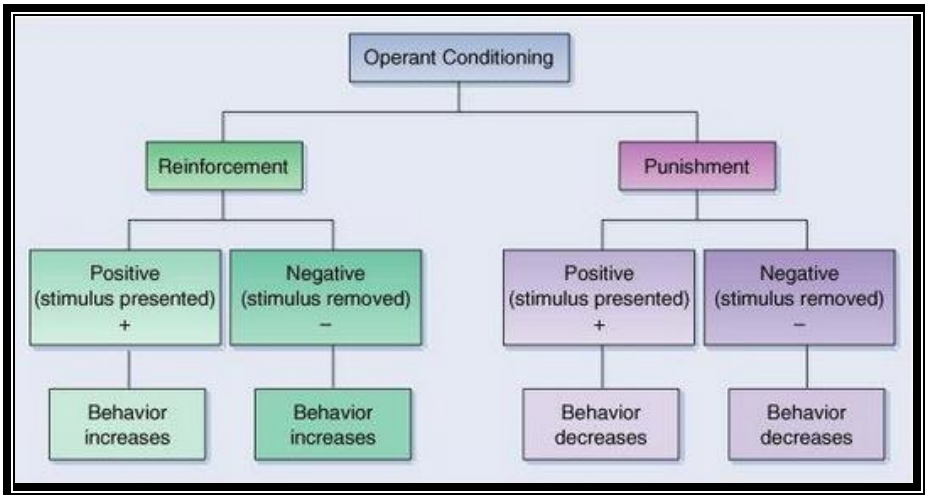


FIGURE 40 | Skinner Operant Conditioning Theory (Parchment, 2016)

The basic idea behind it is that -by extending the spectrum to human and artificial agents, an animal or agent can learn to optimize its behavior by learning from past experiences. One of the most commonly used examples of how this type of learning works is how children explore the world around them and learn from actions that help them reach a goal. Without a supervisor, the learner must independently discover the steps that maximize the reward.

FIGURE 41 | Skinner's Operant Conditioning and Learning / communicationtheory.org

This process of discovery is equivalent to what is known as a trial-and-error search. The quality of the action is not only the immediate reward but also the advantages of delaying this reward. Fortunately, this mechanistic approach to understanding learning of behaviorist psychologists like Hull (1935), who thought that learning is a strict connection of stimulus-response, has been partly overcome.

Thus, we arrive at the concept of latent learning developed by Edward Chance Tolman (1886-1959). Tolman does not break entirely with the behaviorist current of the time, which in that period was dominated by the ideas of J.B. Watson (1878-1958), who shared the same ideas of learning and human behavior as Thorndike and Hull (Jensen, 2006).

According to Jensen, Tolman (1930) dares to refute the idea that learning consists simply of mechanical conditioning processes, and his claims go in the direction that learning is related to complex mental problems. To test his theory, Tolman conducted experiments in which he emphasized the role of reinforcement in mice learning their ways in complex mazes. These experiments caused the birth of "The Theory of Latent Learning"; this theory expresses that learning occurs in situations without a sure reward (Jensen, 2006).

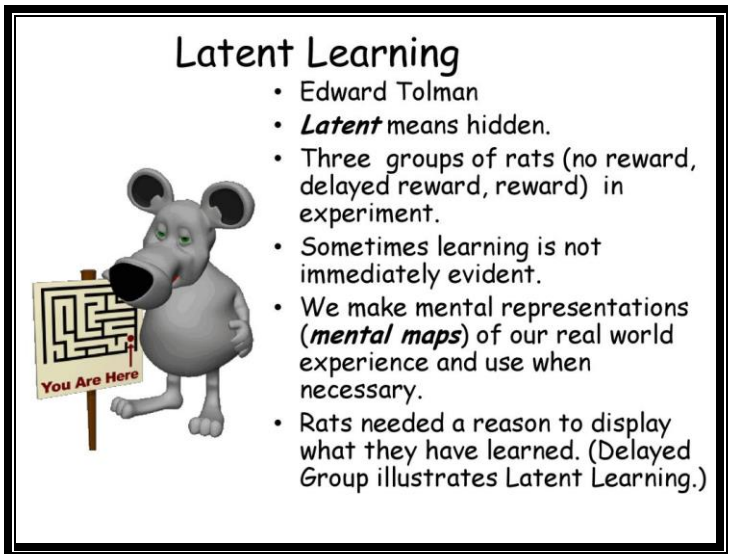


FIGURE 42 | Latent Learning / (Arlene, 2019)

We had to rescue Tolman's latent learning theory because, as Lee argues:

> "The reinforcement learning theory can account for the same phenomenon as latent learning using the concept of mental simulation. In mental simulation, animals can predict the hypothetical outcomes of various actions based on their knowledge of the environment and adjust the values of corresponding actions by comparing such hypothetical outcomes to previously expected outcomes" (Lee, 2017).

We can also find these abilities in how humans make daily decisions. An example to reflect this can be weather forecasting. Because weather forecasts are accurate but not 100% accurate, imagine you are getting ready to ride your bike to work because the forecast was for excellent sunshine. When it's time to get the bike, you look at the sky; clouds cover part of it, and the wind is present. You have to decide what to do, whether to risk 20 km on the bike and arrive wet at work or go by car and be safe. Which is the safer solution?

The answer is that, in the end, it rained, and the person decided to take precautions and go by car. However, this example does not mean everyone would decide the same thing in the same situation. It is precisely that reasoning process, in this case, called Lee mental simulation, that leads you to put into action another previously acquired knowledge of the current situation. Indeed, following the statement of this neuroscientist, “Without mental simulation, any knowledge acquired about a new environment would be useless to a decision maker because it cannot influence their actions” (Lee, 2017).

Hence, we have that in the theory of learning by reinforcement. There are two clear branches. On the one hand, there is model-based reinforcement learning, whose purpose is to adjust mental simulation and knowledge acquired in the past to update decision strategies.

On the other hand, we have the simple and classical instrumental conditioning typical of behaviorism theories (HEBB, 1956), where the learning process doesn’t require any mental simulation. This type refers to model-free reinforcement learning (Lee, 2017).

In this research, we deal with the advantages presented by the theory of model-based reinforcement learning because it focuses on the hypothesis that part of the decision-making that human beings have to make requires mental simulations that can be acquired through latent learning techniques, for example.

Of course, the challenges remain at this level, which are the two main constraints mentioned at the beginning of this research in determining what knowledge is adequate or serves to make a better decision and depend on the one hand by defining what world to represent and on the other hand, what information or values are to be prioritized to achieve the best result.

This dissertation supports Lee's assumption:

> "For successful model-based reinforcement learning, acquiring and revising the knowledge about the environment should never stop. It might not always be possible to know which environmental changes or events become relevant later". (Lee, 2017)

However, it would be short-sighted not to update the knowledge of your environment when it changes" (Lee, 2017). We add that it would be unethical or unfair not to adapt to change due to laziness, arrogance, fear, or other conscious or unconscious emotions or reasons.

8.2.3 In model-reinforcement learning, how much regret is enough?

"Model-based reinforcement learning and mental simulation never stop in humans" (Lee, 2017). If we consider it carefully, a particular statement is true even if we are unaware. As Lee dares to say again, we are always running mental simulations based on new information from our environment. Even when there are no factual reward prediction errors, we constantly adjust the values of different actions based on hypothetical reward prediction errors.

For this neuroscientist, all our thoughts might be mental simulations for model-based reinforcement learning, and regret plays a significant role in this learning process called "model-based reinforcement learning. In addition, as discussed in previous chapters, regret is a complex and even paradoxical emotion, and its real function and benefit might not be obvious" (Lee, 2017).

On the one hand, without it, one cannot detect, see, or learn from one's mistakes, but on the other hand, too much regret can cause thoughts about what we could have done to avoid the negative consequences of our actions. The technical concept for this type of information processing is called rumination.

One of the key points to keep in mind is that we can start from the idea that when we decide on something or perform a particular action, we usually take it without sufficient information to determine with certainty which action will produce the best result. Most of the time, it is clear that only sometime later we would know what would have been the best decision.

However, on many occasions, only if we have asked ourselves the right questions and gone to diverse sources of knowledge.

In addition, other elements may emerge in this search for information that may remain as "latent learning" (Jensen, 2006) for future encounters. But this ability to reflect on what we could change to improve the future is a product of regret because we might experience it.

So, "human experience regrets not because we do not understand that past actions cannot be undone. Instead, these emotions are common because humans are constantly engaged in model-based reinforcement learning to improve our behavior in the future" (Lee, 2017).

8.2.4 Regret and decision making: a neuroanatomical view

From the strictly neuroanatomical point of view, a study was published in 2004 showing that activity in the orbitofrontal cortex (OFC) is closely related to regret (Camille et al., 2004). To focus more closely, the OFC is a part of the prefrontal cortex.

Thus, for several decades, neuroscientists worked with the hypothesis that the prefrontal cortex played an essential role in many high-order cognitive processes related to thoughts and emotions.

This hypothesis is corroborated by studies conducted some years later, in 2007, by Coricelli and co, which arrived at the same results as the previous studies. In this way, neuropsychological and neuroimaging data have stressed the fundamental role of the orbitofrontal cortex in mediating the experience of regret.

"Functional magnetic resonance imaging data indicate that reactivation of activity within the orbitofrontal cortex and amygdala during the phase of choice when the brain is anticipating possible future consequences of decisions characterizes the anticipation of regret" (Coricelli et al., 2005, 2007).

Another aspect that we need to highlight, already discussed in Chapter 6, is the relationship between the ability to make decisions and brain lesions, i.e., whether specific abilities are compromised after a brain injury, like in the case of Phineas Cage (Damasio, 1994).

Since then, several studies have continued to study this phenomenon. However, because the topic of this research is regret, the study mentioned above (Camille et al., 2004) is more relevant because it focused, in detail, on measuring whether the mood of the OCF patients and normal subjects who participated in the experiment reported that their mood worsened with the magnitude of negative reward prediction errors in the outcomes of their choices.

The results were that normal subjects reported more negative moods when there was regret, namely, when the hypothetical outcome from the unchosen target was better than the actual outcome from their choice.

Negative hypothetical reward prediction errors were aversive to normal subjects. On the other hand, the hypothetical outcomes did not affect the mood of OFC patients. They did not report negative emotions, even when they learned that the hypothetical outcome from the unchosen target was better than the outcome from their choice. In addition, Coricelly and co (2005) observed enhanced responses in the right dorsolateral prefrontal cortex, right lateral OFC, and inferior parietal lobule during a choice phase after the experience of regret and subsequent choice processes induced reinforcement, or avoidance, of the experienced behavior (Coricelli et al., 2005).

Similarly, the results from Simon-Thomas and Knight show that:

> "Negative emotions can recruit cognitive-based right hemisphere responses. Negative affective consequences (regret) induce specific mechanisms of cognitive control on subsequent choices" (Simon-Thomas & Knight, 2005).

Nevertheless, according to Lee (2017), although observing behavioral and psychological changes following damage to a brain area provides valuable information about how different brain regions might contribute to specific aspects of cognition, a solid understanding of brain damage and lesions is still limited.

In the same vein, lesions or studies of brain damage can tell us that the OFC might be involved in emotional and cognitive processes related to regret, but that is the crucial element that Lee expresses with the following words: “These studies do not provide much information about the precise nature of OFC functions related to regret” (Lee, 2017). Since individual neurons exchange information using action potentials, it would be beneficial if we could monitor the action potentials of individual neurons in the human OFC, but such an experiment is presently impossible to do in healthy subjects. In addition, these neuroscience findings lead us to other theories about the role of emotions in decision-making, which have been mentioned in Chapter 6.

Furthermore, if these theories could be proven, they would validate Damasio's concept (1994) that a "sick" individual, which we have explained in detail in the previous chapter, could be fitted with a person with no activity in the OFC in terms of not experiencing regret. On the other hand, these neuroscience approaches to regret fit more with the traditional theory of emotions, in the sense that Feldman Barret understands them (Feldman Barret, 2017; Feldman Barrett, 2017) because she argues that looking only for some "fingerprints" and not studying the variability of the regret experience itself is a fundamental and categorical error.

Is regret only expressed as a brain function and activated in a particular brain area? Or does it instead manifest itself as bodily changes? Many more studies are needed to answer these questions, which Lee (2017) makes clear could not be performed on healthy subjects because the techniques to measure them are invasive.

However, regrets seem to be involved in continuous learning processes as we review our actions repeatedly to improve our future choices. But here, another even more profound and more complex set of questions arises about the impact of our decision-making on the lives of others. What happens when we have to make decisions for others?

How can we prove this person is suitable if we cannot detect it yet? What happens when someone decides for us such vital aspects as our future health or future job?

What if a developer automates that decision, and nobody knows why? Moreover, what happens when a group has to decide the future of millions but cannot experience the anticipated regret described in the previous section?

We are talking about using all our resources today to discover new patterns that allow us to understand reality from a broader perspective, making our decisions reach a new, more ethical, humble, and socially responsible level.

8.2.5 Can children be trained to anticipate their regrets?

According to the study conducted by McCormack and colleagues, the ability to experience regret emerges relatively late in children, around 8 years old. In addition, the experiment concluded that children who experienced regret tended to make better choices in the future, and the other important element was that this emotion helps children learn to delay gratification and behave more prosocially. The study begins with the following example: Suppose you move to an unfamiliar country. It is a risk, but you decide to go because you predict that you will regret staying home forever: Anticipating future regret guides your choice. However, this emotion, whose beneficial effects are not visible at first glance, provides meaningful points for correcting past mistakes and those likely to occur (McCormack et al., 2020).

However, related to the previous section and how humans construct counterfactual scenarios rejecting the regret experience (Byrne, 2019) or according to Zeelenberg (2008), we feel an aversion to regret. So much so that research on how adults experience regret points to the fact that "they adopt a strategy of anticipating the regret they would feel after choosing options that appear attractive but are ultimately disadvantageous" (…); therefore, "anticipation of regret leads them to avoid such choices" (McCormack et al., 2020).

The decisive aspect also of the study is that although it was found that, on the one hand, the experience of regret directly affects the decision-making of children, on the other hand, regret anticipation and its relationship with delaying gratification may be because they decided to wait because they anticipated that they would later regret taking the smaller, more immediate reward. While regret helps children make decisions, how exactly this happens is still unknown.

In the same way, another observation in the paper written by MacCormack et al. (2020) is that it remains to be proven whether anticipation of regret is likely to require additional cognitive skills because it involves not only thinking counterfactually but also considering counterfactuals that one might entertain in the future and their emotional impact.

In the next section, we address this question based on the studies conducted by Kahneman and colleagues (2021) on biases, but as biases have already been discussed in Chapter 5, we focus more specifically on the term noise and how to prevent it. It is not a matter of presenting their work but of reviewing it and modifying some of its elements or at least justifying why their theory could be improved.

8.3 Understanding to prevent potential unintended errors

In recent years, the need to reduce and understand both bias and noise in decision-making has led many researchers and institutions, both private and public, to invest both time and money in this task. Specifically, this section presents different ways that Kahneman and his colleagues (2021) point out in their book *Noise* that pursue this same goal. Remember that Kahneman's work has been present throughout this research, and this section takes as a starting point the ideas of Kahneman, Sibony, and Sunstein (2021) in trying to reduce biases and noise in human judgments. Then, this section considers some of their suggestions as a good starting point to improve decision-making and make it more ethical. Nevertheless, this research cannot entirely agree with their recommendations. We explain what these disagreements are.

One controversial point is the one proposed by Sunstein (2022), who, in his paper entitled: *Governing by algorithm? no noise and (potentially) less bias* claims that the noise and biases in human judgments can be almost wholly reduced if we change the role of algorithms from mere advisors to decision-makers (Sunstein, 2022). After what has been explained in the different chapters of this dissertation, Sunstein's proposal is perilous, and the impacts on society without further study could be catastrophic. Of course, algorithms could minimize noise and biases. The fact is that their mood remains unchanged, they do not vary their opinions in similar contexts that require similar judgments, and they can follow the rules perfectly. However, it is impossible today to implement algorithms that capture the complex context of each situation they must decide.

Consequently, their judgments would be accurate according to the rules and operations they have to perform. But would it be possible for algorithms to detect exceptions to understand contexts beyond the data with which the system has been trained? Moreover, could a system help to represent data that has not previously been fed into it?

For example, one could be a doctor who used to solve complicated cases, not only what is most probable statistically speaking, but when it comes to helping a patient, many types of reasoning are involved, including reasoning about how particular emotions trigger one decision or another. The same happens in situations such as a war when, for example, a soldier is in a new situation of maximum risk and must make a decision in a few seconds to save the life of a civilian.

An action that the soldier may not even be able to explain because of the complexity of the reasoning or feelings involved in making such a decision. Can we give explanations that rationalize what we feel when this pushes us to make one decision and not another? In short, are algorithms involved with the motivation and intention to do their job to the best of their ability? Algorithms have not yet reached a state of meta-reflection. As explained in Chapter 6 on emotions, it is still up to humans to understand the variability of their emotions and their impact on how humans make decisions.

Therefore, denying this role to emotions as Kahneman and his colleagues do (2021), giving it only the negative aspect of provoking noise, even bias may cause more noise and new bias in other directions that we cannot even imagine yet, but that would go almost in the direction of what Coeckelberg (2010) calls as psychopathy in heartless AI systems.

Thus, this dissertation takes bias and noise as human shortcomings in the sense that we are currently not able to fully understand the impact that these terms have on the decisions or choices that we make, especially in interpersonal decisions, which are those made by an individual or group of individuals for others. To give an example, it has been more than demonstrated that the experience of anticipated regret, whether in the form of cognition or emotion, involves a kind of meta-reflection on what kind of decisions would be better in the long term, even if this means delaying the reward (Breugelmans et al., 2014; Pieters & Zeelenberg, 2007; Shani & Zeelenberg, 2007; Zeelenberg, 2018). A relevant issue to resolve in the future is that studies or experiments that aim to understand the formation of "mental simulations" (Lee, 2017) as occurs in the experience of anticipating regret (Shani & Zeelenberg, 2007; Zeelenberg, 2018) or how humans formulate their counterfactual scenarios or imagine future scenarios (Byrne, 2019, 2020; Byrne & Timmons, 2018; Saffrey et al., 2008) focus on rather intrapersonal and not interpersonal issues.

What is interesting about improving responsible decision-making is understanding how so-called "social decision-making" is made (Lee, 2017) or even more focused on the objectives of this thesis, how our decision-making affects the relationships we establish with the environment in which we move, understanding the concept of environment in all its breadth.

8.3.1 Why do we design an AI system as a “Decision Observer” for detecting bias?

After all the theories analyzed throughout this dissertation, a good starting point for improving responsible decision-making in a human-AI interaction is the one proposed by Kahneman and his colleagues (2021) called “The Decision Observer.” However, this research gives this "decision observer" a different role. While the authors of *Noise* have opted for this position for a human agent, the framework of this dissertation argues that an AI system would be more beneficial to improve our decision-making because rather than continuing to look for ways to reduce noise and biases with Ex Post and Ex Ante strategies, the crucial point would be to track the processing of decision making in “real-time.” It is important to note that this idea of tracking decision-making in “real-time” also belongs to the authors of *Noise*, but the novelty is the approach of opting for an AI system for this role.

Then, how would an algorithm be implemented to monitor our decision-making over time? Couldn't it be considered necessary, even legally, to implement a more objective “decision observer" that helps us detect when biases arise? However, where is our privacy? How do we create a balance? Is our fight for privacy sometimes a way to hide the real reasons for our decision-making? Then, could these systems teach us to understand our reasoning regardless of whether it is caused by reason or emotional matters? Which of these questions should we start with? Could systems even contribute to shaping new ontologies and new meanings? Would they change our understanding of reality, normality, and our way of perceiving human nature more transparently?

It is important to remark that this dissertation has supervisors more on the technical side, which gives this research an insider’s understanding of the advantages and disadvantages of both branches of AI (without these approaches and this interdisciplinary support, this research work would not have been possible).

However, due to my limited training in these techniques and my background, this doctoral thesis has aimed to explore the theoretical aspects of decision-making science from an interdisciplinary viewpoint in which emotions critically connect actions with ethical values. Specifically, a consistent multi-ethical framework is proposed to improve responsibility in human-AI interaction with the study of regret and its connection with the principle of responsibility within the development model of Confucian ethics.

In this way, the controversy at this point is not so much whether this technology can be implemented in the next few years but whether we can understand the phenomenon of biases and noise by reaching a degree of maturity in the emotional and rational mechanisms that drive aspects as moral human judgments, natural explanations, and human behavior.

Having said all this, we begin with the theoretical proposal for detecting biases. Hence, how do you get an AI system to observe them in real-time and classify them afterward? We need a list of rules (symbolic AI) that contains, in a well-defined way, the biases we want to detect. The authors of *Noise* have good advice in this regard. They suggest using a checklist (D. Kahneman et al., 2021). The reason is that this technique has a long tradition of improving decision-making in high-risk contexts and is particularly useful in preventing the repetition of past errors.

Moreover, the authors of *Noise* have another idea that is interesting to implement in an AI system: "This checklist does not have to be an exhaustive list of all the cases of biases that can affect a decision but focus on the most frequent and those that can have the most negative impact in the concrete context" (D. Kahneman et al., 2021). In this process of observation, collection, and classification of data by the AI system with machine learning techniques (sub-symbolic AI), the system may encounter recurring data that could not be classified in the system of rules embedded in the checklist encountered before. And it is this phenomenon that provides new ways to understand biases in a novel and groundbreaking way, as explained below.

8.3.2 An AI System explains (new) observations of bias systematically

The novelty of these techniques is that this AI system could contribute to generating new epistemological categories that describe new properties from previously unknown contexts. In this way, new rules emerge, and we need to categorize this further information observed and collected by the system.

This approach would help us more closely understand the variability of judgments and which components of "reasons-emotions" decision-making are most prevalent in addition to other findings. This way of proceeding is similar to one or two purposes of explainability that Adadi and Berrada (2018) point out. In this particular case, one of these goals is that Explainable AI in the role of a “Decision Observer” “could be the foundation for ongoing iteration and improvement between humans and machines” (Adadi & Berrada, 2018).

The other aim, cited above by the authors, is to discover effective ways of learning new facts, gathering information, and thus, gaining knowledge. In addition, “So it will come as no surprise if, in the future, XAI models teach us about new hidden laws in different fields” (Adadi & Berrada, 2018). To these statements, we propose modifying or expanding the role of these systems from mere "explainers" to active observers who can show us causal elements or correlations that were never seen before with their observation methods and excellent memory capacity. Furthermore, an explainable component in an AI system would no longer be a passive element but a way of understanding new aspects of reality previously not perceptible by human capacities.

The fact that these systems may be able to make autonomous decisions in high-risk decisions is an issue beyond the scope of this research work, although my answer to this question is clear: not at the moment. Nevertheless, this technical panorama makes a call, almost a cry for a robust social debate, to the point of starting a new paradigm (Van Dyck, 2018). To continue building ethical frameworks or legislations without knowing human nature is to continue building sandcastles as in the past, but nothing solid and unfair. This hypothesis is the philosophical core of this research: the human being is his nature, and it is in our hands to understand it as it is because it is the only way to learn from our mistakes, and it is what allows us to create new solutions to problems that are repeated over and over again.

Once again, some compelling reasons for creating a position known as "Decision Observer" are those provided by the authors of *Noise* in the following way:

> "Of course, people are rarely aware of their own biases when they are misleading them. This lack of awareness is itself a known bias, the bias blind spot. People often recognize biases more easily in others than they do in themselves. We suggest that observers can be trained to spot, in real-time, the diagnostic signs that one or several familiar biases are affecting someone else’s decisions or recommendations” (D. Kahneman et al., 2021).

In addition, our understanding of an AI system's role as a “Decision Observer” is similar to Rosalind Picard's (1995) idea to design and create an AI system capable of being the best piano teacher. According to Picard, an AI system could be a better piano teacher than a human if it detects in real-time (I do not know if Picard and his colleagues have achieved this) when a student loses interest and how to restore one’s motivation.

Detecting these changes in a body or mind with an "objective" observer built just for this purpose can be a solution to measure the quality of decisions regarding their ethical impacts, especially to continue learning what keeps humans motivated beyond their immediate needs.

In addition, the aim of this dissertation is not to replace human piano teachers with AI systems; these systems or AI agents could be beneficial as assistants to human agents to optimize learning processes and to understand in a personalized and contextualized way the different mechanisms of motivation that exist in the learning processes and of course, extrapolated to the issue of behavior.

On the other hand, it is essential to highlight that the authors of *Noise* have taken the position of decision observer only for bias issues. Well, in this dissertation, we see that the AI system as a “Decision Observer” plays a vital role in the understanding of the processes that make up both bias and noise because while Kahneman et al. (2021) interpret only emotions, moods and feelings as sources of noise, we see these components also as sources of bias.

Therefore, what has been described in this section can be extrapolated to study both sources of error in human judgments. These sources of error, we repeat, are considered for the moment as an impediment to improving our decision-making, but the moment we understand these mechanisms, they might cease to be an error and become an advance in the understanding of our behavior and human nature.

8.3.3 Preventing noise: the role of anticipated regret as recommendation norm

This section closes this dissertation with the proposal of an interdisciplinary multi-ethical framework divided into three phases, further subdivided depending on the type of explanations required in each phase.

The framework's three main groups correspond to the first phase of detection and observation, which culminates in the generation of explanations of the detected and observed data. The second phase is concerned with intervention and understanding whether the goals set for the tasks that human and artificial agents have to perform are what is expected.

If not, the third phase has to intervene with specific strategies of reinforcement learning to achieve the objectives and expectations set to complete this third phase. It should be emphasized that the object manipulated throughout this framework is the emotion of regret in both prospective and retrospective variants.

However, before starting with our proposal, why has it been chosen to establish a strategy to reduce the noise phenomenon to improve responsible decision-making in a human-AI interaction? According to Kahneman et al., the noise phenomenon is much more complicated to detect than biases, and its study, although it has begun to be systematized, is still a very new field with knowledge gaps.

For this reason, and given the importance of noise's impact on social or interpersonal decisions, it is crucial to address this issue rigorously and with a well-structured methodology.

Because we want to prevent noise, starting with one way of defining a strategy is convenient. We have again opted for the strategy that Kahneman and his colleagues offered in their book *Noise*: "The goal is to prevent an unspecified range of potential errors before they occur" (D. Kahneman et al., 2021). However, this definition is vague; we aim to concretize or further deepen the meaning of these terms with a strategic plan.

In addition, we must implement new interdisciplinary methodologies that provide a more precise and transparent vision of this noise phenomenon. Given the scenario described above, we assume that noise is present both in human judgments and in the algorithms that replicate the noise inherent from their programmers in a non-visible and non-understandable way because of our inherent limitations.

This potential source of error does not mean there is an explicit failure in their reasoning. Still, as we have tried to demonstrate throughout this dissertation, there is more than one way of interpreting reality, and depending on which models and data are chosen to work, the outputs might be completely different. For this purpose, some theories or frameworks cited in this dissertation's chapters serve as inspiration sources and make this interdisciplinary multi-ethical framework possible. So, it is time to rescue the research question we are trying to answer:

> How could we design an interdisciplinary theoretical framework for responsible AI, where developing different levels of explanations in human-AI interaction validates whether the somatic marker hypothesis could serve as a recommendation norm for preventing unspecified errors before they occur?

Before starting with the description of the framework and the different phases, it is vital to highlight the importance of continuous learning throughout the process. It means that the central angle is the continuous learning of how a set of specifications in the form of a priori rules lends itself to identifying variations or new patterns that might introduce improvements to the established rules, which need to be updated if necessary.

For example, if it is discovered when analyzing a large amount of medical data that certain types of new drugs of which there was previously no awareness have significantly improved the health of many patients, these behaviors might function in the future as rules to be followed in the subsequent treatment of patients with that pathology.

In this way, the rules and the data update serve as new feedback to continue adjusting these rules and behaviors. It is imperative to consider the accessibility to quality data and its traceability over a long period to detect new patterns from which new knowledge that can be considered valuable can be inferred.

Before making any changes, it is crucial to bring together the human experts in the field to assess the system's results. In this way, the AI system's explanatory component becomes a support tool that explains why it has arrived at its result based on the theory of mind implemented in the AI system.

It is also cardinal to emphasize that the human agents must perceive AI systems as teammates and not as rivals because they contribute to the expansion of knowledge in a particular area of work, and they are programmed for well-defined tasks with clear rules, which, with the learning of actual and factual data, might vary according to the needs of the specific context. This phenomenon resulting from this form of learning can be labeled as emergent (Prigogine & Stengers, 1984) because the information or data that can be fed into the system tends to become more and more precise and specific, and this design of the technology leads to an adjustment of some of the rules when the situation or new specifications to solve a new problem require it.

To move from a strictly theoretical sphere into the experimental world would be required at the technical level of the implementation of a hybrid approach, as explained in Chapter 6, where the cognitive level (symbolic AI) is combined with the sensory level (sub-symbolic AI) in terms of the subfield known as goal-driven Explainable AI (XGDAI). However, before explaining the role of regret in the three phases of the framework, it is worth starting with the multi-ethical framework in which the whole proposal is embedded.

Anticipated regret as a deontic value in the moral development of Confucian ethics

From the point of view of normative ethics, this framework combines two theories: Deontology and Confucianism. These two ethical approaches have been chosen because each benefits from the other, as described below. According to Tolmeijer et al., "Deontology could be implemented by inputting the action (in terms of mental states and consequences), using rules and duties as the decision criteria, and then mechanizing actions via the extent to which they fit with the rule" (Tolmeijer et al., 2020). In this way, the deontological character of regret as value is expressed in its capacity to assume decision-making responsibility. In other words, the capacity to be aware of the salient role that this concrete emotion plays in human behavior to prevent various errors before they occur.

However, are humans communicating or expressing their regrets the same in different cultures, societies, and member communities? Moreover, are the future consequences of actions assessed in the same way? This limitation is one of the challenges facing deontological ethics because this ethical approach focuses on the intrinsic nature of an action and fails to consider the most likely consequences (Woodgate & Ajmeri, 2022).

This assumption does not mean that deontology takes a back seat as an ethical theory. However, to understand the maxim that "variation is the norm," we should look for a possible combination. We require another ethical theory that focuses on the consequences of a diversity of actions with a dynamic character and where the normative value of regret can be tested in a contextual variety and over time because of human behavior's dynamic and changing nature. Therefore, the role-based Confucian approach could bring this new perspective because it fits very well with the Western approach to deontological ethics to achieving our goals.

This idea of combining Western and Eastern values has its novel side. In addition, the reasons for combining ethical theories are based on the following assumption as stated by Zhu et al. because they “enrich the moral imagination of AI experts to enhance their capabilities to define and engage ethical issues in designing AI systems from culturally diverse perspectives” (Zhu et al., 2019).

However, the most disruptive aspect that Zhu et al. point out and which represents a turning point for this dissertation is reflected in the following quote:

> “Confucian role ethics focuses on the premise that humans were born into a web of social relationships, and their interactions have normative implications, and they prescribe specific moral responsibilities for us in the communities we belong to" (Zhu et al., 2019).

Hence, for Zhu et al., the Confucian ultimate goal of becoming a good person depends on how well we live and practice the moral responsibilities prescribed by these social roles. In this dissertation, these social roles are nurtured by social norms, which justify the necessity, even the urgency, to generate a new social recommendation norm that enforces the maxim of experiencing a sense of responsibility expressed in the ability to experience regret. Another relevant aspect is the very concept of dignified behavior or simply being a good person. Zhu argues that:

> "Western ethical approaches focus on moral reasoning and justification while Confucian ethics emphasizes ethical practice and practical wisdom” (Zhu, 2018).

This research work is based on the idea that both approaches combine harmoniously in shaping and understanding human behavior and changing priorities in moral values according to new knowledge about specific properties of reality to improve our responsible and social decision-making over time.

For example, regret and its connection with the principle of responsibility could play a central role in the model of moral development that is central to Confucian ethics, consisting of these three interrelated components: observation, reflection, and practice" (Zhu, 2018).

It is important to note that these three phases of the model of moral development of Confucianism are very similar or fit very well with the way our work proposal is structured, which is primarily inspired by the three levels described in the framework proposed by Pearl and Mackenzie (2018) entitled: "The Ladder of Causality," presented at the beginning of this chapter. Although, as already stated in the Pearl and Mackenzie (2018) framework, the differences between levels 2 and 3 are unclear, there might be a substantial difference in our case. While this methodology can be used for multiple purposes because it is very versatile, the one presented here is for preventing noise and has the experience or the emotion of regret as a manipulation object in the three levels that comprise the framework.

Thus, each level of this Ladder of Causation has tasks shared between humans and AI systems, or, in other words, human-AI interaction is required at each stage of the Ladder and only when a level has been completed, i.e., assessed, it is possible to move on to the next level.

The first level: generating explanations of detected patterns

So, in this first level of the Ladder, what are our object of study and our justification for carrying it out? It is important to stress again that this dissertation starts from the somatic marker hypothesis (Damasio, 1994). However, our approach here is based on an updated version of the same hypothesis proposed by Overskeid (2021).

He proposes a more radical reading of it. In other words, emotions cause decision-making from the beginning as a motivating element to the end when decision-making is consummated (Overskeid, 2021). Under this assumption, this framework aims to observe and detect the variability of the emotion of regret.

This purpose refers to Feldman Barrett's (2017; 2020) conclusions from her studies on the variability of emotions or how one single emotion is expressed in different individuals. This view contrasts with the classical theory of emotions, which looks for universal patterns with simple correlations. (Alfaisal & Aljanada, 2018; Ekman, 2005, 2017; Ekman et al., 1972; Izard & Haynes, 1988; R. W. Picard, 1995).

Given that both theories are antagonistic, and it is not the purpose of this thesis to lean wholly towards one of them, the proposed approach is to broaden and deepen into each to understand their concrete impact on human judgments, specifically on interpersonal decision making.

On the one hand, the assumption is that an emotion or emotional category motivates different behaviors and is expressed differently (Feldman Barrett, 2017). This dissertation aims to delve into regret's impact on human behavior. Hence, this first level focuses on reaching an "objectual understanding" (Páez, 2019) about the emotion of regret as a whole in both its anticipatory and retrospective variants to detect when, where, why, how, and what drives also "regret aversion" (Byrne, 2019; Lee, 2017) in counterfactual reasoning in context-dependent scenarios (Oltramari et al., 2020).

In terms of how cognitive explanations in this first level could be generated (Neerincx et al., 2018) in the hybrid system for our goal-driven XAI approach in this first level, it is suggested that the framework Case-Based Reasoning (CBR) proposed by Urdiales et al. They applied a learning and adaptation strategy to help explanation generation by storing, recalling, and adapting knowledge information or experiences (or cases) stored in a case library or memory (Urdiales et al., 2006).

One of the key points discussed above is that one of the characteristics of the way explanations are understood in this framework is the creation of dynamic knowledge.

In this way, in the CBR framework, agent learning to expand knowledge is executed by evaluating and integrating new experiences into the case library and re-indexing and reusing previous experiences (Kolodner, 1993).

According to Urdiales et al. CBR has two primary learning methods: observational learning (supervised) and learning through personal experience. Observational learning is performed by filling the case library with observations from expert demonstrations or actual data.

Learning from one's personal experience happens after a reasoning process that evaluates a potential solution to a challenge. If the solution succeeds, it is saved and applied for future reference (Sado et al., 2020; Urdiales et al., 2006).

However, it is also mandatory to save cases where the anticipation of future regrets seems problematic for the decision-maker and to understand why. In the end, understanding where the aversion to experience or express regret arises results in a specific type of behavior and not the expected one.

Regarding how to study this variability of experience in our specific case of regret at the subsymbolic or in terms of sensory explanations, the techniques in "real-time" in the field of Affective Computing (R. W. Picard, 1995) could be beneficial. They could solve significant challenges we face today. However, this technique also entails several risks that need to be listed and of which one should be aware. One constraint is that the techniques for measuring neural and biometric data are considered highly invasive (Lee, 2017).

Another challenge is data protection from the misuse (White & Katsuno, 2022). This challenge is one of the aspects we return to in the conclusions because although a technology may be implemented or intended to solve a specific problem, once it is on the market, it can be used for other purposes that the creators or designers of the technological devices did not have in mind.

Another essential component of this framework is the necessity to be open to dialogue with the actors that contribute to a project of these conditions in case changes need to be made because, for example, the processing of specific data would contribute to a violation of certain rights.

This kind of aspect should not mean that the project should be dismissed as invalid but rather justify why new legislative conditions must be created for the type of technology to be implemented and its possible unintended consequences.

In addition, as much as we want to anticipate all the possible things we might regret in the future, the uncertainty factor is still present. This form of communication and the involvement of different actors justifies the principles of transparency and trustworthiness.

The first level: generating explanations of detected patterns:

“Variation is the norm”

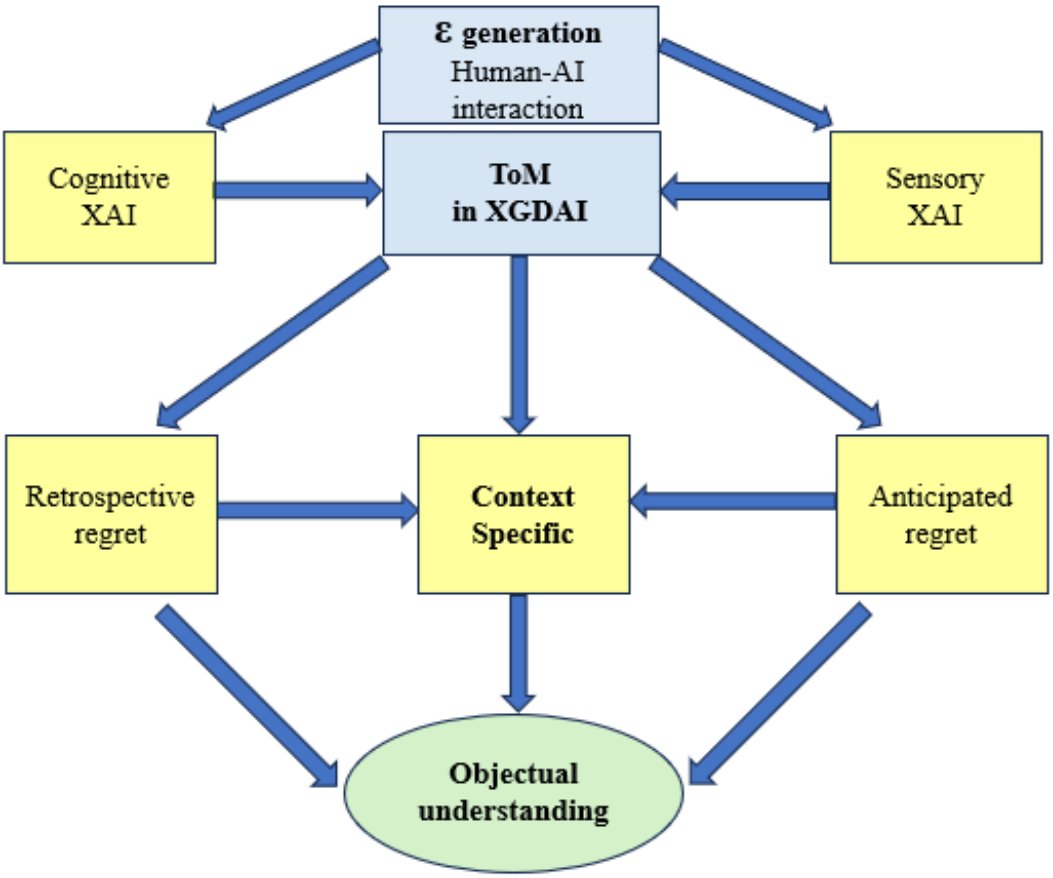


FIGURE 43 | Generating explanations of detected patterns (Own contribution based on the authors cited above)

The second level: explanations for communicating possible future regrets

The second phase or level of this Ladder corresponds to action or intervention. Intervention can only take place when the first level has been completed. This condition is necessary to design the content and the form of the explanations in this framework phase. The aim at this level is to detect whether regret aversion remains even after having intervened with explanations, which try to bring to the surface the possible regrets that decision-makers will be exposed to in the future if they do not plan actions to counteract them before they can occur.

Imagine then that we have completed the first phase with a study of the variability of regret using affective computing techniques in a specific field, for example, in the medical field, where sometimes the decision-making has to do with life and death issues where reflection on the possible regrets that may arise in the future should lead to a well-planned decision making. In theory, we all agree that a doctor can and should consider what effects a previously untested treatment might have or that some symptoms do not match the current patient's diagnosis due to abnormalities in other values.

These data should be used to re-evaluate the cases they are working with. However, in practice, issues related to biases, noise (D. Kahneman et al., 2021), and other factors such as the time available to the doctor to focus on each study mean that this customization is not always possible or is neither cost-effective nor feasible.

Thus, let us assume that we have analyzed this large amount of data through a meta-analysis (Feldman Barret, 2017) that shows how regret, in its retrospective and anticipatory variants, is expressed in doctors of different cultures when they have to make decisions (in this field, every decision could be considered high risk).

However, what we are pursuing at this second level is no longer to detect regularities, causal connections, or simple correlations. In this second phase, we seek to intervene for a purpose-oriented goal, which, in the doctor's case, is to find the most appropriate treatment to avoid major problems. The end pursued is to know if anticipated regret manipulated as a "mental simulation" (Lee, 2017) can play a more active role in responsible decision-making.

Likewise, this second level could fit very well with the definition that Kahneman and his colleagues have put forward in their book *Noise*:

> "The goal is to prevent an unspecified range of potential errors before they occur" (D. Kahneman et al., 2021).

As expressed in this definition, the prospective component of regret pushes us mentally to imagine these potential errors we could cause if we do not put a solution before they happen. However, here is the catch: first, we tend to imagine these possible futures, which are very much conditioned by the factors that Byrne (2019) explains in how humans use their reasoning to create counterfactual scenarios. In this point, we can easily find a knowledge gap centered on the fact that in the formation of these counterfactuals (whether they are projected into the past or the future), what is detected is that human beings have an aversion to anything that leads to experiencing pain or a negative feeling (Byrne, 2019).

This human way of reasoning, which may have to do with other hitherto unknown elements of survival, among other possible reasons, means that the benefits attributed to the ability to experience and express regret are not visible at first glance. (Lee, 2017). However, it is precisely for this reason that this dissertation has focused on studying the variant of regret in its prospective version. There is still much research to be done, and this is precisely what this dissertation aims to do: to propose enough reasons to promote its study and to reinforce the learning strategies of this critical emerging capacity due to its importance in improving ethical and responsible decision-making. All that is to prevent the invisible enemy inherent in making many decisions: "the issue of noise and its harmful impacts invisible to the human eye" (D. Kahneman et al., 2021).

In addition to that, the way we imagine possible worlds (Parry, 1973) to reach somehow this fairness by manipulating some aspects of these possible worlds is also criticized by the authors Kasirzadeh and Smart (2021) this way: "Depending on what kind of possible worlds we choose, we might end up assigning a different truth value to a counterfactual statement" (Kasirzadeh & Smart, 2021).

Therefore, according to the experiments conducted by McCormack et al., “children who experienced regret tended to make better choices in the future, and the other important element was that this emotion helps children learn to delay gratification and behave more prosocially“ (McCormack et al., 2020). Moreover, this way of looking at the prospective counterfactual variant of regrets plays an essential role in our future planning intentions.

So, let’s continue with another example where the manipulation of anticipated regret could greatly help in strategies for improving responsible and ethical decision-making in human-AI interaction in preventing, for example, a new pandemic such as COVID-19.

However, we tend to stay with the most recent events (Byrne, 2019), and this phenomenon has consequences in the way we reason about a new pandemic: for example, we perceive the pandemic has passed, and we don't consider the possibility that something like this could happen again (or at least not so soon or not as long as we live). We may have data confirming that something like this could happen again and soon, but if we consider this, a large-scale planning strategy would be necessary, for which we would have to invest much time, coordination, cooperation, and even imagination to solve a novelty situation again. And it is the very different ways in which humans deal with such dilemmas, such as the different ways in which different countries use different ways to remedy or control a pandemic, that produces the so-called variability of human judgments (D. Kahneman et al., 2021) in interpersonal decision-making in high-risk situations that have inspired mainly this dissertation.

Therefore, what is proposed in this research is to invert certain statements., i.e., we no longer focus on what kind of counterfactuals human agents would be willing to create but on how to get possible regrets to surface in advance in the form of mental simulation in human minds. In other words, what kind of information or data could bring up those experiences of anticipated regret without being averse to them? How is this information systematized? How can the impact be measured and evaluated? Do all human agents have this capacity to feel/think the anticipated regret? What is its variation across humans with the same role as decision-makers in different cultures? What kind of concrete role should explanations of AI systems play in supporting us in this vital task?

Given that Pearl and Mackenzie's (2018) framework levels two and three are not very different, for our proposal, both levels are distinguished from each other as follows. In this second level of intervention, which corresponds to our second phase of how explanations must be communicated to be effective, our aim is that the explanations provided by the AI agents try to bring out in the users that they express or communicate their regrets. In this way, the participants are confronted with various information as tangible examples of prospective counterfactual scenarios (Byrne, 2016; McEleney & Byrne, 2006; Santamaría et al., 2005).

The novelty of this approach is that instead of waiting for an AI system to explain how it has arrived at a particular or global output to us, which is the approach of data-driven XAI, we program them to help us in a particular task. In this case, this would offer counterfactual scenarios about the future regrets we would face if we did not prevent them beforehand.

Thus, the query for this second phase could be formulated as follows: how could an AI system be trained at this level to explain to us in images or other unknown information formats that could contribute to driving the simulation of regret in decision-makers in high-stakes situations?

As explained in the first level, much attention is paid to the fact that the explanatory architecture consists of a hybrid explanatory theory in this framework. Therefore, the abovementioned examples offer two ways of proceeding with this task. One is to combine the work of Borgo et al., which is called the XAI-Plan model. In summary, this model provides an immediate explanation for the decision made by the agent planner. In this way, the model produces explanations by encouraging users to try different alternatives in the plans and compares the subsequent plans with those of the planner. The interactions between the planner and the user enhance hybrid-initiative planning that can improve the final plan (Borgo et al., 2018; Sado et al., 2020). Conversely, in the area of eXplainable Reinforcement Learning (XRL), such as the work by Cruz et al., an RL agent explains to the users why it selected an action over the other possible actions. This model uses an episodic memory (Sequeira & Gervasio, 2020) to save each episode or agent's record of executed state-action combinations, then computes both the likelihood of success (Q-values) as well as the number of transitions within each episode to meet the final target to provide an explanation or reason for selecting an action over the others (Cruz et al., 2019).

It is important to use strategies such as the two immediately mentioned because the ability to create counterfactual worlds or the capacity to create infinite mental simulations about something does not lead to anything. According to the neuroscientist Lee (2017), both too much mental stimulation and too little can be considered mental disorders, and our capacity to measure this is beyond currently available techniques, or if any of them are available, they are still too invasive (Lee, 2017). In addition, along the lines of planning or prioritizing what kind of information to select and even save for the future, we also have the learning theory described above. In this line, we rescue Tolman's latent learning theory, as Lee states:

> "The reinforcement learning theory can account for the same phenomenon as latent learning using the concept of mental simulation. In mental simulation, animals can predict the hypothetical outcomes of various actions based on their knowledge of the environment and adjust the values of corre-sponding actions by comparing such hypothetical outcomes to previously expected outcomes" (Lee, 2017).

But then, back to the topic: how do we get decision-makers to experience an anticipated regret that allows them to plan their action strategies to prevent their deepest regrets?

How can this be achieved? To do so, we must return to this chapter's beginning. In particular, we concentrate on the " affordance " conceptual approach (Gibson, 1979; Greeno, 1994; Norman, 1999; Nye & Silverman, 2012).

The phenomenon of "affordance" is described by Mahadevan (2018) as how a child can perceive a chair. Thus, according to Mahadevan (2018), "a chair for a child may serve as a hiding place, by crouching under it, or a stool, to retrieve another object placed beyond reach on a high table" (Mahadevan, 2018). Ultimately, it is about producing narratives that can even be in art that achieve the impact we have set beforehand to reach our goals.

An example of affordance can be the one proposed by Kahneman and co (2021). According to these authors, the act of washing your hands is a preventive action against a variety of germs that cause diseases. You are not washing your hands to prevent a specific disease, but this intentional act is done to prevent other types of reactions that you, at the present moment, cannot imagine (D. Kahneman et al., 2021). Nevertheless, this type of strategy has a "but," as seen below.

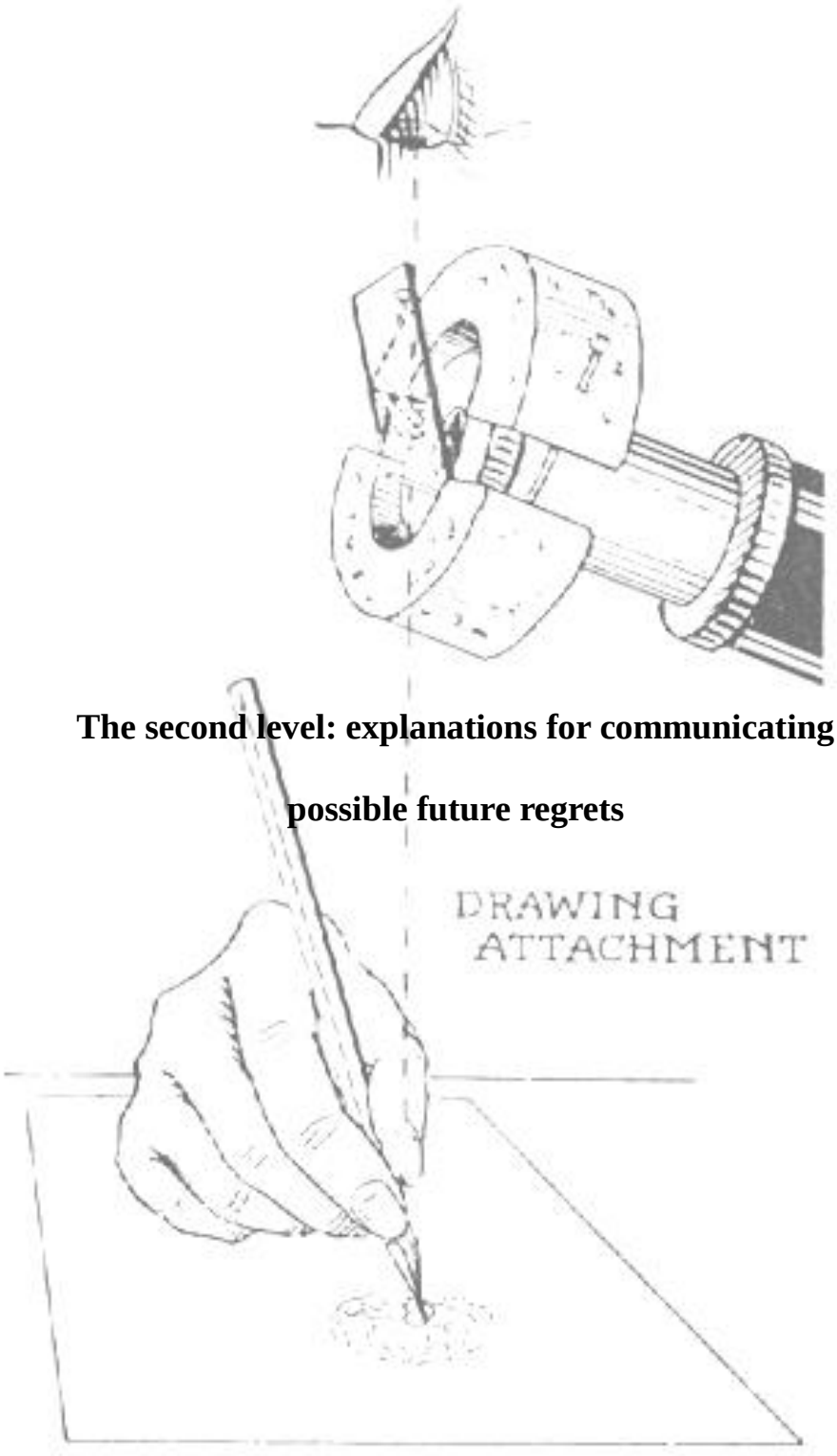
So, how can we create formats or contents in an interdisciplinary group that enables us to communicate possible futures to plan our decision strategies better?

This question has already introduced the other important verb in this framework: to communicate. Hence, how can our anticipated regrets tell us the direction in which our plan strategy should be implemented?

To understand a little more about how the communication of possible future regrets could be reflected in a more precise strategy, we can go back to the years before the COVID-19 pandemic. At this point, the question is why, despite much information pointing to the possibility of a health catastrophe such as the one that occurred, much more was not done earlier (Allam, 2020; Allam & Jones, 2020; Avishai, 2020; Cheng et al., 2007; Merlot, 2020).

Hence, why did this information not make it onto political agendas? This dissertation starts from the premise that these raw data do not communicate directly with decision-makers because they are not trained to anticipate possible regrets, and they are not legally bound by law to feel responsible for the lack of action in the face of events of this magnitude, and there are no sanctions that link their incompetence. However, the most plausible explanation seems to be a lack of systematization or a theory that puts emotions in their rightful place in the decision-making process and, therefore, in the legal and medical systems.

Try to imagine that before the pandemic, we had this kind of technology: what would have happened if an AI system had offered you as an AI explanation for taking a possible pandemic seriously the image of your two-year-old child in a hospital with a mask to breathe? What if an AI system can reconstruct with your voice a situation where you can barely speak because you are short of breath, and this is provided to you by an AI as an audio format? Would it not be a matter of using certain advances in generative AI but for purposes different from what we currently have?



The second level: explanations for communicating possible future regrets

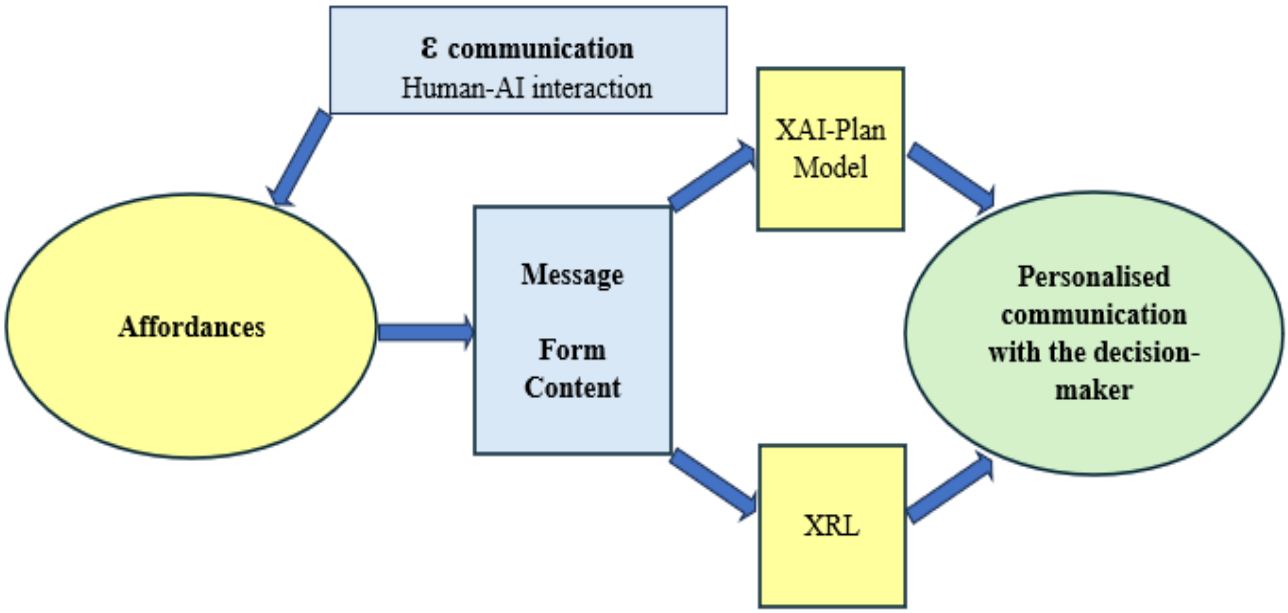


FIGURE 44 | Explanations for communicating possible future regrets (Own contribution based on the authors cited above)

The third level: what if anticipated regret could work as a recommendation norm?

This level concludes the multi-ethical framework for responsible AI. Remember that this level can only be accessed if levels one and two have been completed. So, the question that speaks to this level for this particular case is: what would happen if an AI system could detect that someone has not heeded its suggestions even though the person has understood the seriousness of an issue where many people's lives are involved? Who would be responsible in this case for the undesired effects simply by not having planned the actions required, even knowing what harmful consequences they could have?

This crucial issue brings us back again to Damasio's Somatic Marker Hypothesis (1994), whose definition it is appropriate to repeat here:

> "They are a special instance of feelings generated from secondary emotions. Those emotions and feelings have been connected by learning to predicted future outcomes of certain scenarios" (…). Somatic markers do not deliberate for us. They deliberately highlight some options and eliminate them rapidly from subsequent consideration" (Damasio, 1994).

However, it should be remembered that the version of the somatic marker used in this dissertation is a somewhat more radical view advocated by Overskied (2021). He emphasizes the role of emotion from the beginning of decision-making as a motivating element to the end that corresponds to the behavior or decision taken.

The other specific aspect that somatic markers fulfill is that they are required to form effective personal and social behavior to form an adequate "understanding" of their minds and the minds of others. This phenomenon is called "social intelligence" (Lee, 2017). This aspect fits with the objectives of this dissertation since the so-called "social intelligence" by Lee (2017) is an expression of collective and intentional intelligence.

But again, what if there is no way to bring out the anticipated regrets in human agents or institutions, cultures, or countries? This legal, even medical aspect should be discussed because, according to Damasio (1994), the moment in which a brain or a culture is defective, "somatic markers fail to fulfill its adaptive function" (Damasio, 1994).

The neuroscientist takes as an example "sick individuals," people who develop sociopathy or psychopathy. As "sick cultures," Damasio (1994) mentions German and Russian society in the 1930s and 1940s, China during the Cultural Revolution, and Cambodia during the Pol Pot regime. But how do we measure whether a human being or a culture is healthy enough to make ethical and responsible decisions? How can it be materialized?

This point may be one of the most important of the dissertation because it discusses how anticipated regret as recommendation norm could be put into practice in a social context of a group of members, culture, or political organization. To begin with, the norm recommendation concept applied in this framework is a new and updated version of the same term used by Mahoud et al. in their work. Their recommendation term was initially taken from the OP-RND framework (Ahmad et al., 2011). However, as mentioned above, this framework takes the definition of Mahoud et al. According to them, a recommendation norm "represents actions or behavior of agents judged by the community as noble or altruistic, hence merit for rewards. This norm rewards an agent if the agent exercises it but is not penalized otherwise" (Mahmoud et al., 2014).

As a reminder, this last phase corresponds to the reception of the explanations (Neerincx et al., 2018) or, more specifically, how well the human being understands the explanation offered by the AI system. In our framework, this third phase of explanation is called "ε reception" (Tiddi et al., 2015), which is materialized in two options: on the one hand, the decision-maker takes into consideration the system's explanations with a concrete plan to prevent possible future regrets. On the other hand, if the decision-maker decides to reject the system's recommendations, there is no direct penalty. However, the fact that there is a reward means that there is somehow an indirect penalty.

A positive evaluation over a long period is necessary and sufficient for decision-makers to be rewarded for their performance. Under these conditions, decision-makers become role models in the community they are enrolled in and can transfer their "knowledge" to other communities with similar objectives. Here, “self-enforcement” (Mahmoud et al., 2014) plays an important role.

According to Hollander and Wu, this kind of enforcement is also called “internally directed enforcement.” This learning strategy occurs when decision-making agents punish themselves for violating decision-making already endorsed and accepted by all society, culture, or country members. Some emotions influence this process, as outlined by Neerincx et al. Indeed, according to von Scheve et al. and Staller and Petta, emotion is a critical factor that drives self-enforcement (Hollander & Wu, 2011; Scheve et al., 2005; Staller & Petta, 2001).

However, the emotions that drive decision-makers' self-enforcement to execute one action can vary greatly. Each emotion triggers different actions and carries a different moral burden. Unfortunately, these other emotions are beyond the scope of this research since the emotion that has interested us and accompanied us during the beginning of this doctoral thesis has been regret for its role in regulating decision-making.

We still know little about what motivates human beings to act, but everyone is moved by what matters to them. It is the same as saying that when something causes you pain, you care about it. When it touches you in your body, like when you get burned, you get hurt, you have an accident.

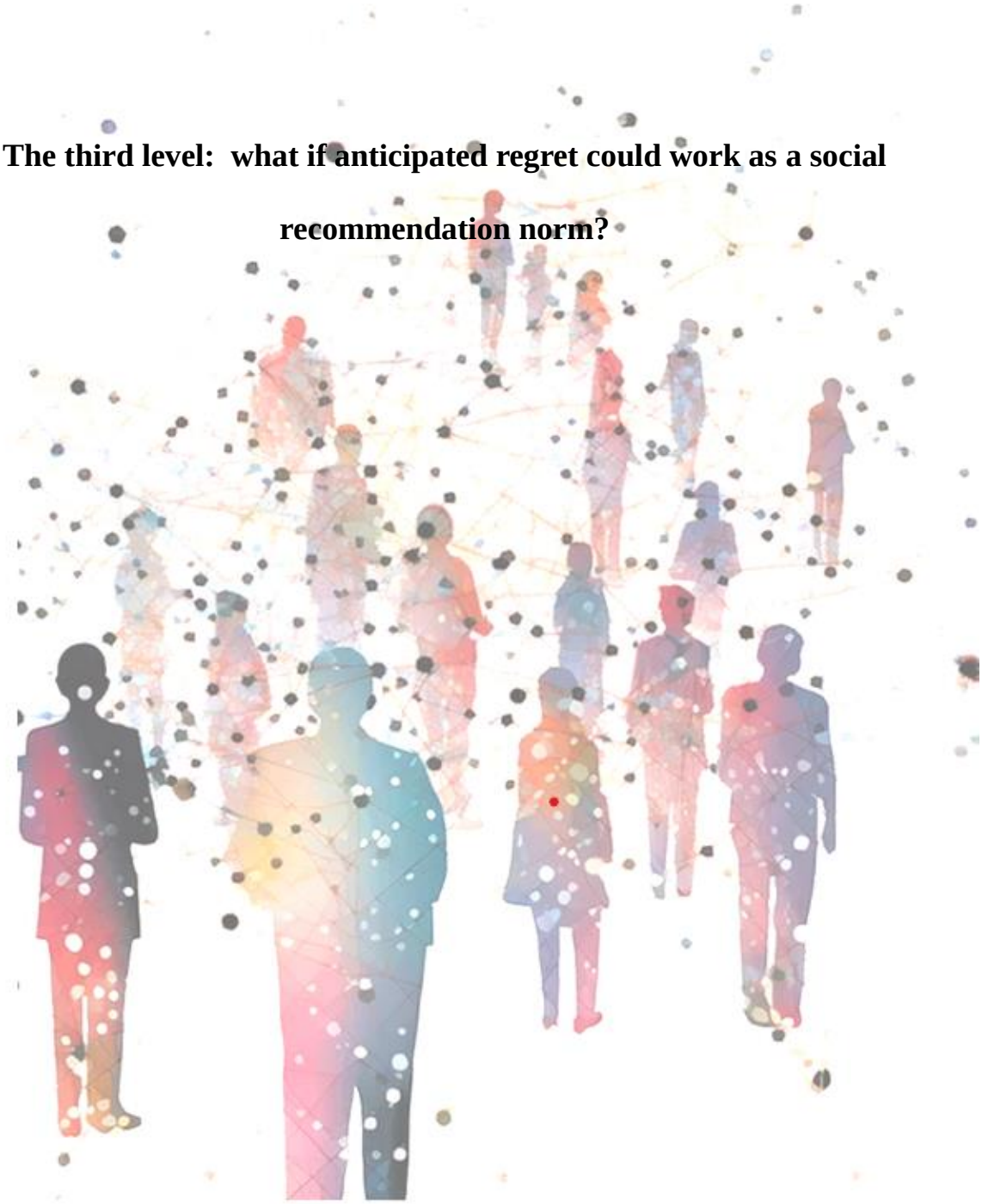
At that moment, you can feel what others think/ feel in the same or a similar situation. We can continue to create “naked technology” because it is still randomly linked data that works in some way like us. They can write texts, even poems, and they can make us laugh, but all of this differs from evolving as a species and understanding what makes us human.

Why not use technology to realize on a deeper level how we make decisions and how others might feel? The rules that emerge from these experiments/cases would be more inclusive and embrace top-down and bottom-up approaches based on considering each case's universality and particularities.

This view of decision-making automatically generates more trustworthy and responsible decisions. In addition, what does seem clear is that the power of emotions in decision-making still seems to be underestimated. Thirty years after Damasio made public his somatic marker

hypothesis, we have not yet bothered to understand scientifically in a coordinated way whether there may be emotions that can help us improve as a species through the materialization of collective responsible intelligence.

The third level: what if anticipated regret could work as a social recommendation norm?



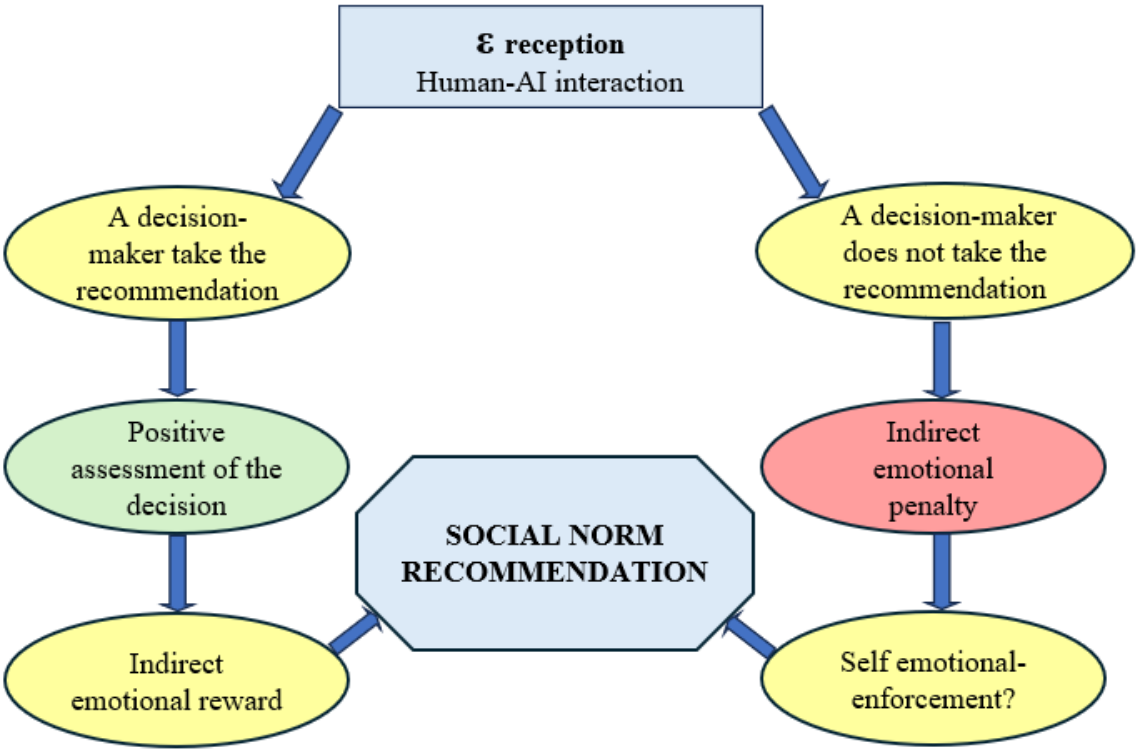
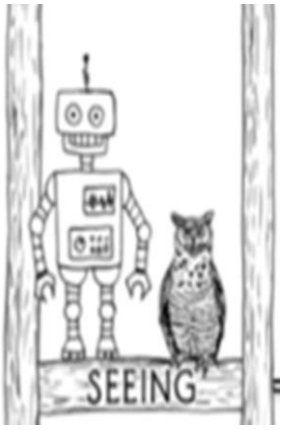


FIGURE 45 | What if anticipated regret? (Own contribution based on the authors cited above)

An approach to the moral development of Confucian ethics:

The Ladder of Regret



1. OBSERVATION/ DETECTION	**ε GENERATION: UNDERSTANDING NEW PATTERNS OF REGRET** **ACTIVITY:** Detecting the variability of regret in its retrospective and anticipatory variants in specific contexts to establish an objective understanding of this emotion. **QUESTIONS:** What would happen if the regret experience differed among individuals of the same culture? What are the differences in behaviors between the anticipatory and retrospective variants of regret? **EXAMPLES:** Does regret impact body states, or is it only a belief or a state of mind? How is the variability of regret aversion expressed by decision-makers with the same information and in the same specific contexts?

TABLE 8 | The first level of the Ladder of Regret (Own contribution based on the authors cited above)

2. PRACTICAL REFLECTION	**ε COMMUNICATION: BRINGING OUT POSSIBLE FUTURE REGRETS** **ACTIVITY:** In terms of messages, what kind of formats and contents do explanations need to have to fulfill their purpose? **QUESTIONS:** What if one human agent cannot experience/express regret? **EXAMPLES:** What if an AI system explains in different formats that a pandemic like COVID-19 could happen and if the decision-maker doesn't react? What should we do?

TABLE 9 | The second level of the Ladder of Regret (Own contribution based on the authors cited above)

3. PRACTICE /NEW SOCIAL SITUATION	**ε RECEPTION: ANTICIPATED REGRET AS A SOCIAL NORM RECOMMENDATION** **ACTIVITY:** Assess the positive or negative impact on trustworthy human-AI interaction using goal-driven XAI that communicates with the decision-maker how beneficial it can be to anticipate regrets and plan concrete actions or behaviors. **QUESTIONS:** What would have happened if we had implemented AI systems differently before the pandemic? How can we train ourselves to anticipate regret? Can the anticipated regret be perceived as a norm recommendation? **EXAMPLES:** Suppose it could be more objectively assessed that a decision-maker cannot experience regret. How could we regulate this phenomenon and include it within what is healthy and normal for a person with decision-making power over others?

TABLE 10 | The third level of the Ladder of Regret (Own contribution based on the authors cited above)

Before turning to the conclusions of this dissertation, it is crucial to close this framework, "the ladder of anticipated regret," with another ladder, which appears in the citation of one of the more representative philosophers in the philosophy of language. I am talking about Wittgenstein and his book *Tractatus Logico Philosophicus*. According to Wittgenstein, only when you have reached one level on the ladder can you move to the next one, which means that once you have passed a level, you no longer need the first and second rungs.

"My propositions are elucidatory in this way: he who understands me finally recognizes them as senseless, when he has climbed out through them, on them, over them. (He must, so to speak, throw away the ladder after he has climbed up on it)."

(Wittgenstein, 2015)

FIGURE 46 | Metaphor the Ladder of Understanding / Shutterstock

CHAPTER 9 CONCLUSIONS

Responsible behavior in intelligent emerging societies

9.1 Testing the proposed framework with a past example

Since the conclusions are one of the most personal parts of any scientific work, I would like to end them with real stories; some belong entirely to the past, others coexist with us at present, and others are possible future directions, some of them are my personal anticipated regrets I hope will never take place. As I wrote in the first chapter, this doctoral thesis has been configured almost entirely to coincide with the origins of the COVID-19 pandemic that left many people and countries in the world isolated. But this isolation and the measures adopted by each country to control and eradicate the pandemic were in some cases paradoxically contradictory for the same phenomenon. How is this possible? This disillusionment led me to wonder what it is that leads human beings to repeat the same mistakes over and over again. What I observed was that even knowing that we cannot control many variables and that uncertainty is a factor that exists naturally, I asked myself: if we cannot predict with total certainty, what can make us anticipate and prevent some dire consequences that are highly probable even if they are not going to happen? At the bottom, my mind was trying to identify elements of regulation of human behavior broadly understood as suffering. Therefore, I became interested in the reasoning mechanisms that allow us to foresee future scenarios in some way, so I landed on the book that has partially shaped the framework proposed by this doctoral research in the previous pages: *The Book of Why* by Pearl and Mackenzie, from which the development of the three phases of the Ladder of Regret has been taken. Well, that form of reasoning that could regulate future behavior to improve past results had a name. The sad thing is that I once again learned about the power of counterfactual reasoning through this dramatic story from the past described by these two authors in their book. And that it also has to do with disease and the death of people. For the authors of *The Book of Why*, counterfactuals are essential to how humans learn about the world and how our actions affect it. The most popular application of counterfactuals in science today is called mediation analysis. In a scientific context, a mediator, or mediating variable, transmits the treatment's effect to the outcome.

The search for mechanisms is critical to science and everyday life because different mechanisms call for different actions when circumstances change. First, it is worth examining the story as the book's authors describe, and then I discuss how this piece of the true story has helped shape the framework of the Ladder of Regret.

One of the earliest examples of controlled experiment was sea captain James Lind's study of scurvy, published in 1747. In Lind's time, scurvy was a terrifying disease, estimated to have killed 2 million sailors between 1500 and 1800. Lind established, as conclusively as anybody could at that time, that a diet of citrus fruit prevented sailors from developing this dread disease. By the early 1800s, scurvy had become a problem of the past for the British navy, as all its ships took to the seas with an adequate supply of citrus fruit. This is usually the point at which history books end the story, celebrating a great triumph of the scientific method. It seems very surprising, then, that this completely preventable disease made an unexpected comeback a century later, when British expeditions started to explore the polar regions. The British Arctic Expedition of 1875, the Jackson-Harmsworth Expedition to the Arctic in 1894, and most notably the two expeditions of Robert Falcon Scott to Antarctica in 1903 and 1911 all suffered greatly from scurvy. How could this have happened? In two words: ignorance and arrogance- always a potent combination. By 1900, the leading physicians in Britain had forgotten the lessons of a century before. Scott's physicians in Britain had forgotten the lessons of a century before.

Scott's physician on the 1903 expedition, Dr. Reginald Koettlitz, attributed scurvy to tainted meat. Further, he added, "the benefit of the so-called "antiscorbutics (i.e., scurvy preventatives, such as lime juice) is a delusion". In his 1911 expedition, Scott stocked dried meat that had been scrupulously inspected for signs of decay but no citrus fruits or juices. The trust he placed in the doctor's opinion may have contributed to the tragedy that followed. All the five men who made it to the South Pole died, two of an unspecified illness that was most likely scurvy. One team member turned back before the pole and made it back alive, but with a severe case of scurvy. With hindsight, Koettlitz's advice borders on criminal malpractice. How could the lesson of James Lind have been so thoroughly forgotten -or worse, dismissed- a century later? The explanation, in part, is that doctor did not really understand how citrus fruits worked against scurvy. In other words, they did not know the mediator.

This piece of the true story highlights many problems that have been dealt with in this doctoral thesis and can be formulated as follows: How is it possible that the same health catastrophe occurred again when the specific problem of scurvy had already been eradicated in the past?

Since we cannot go back in time and question the parties involved, we can only ask ourselves what could have gone wrong and offer a better "roadmap" for future situations that have a strong resemblance:

The most pertinent questions are:

1. Did Dr. Reginald Koetllitz have access to the discoveries made in the passage by James Lind?

    - If the answer was yes and Dr. Reginald Koetllitz or other doctors on the expedition had known about James Lind's eradication of scurvy, why didn't they use it as a prevention strategy? We are not talking about having to plan enormously or having to invest vast amounts of money. The prevention of introducing citrus fruits on the expedition could have saved lives as it was discovered in the past.
    - If the answer is not, and if the organizers and doctors who prepared the 1911 expedition had not had access to Lind's studies published in 1747, wouldn't an explanatory AI system have helped by bringing to light Lind's study published in 1747?
    - This way of proceeding would be part of what has been called latent reinforcement learning. Since this type of learning can bring out from the past a situation that at time x is similar in the present, it can help enormously to prevent the same mistakes already solved in the past.

2. If Dr. Reginald Koetllitz had known the discovery of the mediator in detail to end with the scurvy through the study James Lind published in 1747, how could we explain that he did not introduce a citrus diet on the expedition to prevent scurvy?

    - Since the words used to describe the doctors of the British expedition in the text are those of "ignorance and arrogance." We must start from the idea that the type of affective counterfactual reasoning discussed in this doctoral thesis was blocked, which in the form of mental stimulation is equated with the anticipated regret of the type: if in the past this same problem was solved, let's at least give it another try as a prevention strategy.
    - The type of reasoning employed by the doctors of the British expedition obeys according to the words "arrogance and ignorance" to what Kahneman and his colleagues have discussed in their book *Noise*, that is, that reward of feeling that one is right and that blocks any other type of counterfactual reasoning in which emotions that cause some negative impacts are blocked.

3. The last and perhaps most salient question to answer is: what would have happened if it had been possible to monitor the decision-making of these English physicians in the past four years by having an AI assistance system as described in Chapter 8 corresponding to the framework, i.e., through the three phases?

    - Undoubtedly, the most crucial part of this question is how the general population would have reacted. Would physicians have been held responsible for their decision-making? If the population had managed to dismiss the knowledge of these physicians by holding them responsible for their poor or even wrong decision-making, would they have changed their future decision-making?

9.2 The dilemma of imaginative intelligence

FIGURE 47 | Evolution of human intelligence / Pixabay

Thanks to new studies in animal behavior and other species, it cannot be claimed that the imaginative processes necessary to plan goal-directed behaviors are solely the territory of human intelligence or imagination. However, given the objective to be pursued, this doctoral thesis has had to limit itself to the study of which imaginative processes understood as counterfactual reasoning push human behavior to opt for a specific decision. The most salient part has been the still little systematized study of whether emotions function as architectures of human behavior.

The hypothesis that has been put forward is that emotions regulate behavior, and each emotion is loaded with an ethical or moral component that needs to be explored urgently, as we have tried to justify from the beginning to the end of this research. This view is why biases and noise in decisions should be investigated as part of what we need to understand more thoroughly, not simply labeled as failures in reasoning because, somehow, we are trying to leave out key aspects that can give us answers to how human intelligence has evolved and why catastrophes caused by this same species happen again and again. But then, why does this section of the conclusions refer to the dilemmas of imaginative intelligence and open with the image of artifacts designed, created, and implemented in the Stone Age? No matter how far ahead we may think we are from our ancestors, we still bear remarkable similarities.

The legacy of human culture seems to align with the complex advances that have co-occurred at these three levels: the creation and design of technological tools to adapt to the environment and survive, our symbolic capacity to use language, and our understanding of social content.

However, these complex advances have allowed us to survive, adapt to various situations, and consolidate our cultures from generation to generation. However, the reading that can be made of human imagined intelligence may be quite different. Therefore, I show you another largely ignored use of these capabilities in this series of headlines and other images, emphasizing its ethically controversial aspects.

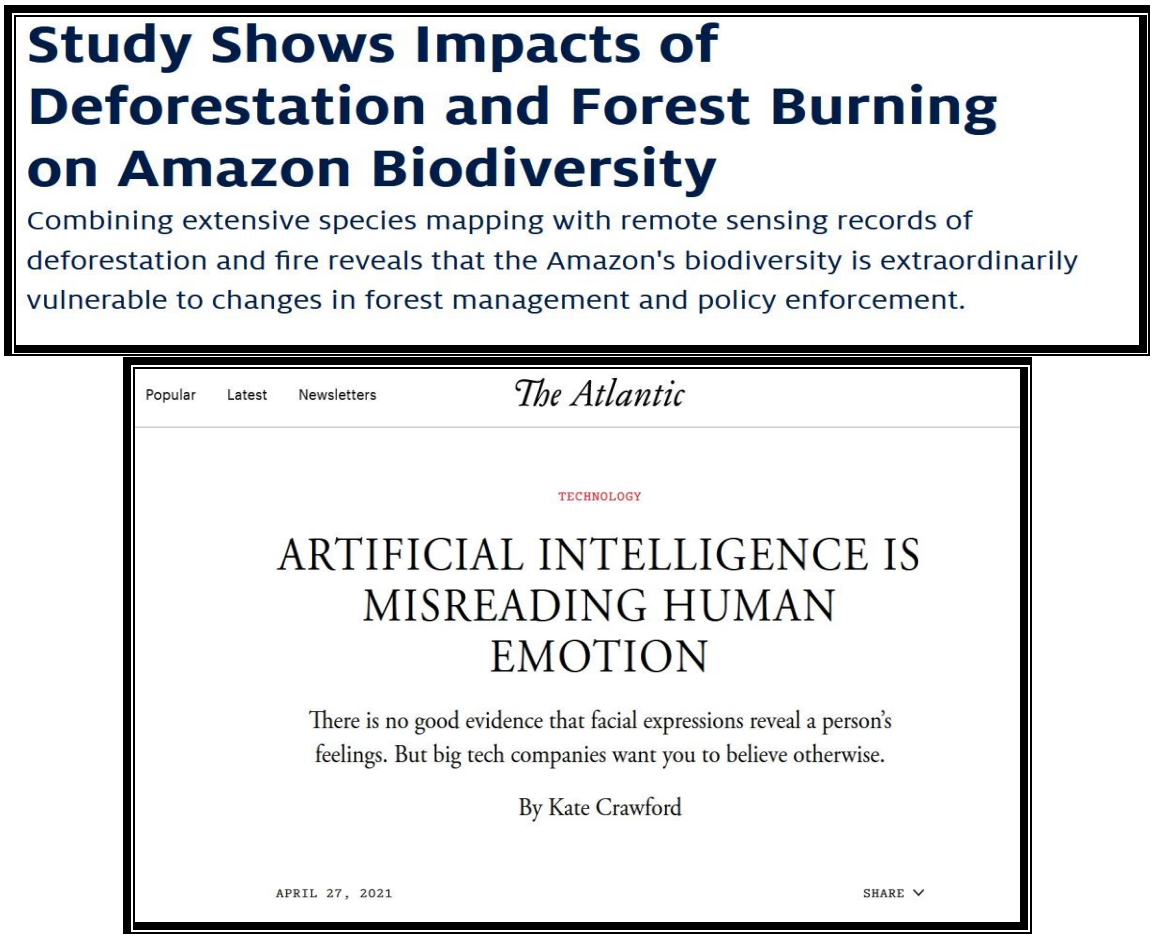


FIGURE 48 | Study shows impacts of deforestation (Stolte, 2021) Artificial intelligence is misreading human emotion (Crawford, 2021)

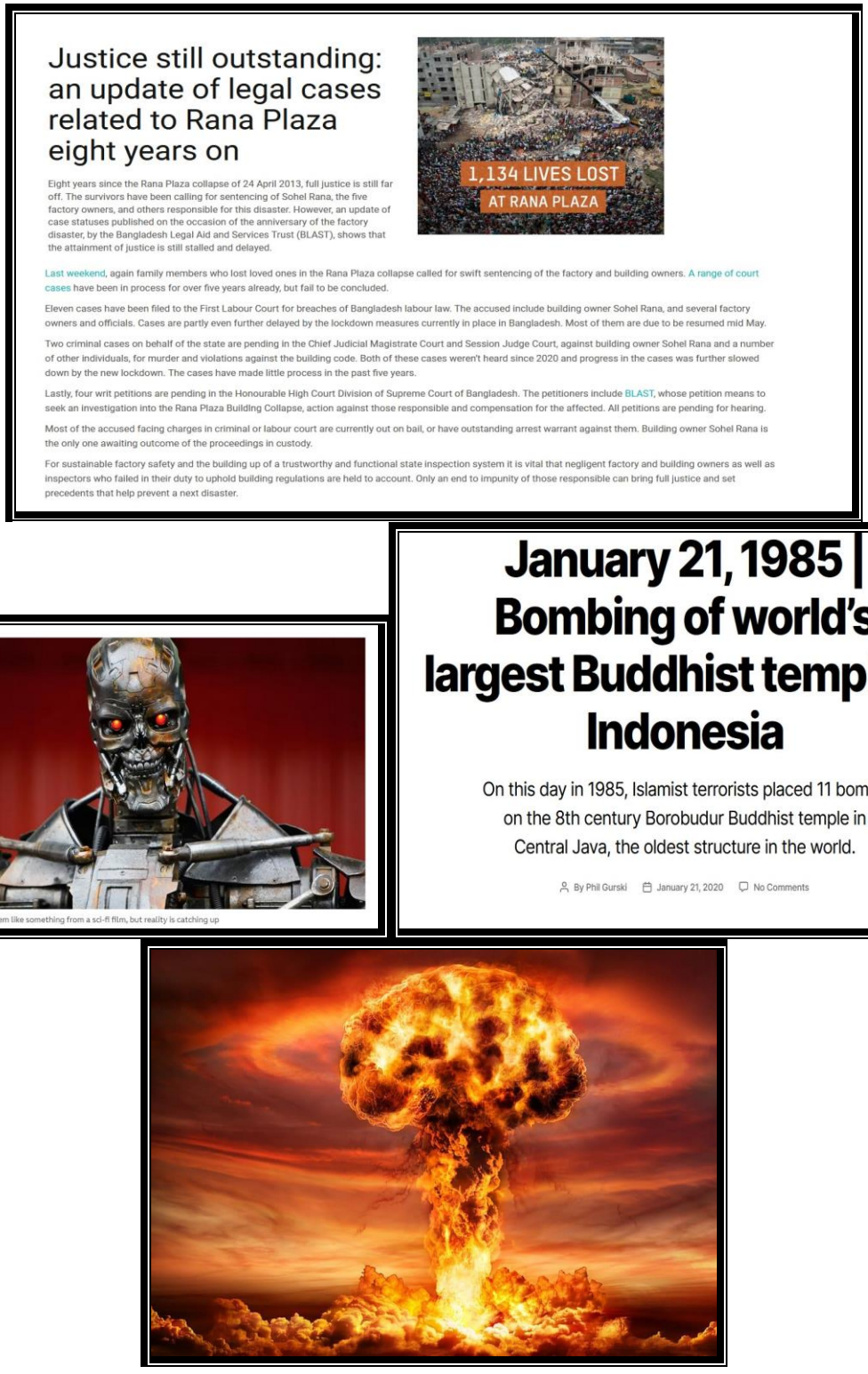


FIGURE 49 | Justice still outstanding (Cleanclothes.org, n.d.) / Killer robot / bbc.com / Bombing of world's largest Buddhist temple (Gurski, 2020) / Atomic Bomb / iStock

What strikes me as a thought-provoking exercise is to compare our stone-age skills with the latest images and news headlines. Both images are products of human behavior and even represent what the intelligent imagination can develop and achieve. No one could say in a time of war that the atomic bomb was considered unethical, but we can say it today.

In the same way, today, we do say that the deforestation of the Amazon may be one of the causes that accelerate the development of climate deterioration and complicate the preservation of the human race as we know it when the climatic effects of droughts, earthquakes, and floods occur more and more often.

It is indisputable that humans have made remarkable discoveries that were unthinkable 100 years ago. However, the unfinished business of predicting the unwanted consequences of our imaginative intelligence remains our problem.

This dissertation has tried to argue the importance of looking inside human beings to understand what guides them in imagining possible futures.

The hypothesis supported from the beginning is that we cannot diagnose our decision-making only by observing how we make decisions. We must understand the world of emotions and how they permeate our imaginative intelligence and ability to predict what we do not want to happen.

In addition, one thing that has permeated the evolution of the human species is the following thought: in resolving one problem, the emergence of others is the dominant tonic. I therefore dare to conclude this doctoral thesis with my anticipated regrets.

9.3 The specific use and misuse of emotions: possible anticipated regrets

This doctoral thesis comes to an end with this section. Every page corresponds to a journey on the ethics of responsibility in trustworthy human-AI interaction. Although the proposed framework approach is deeply theoretical, it could help us understand new elements of reality, how humans make decisions, and other issues such as biases and noise from a not-so-reductionist and more integrative approach. Furthermore, the latest and newest advances in measuring emotions to see how they influence our behavior have led to the European AI legislation, called the AI Act, which proposes specific prohibitions such as measuring emotions in educational settings and work environments to monitor workers' performance. However, as humans know, we will always find loopholes to make the most of these measurements without running up against the law. One of the dangers that I can anticipate is the misuse of the informative power of emotions as data has started to be used since the advent of big data, but we do not know what kind of correlations will be made. I warn you of the importance of a rigorous study of them to avoid unintended consequences in the future by anonymizing these data and providing clear normative frameworks on what is allowed and what is not. To this end, I opt for a new, specific, interdisciplinary study of these kinds of data to which I would like to add another type of data that, although not dealt with directly in this doctoral thesis, could be of enormous importance in the future, not only because of their philosophical significance but also because they will be necessary if we want to understand the cognitive impact of emotions: I am talking about neural data.

What worries me, though, is the so-called generative but embodied AI. Data is undoubtedly the gold of knowledge; the output can be most unexpected depending on what kind of data we put together. I can't predict the future, but we humans have always been brazening and blatant with our intelligence without worrying about the impact of our inventions. I am also aware that I am not free from making the same mistakes as others, which may be the human blind spot, but I hope we will be able to combine intelligence and responsibility. Perhaps if we establish a new science that combines emotions and their ethical impact on human behavior with the help of AI systems such as the one described in Chapter 8, we could reach a new stage in knowledge and understanding of what characterizes us as a human species or at least as individuals who aspire to a responsible collective intelligence.

LIST OF 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

Aday, J., Rizer, W., & Carlson, J. M. (2017). Neural Mechanisms of Emotions and Affect. *Emotions and Affect in Human Factors and Human-Computer Interaction*, 27–87. https://doi.org/10.1016/B978-0-12-801851-4.00002-1

Ahmad, A., Zaliman, M., Yusof, M., Ahmad, M. S., Ahmed, M., & Mustapha, A. (2011). Resolving conflicts between personal and normative goals in normative agent systems. *2011 7th International Conference on Information Technology in Asia: Emerging Convergences and Singularity of Forms - Proceedings of CITA'11*. https://doi.org/10.1109/CITA.2011.5999538

Alfaisal, A., & Aljanada, R. (2018). Universal Emotions. *The International Journal of Humanities & Social Studies*. https://doi.org/10.24940/THEIJHSS/2018/V6/I12/HS1812-051

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable Control and Counterfactual Reasoning in the Psychology of Blame. *Personality and Social Psychology Bulletin*, 34(10), 1371-1381. https://doi.org/10.1177/0146167208321594

Allam, Z. (2020). The Rise of Machine Intelligence in the COVID-19 Pandemic and Its Impact on Health Policy. In *Surveying the Covid-19 Pandemic and its Implications*. https://doi.org/10.1016/b978-0-12-824313-8.00006-1

Allam, Z., & Jones, D. S. (2020). Pandemic stricken cities on lockdown. Where are our planning and design professionals [now, then, and into the future]? *Land Use Policy*, Sep:*97*. https://doi.org/10.1016/j.landusepol.2020.104805

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155.

https://doi.org/10.1007/S10676-006-0004-4/METRICS

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12–17. https://doi.org/10.1109/MIS.2006.83

Anderson, M., & Anderson, S. L. (2011). Machine Ethics. *Machine Ethics*, *9780521112352*, 1–538. https://doi.org/10.1017/CBO9780511978036

Andery, M. A., Micheletto, N., & Sério, T. M. (2005). Meaning and Verbal Behavior in Skinner's Work from 1934 to 1957. *The Analysis of Verbal Behavior*, *21*(1), 163. https://doi.org/10.1007/BF03393018

Angie, A. D., Connelly, S., Waples, E. P., & Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: a meta-analytic review. *Cognition & Emotion*, *25*(8), 1393–1422. https://doi.org/10.1080/02699931.2010.550751

Anjomshoae, S.T., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable Agents and Robots: Results from a Systematic Literature Review. *Adaptive Agents and Multi-Agent Systems.*

Assad, M. L. (1991). From Order to Chaos: Michel Serres's Field Models. *SubStance*, *20*(2), 33. https://doi.org/10.2307/3684967

Avishai, B. (2020). *The Pandemic Isn't a Black Swan but a Portent of a More Fragile Global System | The New Yorker*. https://www.newyorker.com/news/daily-comment/the-pandemic-isnt-a-black-swan-but-a-portent-of-a-more-fragile-global-system

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*. https://doi.org/10.2139/SSRN.2477899

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Barto, A. G. (1994). Reinforcement learning control. *Current Opinion in Neurobiology*, *4*(6), 888–893. https://doi.org/10.1016/0959-4388(94)90138-4

Bauman, Z., & Haugaard, M. (2008). Liquid modernity and power: A dialogue with Zygmunt Bauman1. *Journal of Power*, *1*(2). https://doi.org/10.1080/17540290802227536

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, Decision Making and the Orbitofrontal Cortex. *Cerebral Cortex*, *10*(3), 295–307. https://doi.org/10.1093/CERCOR/10.3.295

Beer, D. (2017). The social power of algorithms. In *Information Communication and Society* (Vol. 20, Issue 1). https://doi.org/10.1080/1369118X.2016.1216147

Beike, D. R., Markman, K. D., & Karadogan, F. (2009). What We Regret Most Are Lost Opportunities: A Theory of Regret Intensity. *Personality and Social Psychology Bulletin*, 35(3), 385-397. https://doi.org/10.1177/0146167208328329

Bentham, J. (1789) *An Introduction to the Principles of Morals and Legislation*. Clarendon Press, Oxford. http://dx.doi.org/10.1093/oseo/instance.00077240

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. Psychological Bulletin, 142(11), 1227–1274. https://doi.org/10.1037/BUL0000067

Biran, O. & Cotton, C. (2017). Explanation and Justification in Machine Learning: A survey. In IJCAI-17 *Workshop on Explainable AI (XAI)*.

Birnhack, M. D. (2008). The EU Data Protection Directive: An engine of a global regime. Computer Law & Security Review, 24(6), 508–520. https://doi.org/10.1016/J.CLSR.2008.09.001

Boella, G., & Torre, L. van der. (2007). A Game-Theoretic Approach to Normative Multi-Agent Systems. *Volume 07122, of Dagstuhl Seminar Proceedings, 2007*. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany

Boffey. (2018). EU border 'lie detector' system criticised as pseudoscience | Migration | The Guardian. https://www.theguardian.com/world/2018/nov/02/eu-border-lie-detection-system-criticised-as-pseudoscience

Borgo, R., Cashmore, M., & Magazzeni, D. (2018). Towards Providing Explanations for AI Planner Decisions. https://arxiv.org/abs/1810.06338v1

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets* 2023 33:1, 33(1), 1–30. https://doi.org/10.1007/S12525-023-00644-5

Breugelmans, S. M., Zeelenberg, M., Gilovich, T., Huang, W. H., & Shani, Y. (2014). Generality and cultural variation in the experience of regret. *Emotion*, 14(6), 1037–1048. https://doi.org/10.1037/A0038221

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 26, 355 - 372.

Brynjolfsson, E. (2018). Where Humans Meet Machines: Intuition, Expertise and Learning | by MIT IDE | MIT Initiative on the Digital Economy | Medium. https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade

Buchanan, B. G., & Shortliffe, E. H. (1984). Knowledge Engineering. Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (*The Addison-Wesley Series in Artificial Intelligence*), 148–159. https://books.google.com/books/about/Rule_based_Expert_Systems.html?hl=de&id=0uZQAAAAMAAJ

Buchanan, J., Summerville, A., Lehmann, J., & Reb, J. (2016). The regret elements scale: Distinguishing the affective and cognitive components of regret. *Judgment and Decision Making*, 11(3). https://doi.org/10.1017/s1930297500003107

Bundy, A. (2017). Preparing for the future of Artificial Intelligence. *Ai & Society*, 32(2), 285–287. https://doi.org/10.1007/S00146-016-0685-0

Butler, A. B. (2009). Triune brain concept: A comparative evolutionary perspective. Encyclopedia of Neuroscience, 1185–1193. https://doi.org/10.1016/B978-008045046-9.00984-0

Byrne, R. M. J. (2016). Counterfactual thought. Annual Review of Psychology, 67, 135–157. https://doi.org/10.1146/ANNUREV-PSYCH-122414-033249

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. IJCAI International Joint Conference on Artificial Intelligence, 2019-August, 6276–6282. https://doi.org/10.24963/IJCAI.2019/876

Byrne, R. M. J. (2020). The Counterfactual Imagination: The Impact of Alternatives to Reality on Morality. The Cambridge Handbook of the Imagination, 529–547. https://doi.org/10.1017/9781108580298.032

Byrne, R. M. J., & Timmons, S. (2018). Original Articles Moral hindsight for good actions and the effects of imagined alternatives to reality. Cognition, 1–10. https://doi.org/10.1016/j.cognition.2018.05.010

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J. R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. Science (New York, N.Y.), 304(5674), 1167–1170. https://doi.org/10.1126/SCIENCE.1094550

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and Risks of Machine Ethics. Proceedings of the IEEE, 107(3), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It Takes Two to Tango: Towards Theory of AI's Mind. https://arxiv.org/abs/1704.00717v2

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A. F. T., & Yampolskiy, R. (2017). Towards Moral Autonomous Systems. http://arxiv.org/abs/1703.04741

Cheng, V. C., Lau, S. K., Woo, P. C., & Yuen, K. Y. (2007). Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. Clinical microbiology reviews, 20(4), 660–694. https://doi.org/10.1128/CMR.00023-07

Chow, S. J. (2013). What's the Problem with the Frame Problem? *Review of Philosophy and Psychology*, 4(2), 309–331. https://doi.org/10.1007/S13164-013-0137-4

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591–621. https://doi.org/10.1146/ANNUREV.PSYCH.55.090902.142015

Clark, E. A., Kessinger, J., Duncan, S. E., Ann Bell, M., Lahne, J., Gallagher, D. L., & O'Keefe, S. F. (2020). The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review. *Frontiers in Psychology*, 11, 507534. https://doi.org/10.3389/FPSYG.2020.00920/BIBTEX

Clynes, M. (1988). Generalised Emotion How it may be produced, and Sentic Cycle Therapy. *Emotions and Psychopathology*, 107–170. https://doi.org/10.1007/978-1-4757-1987-1_6

Cobb, C. D., & Mayer, J. D. (2000). Emotional intelligence. *Educational Leadership*, 58(3), 14–18. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221. https://doi.org/10.1007/S10676-010-9235-5/METRICS

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1), e1391. https://doi.org/10.1002/WIDM.1391

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature Neuroscience*, 8(9), 1255–1262. https://doi.org/10.1038/NN1514

Coricelli, G., Dolan, R. J., & Sirigu, A. (2007). Brain, emotion and decision making: the paradigmatic example of regret. *Trends in Cognitive Sciences*, 11(6), 258–265. https://doi.org/10.1016/J.TICS.2007.04.003

Crawford, K. (2021). Artificial Intelligence Is Misreading Human Emotion - The Atlantic. https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696/

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature* 2016 538:7625, 538(7625), 311–313. https://doi.org/10.1038/538311ª

Cristianini, Nello (2023). The Shortcut - Why Intelligent Machines Do Not Think Like Us. *Boca Raton*, Florida: CRC Press.

Cruz, F., Dazeley, R., & Vamplew, P. (2019). Memory-Based Explainable Reinforcement Learning. *Lecture Notes in Computer Science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11919 LNAI, 66–77. https://doi.org/10.1007/978-3-030-35288-2_6/COVER

Damasio, A. R. (2005). *Descartes' error: emotion, reason, and the human brain*. London; New York, Penguin

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray. http://dx.doi.org/10.1037/10001-000

Davidsson, P. (1996). Autonomous Agents and the Concept of Concepts. https://portal.research.lu.se/en/publications/autonomous-agents-and-the-concept-of-concepts

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299, 103525. https://doi.org/10.1016/J.ARTINT.2021.103525

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2), 101666. https://doi.org/10.1016/J.GIQ.2021.101666

de Paiva-Silva, A. I., Pontes, M. K., Aguiar, J. S. R., & de Souza, W. C. (2016). How do we evaluate facial emotion recognition? Psychology and Neuroscience, 9(2), 153–175. https://doi.org/10.1037/PNE0000047

de Sousa, R. (1987). *The Rationality of Emotion*. MIT Press. https://doi.org/10.7551/MITPRESS/5760.001.0001

Dewey, J. (1925). *Experience and Nature*. Chicago: Dover.

Diakopoulos, Nicholas. (2015). Accountability in algorithmic decision-making. *Queue*. 13. 10.1145/2857274.2886105.

Dietrich, Eric. (2007). After the humans are gone. J. Exp. Theor. *Artif. Intell*.19. 55-67.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. https://arxiv.org/abs/1702.08608v2

Dreyfus, H. L. (1978). Cognitivism vs. Hermeneutics. *Behavioral and Brain Sciences*, 1(2), 233–234. https://doi.org/10.1017/S0140525X00074239

Du, M., Liu, N., & Hu, X. (2018). Techniques for Interpretable Machine Learning. *Communications of the ACM*, 63(1), 68–77. https://doi.org/10.1145/3359786

Damiano, Luisa & Dumouchel, Paul (2017). *Living with Robots*. Harvard University Press.

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2017). Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 81–87. https://doi.org/10.1145/3278721.3278736

Ekman, P. (2005). Basic Emotions. *Handbook of Cognition and Emotion*, 45–60. https://doi.org/10.1002/0470013494.CH3

Ekman, P. (2017). *Facial expressions*. In J.-M. Fernández-Dols & J. A. Russell (Eds.), The science of facial expression (pp. 39–56). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190613501.003.0003

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: guidelines for research and an integration of findings.* Pergamon Press. http://www.sciencedirect.com:5070/book/9780080166438/emotion-in-the-human-face

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). *Pan-cultural elements in facial displays of emotion*. Science (New York, N.Y.), 164(3875), 86–88. https://doi.org/10.1126/SCIENCE.164.3875.86

Eknian, P. (1980). Biological and Cultural Contributions to Body and Facial Movement in the Expression of Emotions. In A. O. Rorty (ed.), *Explaining Emotions*. Univ of California Pr.

Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, 132(1), 33–42. https://doi.org/10.1007/S11098-006-9054-Z

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2), 168–192. https://doi.org/10.1177/1088868308316091

Epstude, K., & Roese, N. J. (2011). When goal pursuit fails: The functions of counterfactual thought in intention formation. *Social Psychology*, 42(1), 19–27. https://doi.org/10.1027/1864-9335/A000039

Etzioni, O. (2017). *How to regulate artificial intelligence*. New York Times, September 1.

European information commisisioners office. (2021). Guide to the General Data Protection Regulation (GDPR). European Data Protection.

Evans, J. S. B. T., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory and Cognition*, 35(7), 1772–1784. https://doi.org/10.3758/BF03193509/METRICS

Feenberg, Andrew (1991). *Critical theory of technology*. New York: Oxford University Press. Edited by Jan Kyrre Berg Olsen Friis, Stig Andur Pedersen & Vincent F. Hendricks.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain.* Houghton Mifflin Harcourt.

Feldman Barret, L. (2020). *Seven and a Half Lessons About the Brain*. Houghton Mifflin Harcourt.

Feldman Barrett, L. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1–23. https://doi.org/10.1093/SCAN/NSW154

Feldman Barrett, L., & Satpute, A. B. (2017). ARTICLE IN PRESS G Model Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience Letters*. https://doi.org/10.1016/j.neulet.2017.07.045

Finnemore, M., & Sikkink, K. (1998). International Norm Dynamics and Political Change. *International Organization*, 52(4), 887–917. https://doi.org/10.1162/002081898550789

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608F92.8CD550D1

Fölster, M., Hess, U., & Werheid, K. (2014). Facial age affects emotional expression decoding. *Frontiers in Psychology*, 5(FEB). https://doi.org/10.3389/FPSYG.2014.00030/FULL

Forgas, J. P. (Ed.). (2006). *Affect in social thinking and behavior*. Psychology Press.

Forster, M. (1993). Hegel's dialectical method. *The Cambridge Companion to Hegel*, 130–170. https://doi.org/10.1017/CCOL0521382742.006

Freitas, A. A. (2014). Comprehensible classification models. *ACM SIGKDD Explorations Newsletter*, 15(1), 1–10. https://doi.org/10.1145/2594473.2594475

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems* (TOIS), 14(3), 330–347. https://doi.org/10.1145/230538.230561

Frijda, N. H. (2017). The laws of emotion. *The Laws of Emotion*, 1–352. https://doi.org/10.4324/9781315086071/LAWS-EMOTION-NICO-FRIJDA

Frijda, N. H., & Scherer, K. R. (2009). Emotion definitions (psychological perspectives). *Oxford Companion to the Affective Sciences*, 186–187.

Gendron, M., & Barrett, L. F. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1(4), 316–339. https://doi.org/10.1177/1754073909338877

Gentzel, M. (2021). Biased Face Recognition Technology Used by Government: A Problem for Liberal Democracy. *Philosophy and Technology*, 34(4), 1639–1663. https://doi.org/10.1007/S13347-021-00478-Z/METRICS

Gert, Bernard (2004). *Common morality: deciding what to do*. New York: Oxford University Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton, Mifflin and Company.

Gilbert, P. (2009). Introducing compassion-focused therapy. *Advances in Psychiatric Treatment*, 15(3), 199–208. https://doi.org/10.1192/APT.BP.107.005264

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics*, DSAA 2018, 80–89. https://doi.org/10.1109/DSAA.2018.00018

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence*, 30(1), 35–79. https://doi.org/10.1016/0004-3702(86)90067-6

Girotto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional counterfactual thinking by actors and readers: Research article. *Psychological Science*, 18(6), 510–515. https://doi.org/10.1111/J.1467-9280.2007.01931.X

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.

Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making*, 9(1), 65–76. https://doi.org/10.1017/S193029750000499X

Goldie, Peter (2000). *The Emotions: A Philosophical Exploration*. Oxford, GB: Oxford University Press.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a 'right to explanation'. *AI Magazine*, 38(3), 50–57. https://doi.org/10.1609/AIMAG.V38I3.2741

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science (*New York, N.Y.), 293(5537), 2105–2108. https://doi.org/10.1126/SCIENCE.1062872

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517–523. https://doi.org/10.1016/S1364-6613(02)02011-9

Greeno, J. G. (1994). Gibson's Affordances. *Psychological Review*, 101(2), 336–342. https://doi.org/10.1037/0033-295X.101.2.336

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly: Management Information Systems*, 23(4), 497–530. https://doi.org/10.2307/249487

Grice, H. P. (1975). *Logic and Conversation*. In Speech Acts [Syntax and Semantics 3], Peter Cole and Jerry Morgan (eds), 41-58. New York: Academic Press.

Gross, J. J., & Barrett, L. F. (2011). Emotion Generation and Emotion Regulation: One or Two Depends on Your Point of View. *Emotion review : journal of the International Society for Research on Emotion*, 3(1), 8–16. https://doi.org/10.1177/1754073910380974

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). https://doi.org/10.1145/3236009

Gunkel, D. J. (2018). 24. Communication Technology and Perception. *Communication and Media Ethics*, 451–468. https://doi.org/10.1515/9783110466034-024/HTML

Gunkel, D. J. (2020). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, 22(4), 307–320. https://doi.org/10.1007/S10676-017-9428-2/METRICS

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

Gurski, P. (2020). Today in Terrorism | Bombing of world's largest Buddhist temple in Indonesia. https://borealisthreatandrisk.com/january-21-1985-bombing-of-worlds-largest-buddhist-temple-in-indonesia/

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, Vol. 2. The cognitive science of morality: Intuition and diversity (pp. 181–217). Boston Review.

Hall, L., & Clapton, W. (2021). Programming the machine: gender, race, sexuality, AI, and the construction of credibility and deceit at the border. *Internet Policy Review*, 10(4). https://doi.org/10.14763/2021.4.1601

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4), 843–887. https://doi.org/10.1093/BJPS/AXI147

Hamilton, I. A. (2018). Amazon Built AI to Hire People, but It Discriminated Against Women. https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10

Harari, Yuval Noah. (2015). *Sapiens: a Brief History of Humankind.* New York: Harper Perennial. Chicago Style. Harari, Yuval Noah

Harding, Jennifer. (2007). Evaluative stance and counterfactuals in language and literature. *Language and Literature* - LANG LIT. 16. 263-280. 10.1177/0963947007079109.

Hayes, B., & Shah, J. A. (2017). Improving Robot Controller Transparency Through Autonomous Policy Explanation. *ACM/IEEE International Conference on Human-Robot Interaction*, Part F127194, 303–312. https://doi.org/10.1145/2909824.3020233

HEBB, D. O. (1956). The distinction between classical and instrumental. *Canadian Journal of Psychology*, 10(3), 165–166. https://doi.org/10.1037/H0083677

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1), 19–29. https://doi.org/10.1007/S10676-008-9167-5/METRICS

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for "explainable AI. *Proceedings of the Human Factors and Ergonomics Society*, 1, 197–201. https://doi.org/10.1177/1541931218621047

Hollander, C. D., & Wu, A. S. (2011). The current state of normative agent-based systems. *JASSS, 14(2)*. https://doi.org/10.18564/JASSS.1750

Holmes, J. (2008). Mentalizing from a Psychoanalytic Perspective: What's New? *Handbook of Mentalization-Based Treatment*, 31–49. https://doi.org/10.1002/9780470712986.CH2

Hook, D. (2007). *Foucault, psychology and the analytics of power*. New York: Palgrave McMillan. https://doi.org/10.1057/9780230592322

Howard, A., Zhang, C., & Horvitz, E. (2017). Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. *Proceedings of IEEE Workshop on Advanced Robotics and Its Social Impacts, ARSO*. https://doi.org/10.1109/ARSO.2017.8025197

Hu, Xingming (2021). Hempel on Scientific Understanding. *Studies in History and Philosophy of Science Part A 88* (8):164-171.

Hume, D. (1978). *Treatise of human nature* (L. A. Selby-Bigge, Ed.; 2nd ed.). Oxford University Press.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1), 141–154. https://doi.org/10.1016/J.DSS.2010.12.003

Igou, E. R., van Tilburg, W. A. P., Kinsella, E. L., & Buckley, L. K. (2018). On the existential road from regret to heroism: Searching for meaning in life. *Frontiers in Psychology*, 9(DEC). https://doi.org/10.3389/FPSYG.2018.02375/FULL

Ivanhoe, Philip J. (2000). *Confucian Moral Self Cultivation*. Hackett Publishing Company.

Izard, C. E., & Haynes, O. M. (1988). On the form and universality of the contempt expression: A challenge to Ekman and Friesen's claim of discovery. *Motivation and Emotion*, 12(1), 1–16. https://doi.org/10.1007/BF00992469/METRICS

Jain, A. (2018). Book Review: Weapons of Math Destruction by Cathy O'Neill. SSRN *Electronic Journal*. https://doi.org/10.2139/SSRN.3187660

James, W. (1948). What is emotion? 1884. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 290–303). Appleton-Century-Crofts. https://doi.org/10.1037/11304-033

Jensen, R. (2006). Behaviorism, Latent Learning, and Cognitive Maps: Needed Revisions in Introductory Psychology Textbooks. *The Behavior Analyst*, 29(2), 187. https://doi.org/10.1007/BF03392130

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8(4), 195–204. https://doi.org/10.1007/S10676-006-9111-5/METRICS

Johnson, Mark (2014). *Morality for Humans: Ethical Understanding From the Perspective of Cognitive Science*. Chicago: University of Chicago Press.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., & Ghosh, J. (2019). Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. https://arxiv.org/abs/1907.09615v1

Jowett, B. (1936). *The Republic by Plato (370 BC)*: translated by Benjamin Jowett. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Miller, D. T. (1986). Norm Theory. Comparing Reality to Its Alternatives. *Psychological Review*, 93(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: a flaw in human judgment. First edition. New York, Little, Brown Spark.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Experiments in Environmental *Economics*, 1, 143–172. https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (2000). Choices, Values, and Frames. *American Psychologist*, 39(4), 341–350. https://doi.org/10.1037//0003-066X.39.4.341

Kahneman, Daniel ; Slovic, Paul & Tversky, Amos (eds.) (1982). J*udgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kaminski, M. E., Bertolini, A., Brennan-Marquez, K., Comandé, G., Cushing, M., Helberger, N., Van Drunen, M., Van Eijk, N., Eskens, S., Malgieri, G., Price, N., Sax, M., & Selbst, A. (2019). The Right to Explanation, Explained. Berkeley *Technology Law Journal*, 34, 189. https://doi.org/https://doi.org/10.15779/Z38TD9N83H

Kaptein, F., Broekens, J., Hindriks, K., & Neerincx, M. (2018). The role of emotion in self-explanations by cognitive agents. *7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, ACIIW 2017, 2018-January, 88–93. https://doi.org/10.1109/ACIIW.2017.8272595

Kasirzadeh, A., & Smart, A. (2021). The Use and Misuse of Counterfactuals in Ethical Machine Learning. FAccT 2021 - *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 228–236. https://doi.org/10.1145/3442188.3445886

Khalifa, K. (2012). Inaugurating Understanding or Repackaging Explanation? *Philosophy of Science*, 79(1), 15–37. https://doi.org/10.1086/663235

Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. J. *Artif. Int. Res. 70* (May 2021), 871–890. https://doi.org/10.1613/jair.1.12481

Kitcher, P., & Dennett, D. C. (1990). The Intentional Stance. *The Philosophical Review*, 99(1), 126. https://doi.org/10.2307/2185215

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a 'team player' in joint human-agent activity. *IEEE Intelligent Systems*, 19(6), 91–95. https://doi.org/10.1109/MIS.2004.74

Klingele, C. (2016). The Promises and Perils of Evidence-Based Corrections. *Notre Dame Law Review*, 91(2). https://scholarship.law.nd.edu/ndlr/vol91/iss2/2

Koch, E. J. (2014). How Does Anticipated Regret Influence Health and Safety Decisions? A Literature Review. *Basic and Applied Social Psychology*, 36(5), 397–412. https://doi.org/10.1080/01973533.2014.935379

Kolodner, J. L. (1993). *Case-based reasoning*. Morgan Kaufmann Publishers. http://www.sciencedirect.com:5070/book/9781558602373/case-based-reasoning

Küchle, G., & Ríos, D. (2008). *The Grammar of Society: The Nature and Dynamics of Social Norms*, by Cristina Bicchieri. Cambridge and New York: Cambridge University Press2006, xvi + 260 pp. Economics & Philosophy, 24(1), 117–123. https://doi.org/10.1017/S0

Kurzweil, R. (2012). *How to create a mind: the secret of human thought revealed*. New York, Viking. 266267108001727

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, 2017-December, 4067–4077. https://arxiv.org/abs/1703.06856v3

Lacave, C., & Diez, F. J. (2004). A review of explanation methods for heuristic expert systems. *Knowledge Engineering Review*, 19(2), 133–146. https://doi.org/10.1017/S0269888904000190

Lagioia, F., Rovatti, R., & Sartor, G. (2023). Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI and Society*, 38(2), 459–478. https://doi.org/10.1007/S00146-022-01441-Y/FIGURES/10

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: the effects of intentionality and foreseeability. *Cognition*, 108(3), 754–770. https://doi.org/10.1016/J.COGNITION.2008.06.009

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(2), 4762-4763. https://doi.org/10.1609/aaai.v31i2.19108

Lavanchy, M. (2018). Amazon's sexist hiring algorithm could still be better than a human. https://www.imd.org/research-knowledge/digital/articles/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human/

Lee, D. (2017). *Birth of intelligence: From RNA to artificial intelligence*. Oxford University Press. https://doi.org/10.1093/oso/9780190908324.001.0001

Leslie, A. M. (1987). Pretense and Representation: The Origins of 'Theory of Mind. *Psychological Review*, 94(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Lewis, M. D. (1995). Cognition-emotion feedback and the self-organization of developmental paths. *Human Development*, 38(2), 71–102. https://doi.org/10.1159/000278302

Leys, R. (2017). Paul Ekman's Neurocultural Theory of the Emotions. *The Ascent of Affect*, 76–128. https://doi.org/10.7208/CHICAGO/9780226488738.003.0003

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy 2021*, Vol. 23, Page 18, 23(1), 18. https://doi.org/10.3390/E23010018

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences*, 35(3), 121–143. https://doi.org/10.1017/S0140525X11000446

Lipton, Peter (2009). Understanding without explanation. In H. W. de Regt, S. Leonelli & K. Eigner (eds.), *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press. pp. 43-63.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *Communications of the ACM*, 61(10), 35–43. https://doi.org/10.1145/3233231

Lochmann, T., & Deneve, S. (2011). Neural processing as causal inference. Current Opinion *in Neurobiology*, 21(5), 774–781. https://doi.org/10.1016/J.CONB.2011.05.018

Lokman, Anitawati. (2010). Design & Emotion: the Kansei Engineering Methodology. *Design & Emotion: the Kansei Engineering Methodology*. 1. 1-11.

Loomes, G., & Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92(368), 805. https://doi.org/10.2307/2232669

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-December, 4766–4775. https://arxiv.org/abs/1705.07874v2

Maclure, J. (2021). Correction to: AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind (*Minds and Machines*, (2021), 31, 3, (421-438), 10.1007/s11023-021-09570-x).

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). Explainable Reinforcement Learning Through a Causal Lens. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2493–2500. https://doi.org/10.1609/aaai.v34i03.5631

Mahadevan, S. (2018). Imagination Machines: A New Challenge for Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 7988–7993. https://doi.org/10.1609/AAAI.V32I1.12214

Mahmoud, M. A., Ahmad, M. S., Mohd Yusoff, M. Z., & Mustapha, A. (2014). A review of norms and normative multiagent systems. *Scientific World Journal*, 2014. https://doi.org/10.1155/2014/684587

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, 25(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malone, J. (2022). Reflections and images: A place for art in medical physics? *Physica Medica*, 98, 63–79. https://doi.org/10.1016/J.EJMP.2022.04.004

Manna, Riya & Nath, Rajakishore (2021). Kantian Moral Agency and the Ethics of Artificial Intelligence. *Problemos 100*:139-151.

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. J*ournal of Experimental Social Psychology*, 29(1), 87–109. https://doi.org/10.1006/JESP.1993.1005

Markman, K. D., & McMullen, M. N. (2003). A reflection and evaluation model of comparative thinking. *Personality and Social Psychology Review*, 7(3), 244–267. https://doi.org/10.1207/S15327957PSPR0703_04

Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the Reflection and Evaluation Model. *Journal of Experimental Social Psychology*, 44(2), 421–428. https://doi.org/10.1016/J.JESP.2007.01.001

Markman, K. D., McMullen, M. N., Elizaga, R. A., & Mizoguchi, N. (2006). Counterfactual thinking and regulatory fit. *Judgment and Decision Making*, 1(2), 98–107. https://doi.org/10.1017/S193029750000231X

Markman, K. D., Mizoguchi, N., & McMullen, M. N. (2008). 'It would have been worse under Saddam:' Implications of counterfactual thinking for beliefs regarding the ethical treatment of prisoners of war. *Journal of Experimental Social Psychology*, 44(3), 650–654. https://doi.org/10.1016/J.JESP.2007.03.005

Marx, K., & Engels, F. (1967). *Capital: a critique of political economy.* New York, International Publishers.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. https://doi.org/10.1007/S10676-004-3422-1/METRICS

May, Joshua & Kumar, Victor (2018). Moral Reasoning and Emotion. In Karen Jones, Mark Timmons & Aaron Zimmerman (eds.), *Routledge Handbook on Moral Epistemology*. Routledge. pp. 139-156

Maybee (2020), *Hegel's dialectics,* The Stanford Encyclopedia of Philosophy. Encyclopaedia Britannica, s.v

McCarthy, J. and Hayes, P.J. (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence. In: Meltzer, B. and Michie, D., Eds., *Machine Intelligence, Vol. 4,* Edinburgh University Press, Edinburgh, 463-502.

McCloy, R., & Byrne, R. M. J. (2002). Semifactual "even if" thinking. *Thinking & Reasoning*, 8(1), 41–67. https://doi.org/10.1080/13546780143000125

McCormack, T., Feeney, A., & Beck, S. R. (2020). Regret and Decision-Making: A Developmental Perspective. *Current Directions in Psychological Science*, 29(4), 346–350. https://doi.org/10.1177/0963721420917688/ASSET/IMAGES/LARGE/10.1177_0963721420917688-FIG1.JPEG

McEleney, A., & Byrne, R. M. J. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking and Reasoning*, 12(2), 235–255.

Sperber, Dan & Mercier, Hugo (eds.) (2017). *The Enigma of Reason*. Cambridge, MA, USA: Harvard University Press.

Merlot, J. (2020). Coronavirus: Was der RKI-Katastrophenplan aus 2012 mit der echten Pandemie zu tun hat - DER SPIEGEL. https://www.spiegel.de/wissenschaft/medizin/coronavirus-was-der-rki-katastrophenplan-aus-2012-mit-der-echten-pandemie-zu-tun-hat-a-8d0820ca-95a7-469b-8a6a-074d940543d6

Metz, C. (2016). Google's AI Wins Pivotal Second Game in Match With Go Grandmaster | WIRED. https://www.wired.com/2016/03/googles-ai-wins-pivotal-game-two-match-go-grandmaster/

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 Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. ArXiv. /abs/1712.00547

Miracchi, Lisa (2020). Updating the Frame Problem for Artificial Intelligence Research. *Journal of Artificial Intelligence and Consciousness 7* (2):217-230.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data and Society, 3(2). https://doi.org/10.1177/2053951716679679

Mittelstadt, B. D., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *ACM Conference on Fairness, Accountability, and Transparency* (ACM FAT*), 279–288.

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 46. https://doi.org/10.1145/3387166

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Moore, A. W., & Nagel, T. (1986). The View From Nowhere. *The Philosophical Quarterly*, 37(148), 323. https://doi.org/10.2307/2220404

Morse, D. J. (2010). Dewey on The Emotions. *Human Affairs*, 20(3), 224–231. https://doi.org/10.2478/V10023-010-0023-Y

Moyer, C. (n.d.). How Google's AlphaGo Beat Lee Sedol, a Go World Champion - The Atlantic. 2016. Retrieved 10 July 2023, from https://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/

Muehlhauser, L. (2013). Transparency in Safety-Critical Systems - *Machine Intelligence Research Institute*. https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/

Muggleton, S. (2014). Alan turing and the development of artificial intelligence. *AI Communications*, 27(1). https://doi.org/10.3233/AIC-130579

Nagamachi, M. (2002). Kansei engineering as a powerful consumer-oriented technology for product development. *Applied Ergonomics, 33(3), 289–294*. https://doi.org/10.1016/S0003-6870(02)00019-4

Neerincx, M. A., van der Waa, J., Kaptein, F., & van Diggelen, J. (2018). Using perceptual and cognitive explanations for enhanced human-agent team performance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10906 LNAI, 204–214.* https://doi.org/10.1007/978-3-319-91122-9_18/FIGURES/5

Niiler. (2020). An AI Epidemiologist Sent the First Warnings of the Coronavirus | WIRED. https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/

Norman, D. A. (1999). Affordance, conventions, and design. *Interactions, 6(3), 38–43*. https://www.academia.edu/2849872/Affordance_conventions_and_design

Nussbaum M. C. (2001). *Upheavals of thought : the intelligence of emotions.* Cambridge University Press.

Nuyen, A. (2012). Confucian Role Ethics. *Comparative and Continental Philosophy 4* (1):141-150.

Nye, B. D., & Silverman, B. G. (2012). Affordance. *Encyclopedia of the Sciences of Learning,* 179–183. https://doi.org/10.1007/978-1-4419-1428-6_369

Oltramari, A., Francis, J., Henson, C., Ma, K., & Wickramarachchi, R. (2020). *Neuro-symbolic Architectures for Context Understanding*. http://arxiv.org/abs/2003.04707

Ortony, A. (2021). Are All "Basic Emotions" Emotions? A Problem for the (Basic) Emotions Construct. *Perspectives on Psychological Science,* 17(1), 41-61. https://doi.org/10.1177/1745691620985415

Overskeid, G. (2000). The slave of the passions: Experiencing problems and selecting solutions. *Review of General Psychology,* 4(3), 284–309. https://doi.org/10.1037/1089-2680.4.3.284

Overskeid, G. (2021). Can Damasio's Somatic Marker Hypothesis Explain More Than Its Originator Will Admit? *Frontiers in Psychology,* 11. https://doi.org/10.3389/FPSYG.2020.607310

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. https://doi.org/10.1007/S11023-019-09502-W/METRICS

Pariser E. (2011). *The filter bubble : what the internet is hiding from you.* Viking/Penguin Press.

Parry, W. T. (1973). Counterfactuals. *Journal of Symbolic Logic*, 44(2), 278–281. https://doi.org/10.2307/2273738

Pasquale, Frank author. (2015). *The black box society: the secret algorithms that control money and information.* Cambridge, Massachusetts ; London, England :Harvard University Press,

Pearl, Judea (2009). Causal inference in statistics. An overview. *Statistics Surveys* 3:96-146.

Judea Pearl. 2018. Theoretical Impediments to Machine Learning With Seven Sparks from theCausal Revolution. *In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 3. https://doi.org/10.1145/3159652.3176182

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. New York, Basic Books.

Picard, R.W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W. (1995). Affective Computing. M.I.T *Media Laboratory Perceptual Computing Section Technical Report No. 321*.

Piccardo, E. (2017). Plurilingualism as a catalyst for creativity in superdiverse societies: A systemic analysis. *Frontiers in Psychology*, 8(DEC), 288702. https://doi.org/10.3389/FPSYG.2017.02169/BIBTEX

Pieters, R., & Zeelenberg, M. (2007). A theory of regret regulation 1.1. *Journal of Consumer Psychology*, 17(1). https://doi.org/10.1207/s15327663jcp1701_6

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741), 233–238. https://doi.org/10.1038/22268

Plous, S. (2003). The psychology of prejudice, stereotyping, and discrimination: An overview. In S. Plous (Ed.), *Understanding prejudice and discrimination* (pp. 3–48). McGraw-Hill.

Pomerol, J. C., & Adam, F. (2008). Understanding human decision making - A fundamental step towards effective intelligent decision support. *Studies in Computational Intelligence*, 97, 3–40. https://doi.org/10.1007/978-3-540-76829-6_1

Preece, A.D., Harborne, D., Braines, D., Tomsett, R.J., & Chakraborty, S. (2018). Stakeholders in Explainable AI. ArXiv, abs/1810.00184.

Prigogine, I. & Stengers, Isabelle (1984). *Order Out of Chaos Man's New Dialogue with Nature* /Ilya Prigogine and Isabelle Stengers; Foreword by Alvin Toffler. Bantam Books, 1984.

Prinz, Jesse J. (2004). *Gut Reactions: A Perceptual Theory of the Emotions.* Oxford University Press.

ProPublica. (2016). Response to ProPublica: Demonstrating accuracy equity and predictive parity - equivant. https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

Rachlin, H. (2013). About Teleological Behaviorism. *The Behavior Analyst*, 36(2), 209. https://doi.org/10.1007/BF03392307

Re, R. M., & Solow-Niederman, A. (2019). Developing Artificially Intelligent Justice. I*n Stanford Technology Law Review* (Vol. 22, Issue 2).

Reb, J. (2008). Regret aversion and decision process quality: Effects of regret salience on decision process carefulness. *Organizational Behavior and Human Decision Processes,* 105(2), 169–182. https://doi.org/10.1016/J.OBHDP.2007.08.006

Rhim, J., Lee, J. H., Chen, M., & Lim, A. (2021). A Deeper Look at Autonomous Vehicle Ethics: An Integrative Ethical Decision-Making Framework to Explain Moral Pluralism. *Frontiers in Robotics and AI,* 8, 632394. https://doi.org/10.3389/FROBT.2021.632394/BIBTEX

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session,* 97–101. https://doi.org/10.18653/v1/n16-3020

Ribera, M. & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. *CEUR Workshop Proceedings,* 2327

Roese, N. J. (1994). The Functional Basis of Counterfactual Thinking. *Journal of Personality and Social Psychology*, 66(5), 805–818. https://doi.org/10.1037/0022-3514.66.5.805

Roese, N. J., & Epstude, K. (2017). The Functional Theory of Counterfactual Thinking: New Evidence, New Challenges, New Insights. *Advances in Experimental Social Psychology*, 56, 1–79. https://doi.org/10.1016/BS.AESP.2017.02.001

Roese, N. J., Hur, T., & Pennington, G. L. (1999). Counterfactual thinking and regulatory focus: Implications for action versus inaction and sufficiency versus necessity. *Journal of Personality and Social Psychology*, 77(6), 1109–1120. https://doi.org/10.1037/0022-3514.77.6.1109

Roseman, I. J., Wiest, C., & Swartz, T. S. (1994). Phenomenology, Behaviors, and Goals Differentiate Discrete Emotions. *Journal of Personality and Social Psychology*, 67(2), 206–221. https://doi.org/10.1037/0022-3514.67.2.206

Rosenberg, Alexander (1988). *Philosophy of social science*. Boulder, Colo.: Westview Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 2019* 1:5, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ryan, M. J. (1983). Decision theory and incomplete knowledge. *Metrika 1983 30:1*, *30(1)*, *108–108*. https://doi.org/10.1007/BF02056910

Sado, F., Loo, C. K., Liew, W. S., Kerzel, M., & Wermter, S. (2020). Explainable Goal-Driven Agents and Robots -- *A Comprehensive Review*. *ACM Computing Surveys*, 55(10). https://doi.org/10.1145/3564240

Saffrey, C., Summerville, A., & Roese, N. J. (2008). Praise for regret: People value regret above other negative emotions. *Motivation and Emotion*, 32(1), 46–54. https://doi.org/10.1007/s11031-008-9082-4

Salmon, Wesley C. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton University Press.

Santamaría, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *In Journal of Experimental Psychology: Learning Memory and Cognition* (Vol. 31, Issue 5). https://doi.org/10.1037/0278-7393.31.5.1149

Savarimuthu, B. T. R., & Purvis, M. (2007). Mechanisms for norm emergence in multiagent societies. *Proceedings of the International Conference on Autonomous Agents*, 1104–1106. https://doi.org/10.1145/1329125.1329335

Scheve, C.V., Moldt, D., Fix, J., & Lüde, R.V. (2006). My agents love to conform: Norms and emotion in the micro-macro link. *Computational & Mathematical Organization Theory*, *12*, 81-100.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4), 233–242. https://doi.org/10.1093/IDPL/IPX022

Seng, M., Lee, A., Floridi, · Luciano, & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics 2021* 1:4, 1(4), 529–544. https://doi.org/10.1007/S43681-021-00067-Y

Sequeira, P., & Gervasio, M. (2020). Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence, 288,* 103367. https://doi.org/10.1016/J.ARTINT.2020.103367

Shanahan, Murray (2008). *The frame problem.* Stanford Encyclopedia of Philosophy.

Shani, Y., & Zeelenberg, M. (2007). When and why do we want to know? How experienced regret promotes post-decision information search. *Journal of Behavioral Decision Making*, 20(3), 207–222. https://doi.org/10.1002/BDM.550

Shilton, K., Koepfler, J. A., & Fleischmann, K. R. (2014). How to see values in social computing: Methods for studying values dimensions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work,* CSCW, 426–435. https://doi.org/10.1145/2531602.2531625

Simon-Thomas, E. R., & Knight, R. T. (2005). Affective and cognitive modulation of performance monitoring: *Behavioral and ERP evidence. Cognitive, Affective and Behavioral Neuroscience*, 5(3), 362–372. https://doi.org/10.3758/CABN.5.3.362

Simonite. (2018). Google's AI Guru Wants Computers to Think More Like Brains | WIRED. https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/

Smallman, R., & Mcculloch, K. C. (2012). Learning from yesterday's mistakes to fix tomorrow's problems: When functional counterfactual thinking and psychological distance collide. *European Journal of Social Psychology*, 42(3), 383–390. https://doi.org/10.1002/EJSP.1858

Solomon, R. L. (1980). The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. *American Psychologist*, 35(8), 691–712. https://doi.org/10.1037/0003-066X.35.8.691

Spezialetti, M., Placidi, G., & Rossi, S. (2020). Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. *Frontiers in Robotics and AI*, 7, 532279. https://doi.org/10.3389/FROBT.2020.532279/BIBTEX

Spirtes, Peter ; Glymour, Clark ; N., Scheines & Richard, (1993). *Causation, Prediction, and Search*. Mit Press: Cambridge.

Staller, A., & Petta, P. (2001). Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. *Journal of Artificial Societies and Social Simulation.*

Stalnaker, Robert (1968). A Theory of Conditionals. In Nicholas Rescher (ed.), Studies in Logical Theory *(American Philosophical Quarterly Monographs 2*). Oxford: Blackwell. pp. 98-112.

Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking and Reasoning*, 24(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stephan, W. G., & Finlay, K. (1999). The role of empathy in improving intergroup relations. *Journal of Social Issues, 55(*4), 729–743. https://doi.org/10.1111/0022-4537.00144

Sterling, P., & Eyer, J. (1988). Allostasis: A new paradigm to explain arousal pathology. In S. Fisher & J. Reason (Eds.), *Handbook of life stress, cognition and health* (pp. 629–649). John Wiley & Sons.

Stern, C. D. (1981). Lewis' counterfactual analysis of causation. *Synthese*, *48*(3), 333–345. https://doi.org/10.1007/BF01063984/METRICS

Stolte, D. (2021). Study Shows Impacts of Deforestation and Forest Burning on Amazon Biodiversity | University of Arizona News. https://news.arizona.edu/story/study-shows-impacts-deforestation-and-forest-burning-amazon-biodiversity

Stoye, J. (2009). Minimax Regret. *The New Palgrave Dictionary of Economics*, 1–4. https://doi.org/10.1057/978-1-349-95121-5_2965-1

Sunstein, C.. (2021). Governing by Algorithm? No Noise and (Potentially) Less Bias. *SSRN Electronic Journal*. 10.2139/ssrn.3925240.

Swartout, W. R., & Moore, J. D. (1993). Explanation in Second Generation Expert Systems. *Second Generation Expert Systems*, 543–585. https://doi.org/10.1007/978-3-642-77927-5_24

Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable (Random House, 2007). *In The Review of Austrian Economics* (Vol. 21, Issue 4).

Tamò-Larrieux, A. (2021). Decision-making by machines: Is the 'Law of Everything' enough? *Computer Law and Security Review, 41*. https://doi.org/10.1016/j.clsr.2021.105541

Tegmark, M. (2017). *Life 3.0: being human in the age of artificial intelligence*. First edition. New York, Alfred A. Knopf.

Tiddi, I., D'Aquin, M., & Motta, E. (2015). An ontology design pattern to define explanations. *Proceedings of the 8th International Conference on Knowledge* Capture, K-CAP 2015. https://doi.org/10.1145/2815833.2815844

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in Machine Ethics. *ACM Computing Surveys* (CSUR), 53(6). https://doi.org/10.1145/3419633

Savarimuthu, B. T. R., & Cranefield, S. (2009). A categorization of simulation works on norms (pp. 39–58). *Presented at the Dagstuhl Seminar Proceedings 09121: Normative Multi-Agent Systems.*

Tosoni, L. (2021). The Right To Object to Automated Individual Decisions: Resolving the Ambiguity of Article 22(1) of the General Data Protection Regulation. *European Education*, 31(1), 96–103. https://doi.org/10.2753/EUE1056-4934310196

Turing, A. M. (1950). Computing Machinery and Intelligence. Source: *Mind,* New Series, 59(236), 433–460.

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. Source: *The Journal of Business*, 59(4), 251–278.

Tykocinski, O. E., & Steinberg, N. (2005). Coping with disappointing outcomes: Retroactive pessimism and motivated inhibition of counterfactuals. *Journal of Experimental Social Psychology*, 41(5), 551–558. https://doi.org/10.1016/J.JESP.2004.12.001

unesco. (2023). Building Partnerships to Mitigate Bias in AI | UNESCO. https://www.unesco.org/en/articles/building-partnerships-mitigate-bias-ai

Urdiales, C., Perez, E. J., Vázquez-Salceda, J., Sànchez-Marrè, M., & Sandoval, F. (2006). A purely reactive navigation scheme for dynamic environments using Case-Based Reasoning. *Autonomous Robots*, 21(1), 65–78. https://doi.org/10.1007/S10514-006-7231-8/METRICS

Vallor, Shannon (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York, NY: Oxford University Press USA.

Van De Poel, I. (2020). Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence. *Human Affairs*, 30(4), 499–511. https://doi.org/10.1515/HUMAFF-2020-0042

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Handbook of ethics, values, and technological design: Sources, theory, values and application domains. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 1–871. https://doi.org/10.1007/978-94-007-6970-0

Van Dyck, M. (2018). Kuhn's structures of scientific revolutions at fifty. Reflections on a science classic. *Centaurus, 60(1–2)*. https://doi.org/10.1111/1600-0498.12184

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, 25(3), 719–735. https://doi.org/10.1007/S11948-018-0030-8/METRICS

Varela, F. J., Thompson, E., Rosch, E., & Kabat-Zinn, J. (2016). The embodied mind: Cognitive science and human experience. *The Embodied Mind: Cognitive Science and Human Experience*, 1–322. https://doi.org/10.29173/CMPLCT8718

Verbeek, P.-P. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Ethics and Engineering Design*, 31(3), 361–380. http://www.jstor.org/stable/

Verweij, M., & Damasio, A.R. (2019). *The Somatic Marker Hypothesis and Political Life*. Oxford Research Encyclopedia of Politics.

Vollmer, N. (2023). Article 22 EU General Data Protection Regulation (EU-GDPR).

Wachter, S., Mittelstadt, B., & Floridi, L. (2016). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2903469

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6). https://doi.org/10.1126/SCIROBOTICS.AAN6080

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289

Waldrop, M. M. (1987). A Question of Responsibility. *AI Magazine*, 8(1), 28–28. https://doi.org/10.1609/AIMAG.V8I1.572

Weld, D. S., & Bansal, G. (2018). The Challenge of Crafting Intelligible Intelligence. *Communications of the ACM*, 62(6), 70–79. https://doi.org/10.1145/3282486

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, 72(3), 655–684. https://doi.org/10.1111/1467-8624.00304

White, D. (2017). Affect: An Introduction. *Cultural Anthropology*, 32(2), 175–180. https://doi.org/10.14506/CA32.2.01

White, D., & Katsuno, H. (2022). Artificial emotional intelligence beyond East and West. *Internet Policy Review*, 11(1). https://doi.org/10.14763/2022.1.1618

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Winograd, T. (1990). Thinking Machines: Can There Be? Are We? *Informatica* (Slovenia), 19.

Wittgenstein, L. (1963). *Tractatus logico-philosophicus. Logisch-philosophische Abhandlung.* [Frankfurt am Main: Suhrkamp. ISBN: 3518100122 9783518100127

Wong, P. H. (2019). Rituals and Machines: A Confucian Response to Technology-Driven Moral Deskilling. *Philosophies, 4(*4). https://doi.org/10.3390/PHILOSOPHIES4040059

Woodgate, J., & Ajmeri, N. (2022). Normative Ethics Principles for Responsible AI Systems: *Taxonomy and Future Directions.* 1. https://arxiv.org/abs/2208.12616v2

Woodward, J. F. (2003). Making things happen: a theory of causal explanation. *Philosophy and Phenomenological Research*, 74(1), 233–249. https://doi.org/10.1111/J.1933-1592.2007.00012.X

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11839 LNAI, 563–574. https://doi.org/10.1007/978-3-030-32236-6_51

Yang, C., Rangarajan, A., & Ranka, S. (2018). Global Model Interpretation via Recursive Partitioning. *Proceedings - 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems*, HPCC/SmartCity/DSS 2018, 1563–1570. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00256

Young, H. P. (2008). *Social Norms*. The New Palgrave Dictionary of Economics, 1–7. https://doi.org/10.1057/978-1-349-95121-5_2338-1

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science Technology and Human Values*, 41(1). https://doi.org/10.1177/0162243915605575

Zeelenberg, M. (2015). Robust Satisficing via Regret Minimization. J*ournal of Marketing Behavior*, 1(2), 157–166. https://doi.org/10.1561/107.00000010

Zeelenberg, M. (2018). *Anticipated regret: A prospective emotion about the future past*. In G. Oettingen, A. T. Sevincer, & P. Gollwitzer (Eds.)

Zeelenberg, M., Van Dijk, W. W., Manstead, A. S. R., & Van Der Pligt, J. (1998). The Experience of Regret and Disappointment. *Cognition & Emotion*, 12(2), 221–230. https://doi.org/10.1080/026999398379727

Zeelenberg, M., Van Dijk, W. W., Van Der Pligt, J., Manstead, A. S. R., Van Empelen, P., & Reinderman, D. (1998). Emotional Reactions to the Outcomes of Decisions: The Role of Counterfactual Thought in the Experience of Regret and Disappointment. *Organizational Behavior and Human Decision Processes*, 75(2), 117–141. https://doi.org/10.1006/OBHD.1998.2784

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy and Technology*, 32(4), 661–683. https://doi.org/10.1007/S13347-018-0330-6/METRICS

Zhu, Q. (2018). Engineering ethics education, ethical leadership, and Confucian ethics. *International Journal of Ethics Education*, 3(2), 169–179. https://doi.org/10.1007/S40889-018-0054-6

Zhu, Q., Williams, T., & Wen, R. (2019). Confucian Robot Ethics. *Computer Ethics - Philosophical Enquiry (CEPE) Proceedings, 2019*(1), 12. https://doi.org/10.25884/5qbh-m581

Zoshak, J., & Dew, K. (2021). Beyond kant and bentham: How ethical theories are being used in artificial moral agents. *Conference on Human Factors in Computing Systems* - Proceedings. https://doi.org/10.1145/3411764.3445102