|NST|TUTO
UN|VERS|TÁR|O
DE L|SBOA

AI MINDS MORALS 00111111

Beatriz de Almeida Pereira Nunes Raposo

Master in Social and Organizational Psychology

Supervisor:

Doctor Nelson Campos Ramalho, Associate Professor, Iscte – University Institute of Lisbon

July, 2021

CIÊNCIAS SOCIAIS
E HUMANAS

AI MINDS MORALS 00111111

Beatriz de Almeida Pereira Nunes Raposo

Master in Social and Organizational Psychology

Supervisor:

Doctor Nelson Campos Ramalho, Associate Professor, Iscte – University Institute of Lisbon

July, 2021

Acknowledgments

During my academic journey, writing a dissertation was a phase which I was quite reluctant on, since it is the final step of the journey and, even though I was anticipating it, I was also nervous. However, it was a pleasure to write this dissertation, which I am proud of and wrote with much delight.

I would like to thank my family, to whom I am eternally grateful for all the love, unconditional support and understanding. My parents, my brother and my grandparents are essential people in my life, who believe in me and drive me to do better every day. To one special person, that believed in me until his last day, I leave a special acknowledgment- thank you, Grandpa.

This dissertation would not be possible without the help and supervision of Professor Nelson Campos Ramalho, to whom I am truly grateful for all the support and help, from the begging to the end of this study. From theorical help with the constructs to the creation of the questionnaire and data analysis, he was the best supervisor I could ask for (and more). A special acknowledgment and thanks for all the help and for all the times I was doubting my work and he was not only a supervisor, but also a friend and supporter.

Lastly, I want to thank ISCTE-IUL for allowing me to have the opportunity of exploring a topic of my choice and for always providing me access all the literature concerning the topic. Furthermore, ISCTE allowed me to meet my colleagues and friends, to whom I also thank for the support throughout this journey. I also want to mention Beatriz Lebre, who could not physically participate in this academic ride but is, undoubtably, still present in the lives of the people that had the pleasure to be around her.

(this page purposively left blank)

Abstract

The growing frontiers of technology are also pushing the frontiers of social science as regards the challenges AI poses to assumptions about human being and human society. The increasing surrogate status of AI agents, namely those endowed with autonomous action, is of special interest in the intersection of psychology and ethics. Ascribing morality and mind may read as fictional, but a deeper understanding of these constructs may explain the already emerging literature on AI agency and morality. This literature is still scarce and has not treated both topics conjointly, which we believe would be expectable as morality and mind seem to go hand-in-hand.

With this objective, we designed a 2x2 experiment to test the effects that ascribed mind dimensions (agentic-experiential & metacognitive) crossed with ascribed morality to agents (authority & harm motivations) have on explaining the degree of blame depending on the nature of the agent (Human vs. AI). With a sample of 137 individuals, controlling for age, gender, wrongfulness, and displacement of responsibilities, we tested this three-way interaction to find fundamental differences between Human and AI agents. Humans were more blamable when they were seen as being less authority motivated (uncompliant) while having low metacognitive mind (less reflexive). Conversely, AI agents were more blamable when they were seen as being authority motivated while having low metacognitive mind. Findings are discussed at the light of the theory highlighting implications for theory, practice and the human-AI interaction.

Keywords: moral attributions, mind perception, AI, Human, agency

(this page purposively left blank)

Resumo

A extensão das fronteiras da tecnologia tem estimulado o crescimento das ciências sociais no que concerne os desafios que a IA coloca aos pressupostos sobre o Ser Humano e a sociedade. O crescente estatuto dos agentes de IA, nomeadamente daqueles dotados de ação autónoma, é de especial interesse na interseção da Psicologia e da Ética. Atribuir moralidade e mente pode parecer ficcional, contudo a compreensão aprofundada destes construtos pode explicar a emergência de literatura sobre IA e moralidade. Esta literatura é ainda escassa e não conjugou ainda os tópicos, o que acreditamos ser expectável dado a moralidade e a mente serem indissociáveis.

Com este objetivo, foi desenhada uma experiência 2x2 para testar os efeitos que a atribuição de mente (agêntica-experiencial & metacognitiva) cruzada com a atribuição de moralidade a agentes (motivações para autoridade & prejuízo) têm na explicação do nível de culpa dependendo na natureza do agente (Humano vs. IA). Com uma amostra de 137 indivíduos, controlando a idade, sexo, iniquidade/injustiça, e deslocação de responsabilidade, foi testada esta interação tripla, a fim de testar diferenças significativas entre agentes Humanos e de IA. Os agentes humanos foram mais culpados quando percecionados como sendo menos motivados para a autoridade (desobediente) e possuem menor mente metacognitiva (menos reflexivos). Contrariamente, agentes IA foram mais culpados quando percecionados como sendo motivados para a autoridade e menos mente metacognitiva. Os resultados são discutidos à luz das teorias centrando as implicações para a teoria, a prática e as interações humano-IA.

Palavras-chave: atribuições morais, perceção da mente, IA, Humano, agência

(this page purposively left blank)

Index

Table Index

Figure Index

Introduction

The moral attribution to Artificial Intelligence (AI) action is an uncharted territory that some would discard based on the non-sentient nature of synthetic agents. However, attributions are not intrinsic to the agent but instead to the eyes of the beholder, and thus, inanimate objects can be reified to represent anthropomorphic entities. Despite the recency of the topic there are already contributions that may offer an understanding about the process of moral attribution to AI agents.

In searching for moral attributions linked to the nature of the agent (synthetic versus human) Malle et al. (2019) found patterns of differential moral reasoning and that individuals were ready to attribute blame to AI agents when norms are broken. However, blame was not attributed to synthetic agents in the same manner it was for human agents. When obeying a hierarchical command, human agents blame was mitigated by displacement of responsibility, i.e., due to a sense of duty to follow orders or being obedient. A very small proportion (1:10) of individuals applied the same reasoning on synthetic agents and either mitigated blame for obeying or aggravated blame for failing to comply with the hierarchical command. A missing piece of this puzzle concerns the attribution of moral intention, i.e., why did agents act in the way they did (obeying / disobeying)? A tentative answer can be found on Graham et al. (2011) review on moral foundations. These authors state there are two contrasting moral motivations that stand out when judging morality in synthetic versus human agents: authority (obligations, obedience, and respect for superiors) and harm (caring, compassion for others). An additional missing point in Malle et al. (2019) reasoning concerns the epistemic status of the agents, i.e., how much are they autonomous thinkers (Schreck et al., 2019).

The perception of mind, whether concerning human or AI agents, involves two dimensions: agency and experience. While agent minds are defined by self-control, morals, memory, recognition of emotions, communication, thoughts and, specifically, intentions; experiential minds, on the other hand, are characterized by the capability to experience emotions, feelings and sensations (Gray et al., 2007). However, the agentic mind does not fulfil all the complex tasks and processes that a normal human mind can do, and, for this purpose, the metacognitive mind was formulated (Flavell, 1979). The metacognitive mind is a part of the metacognitive realm and translates the individual's deliberate organization of data into cognitive processes while the learning process or fulfillment of a task still occurs (Aktrk & Sahin, 2011). In this

study, the agentic and experiential dimensions of mind were fused into one singular dimension, named the agentic-experiential mind, and a second one concerning the metacognitive mind was created.

When facing a moral dilemma, especially when involving a life-or-death situation, individuals tend to show aversion towards an AI agent making a moral decision (Bigman & Gray, 2018). Different dimensions of mind perception define different attributions of morality (Gray et al., 2007). A high agency translates an intentional moral agent and a high experience translates a suffering moral patient (Gray et al., 2012) and, accordingly, the stronger the intention to create harm involved in a moral violation, the stronger the moral judgement about it (Gray et al., 2012). As regards the metacognitive mind, results show that this superior cognitive development might indicate a higher quantity of resources when considering and regulating the process of making an ethical decision (McMahon & Good, 2016).

Additionally, moral behaviour is influenced by our social guides of morality. A moral violation can be adapted so that the agent is not held accountable by switching a negative self-perception into a positive one, ignoring the consequences of the act. This is a part of a self-regulatory process and its activation is referred to as Moral Disengagement. It converts the immoral conduct into a rightful one, through many mechanisms. One of its mechanisms, Displacement of Responsibility, is the focus of this study because it enables the agent to perceive the act as mere obedience or dictation by authorities (Bandura, 2002).

Considering the topics above, the research topic of this study revolves around ascribed ethical responsibility of synthetic agents (AI-based algorithm) decisions based on the attribution of mind (Shank & DeSanti, 2018), including the processes and conditions related to experienced cognitions and emotions (e.g. Shank et al., 2019). The goal is to understand the different perspectives that participants might have on human or synthetic decision makers (Malle et al., 2019). When it comes to the topic of mind and morality, the aim is to understand if individuals are able to attribute a mind - and exactly what kind of mind - to an AI agent, and furthermore if the participants attribute morality when in case of a moral violation (e.g. Shank & DeSanti, 2018). The purpose is to offer an account on the different perspectives that individuals have on human or synthetic decision makers (Malle et al., 2019) as potential moral agents.

The thesis is organized so the reader can primarily understand the main issues in extant literature and the essential definitions on the topics of AI, mind perception, and attribution of morality. It is important to gain awareness about the different sort of agents and the existing typology of the mind, and how these relate to distinct levels of moral judgement. Considering this, a conceptual model is proposed that is followed by an explanation of the methodological

apparatus. After showing findings, the study proceeds with discussing them at the light of theory and concludes on the proposed thesis acknowledging its limitations as well as future research avenues.

1. Review of the Literature

1.1. Artificial intelligence

With the upcoming rise of technology today, AI became a notion that is increasingly referred and used. AI agents are becoming essential in modern society, from navigation, to advertising and even dating, they are important influencers on our interaction with others (Bigman & Gray, 2018).

The concept of AI was introduced by John McCarthy when studying the topic of artificial intelligence, which he defined as the use of machinery on human tasks, such as language, abstractions or concepts, which would be able to improve themselves through these actions. McCarthy explains AI as the transfer of human capacities, considered intelligent, to machines which are also capable of such capacities (McCarthy et al., 2006). With the rising of the AI topic, its applications also increase, transferring AI to more complex jobs and tasks that were typically entrusted only to humans (Shank & DeSanti, 2018).

Although both are agents, and therefore can intentionally act, humans and AI agents differ in their origin. Humans, and other animals, are considered natural agents because there are a product of biological evolution and reproduction, and they are biologically alive. AI agents are artificial agents, since their origin comes from an intentional manufacturer that built these agents from scratch with external materials. However, it is essential to understand these concepts are not mutually exclusive and can therefore be both present and absent in an agent. An agent can be both natural and artificial- as for example a type of clone, since it is biologically and artificially alive, or neither of them - like a God - which is by definition an entity that transcends biological or artificial realms (Himma, 2009).

1.2. Mind perception

The presence of mind is often attributed to the capacity to feel and experience emotions consciously, while the absence of mind, typically attributed to objects or nonhuman agents, is associated with the lack of these. Given this, the perception of mind is essential to comprehend other social phenomena, such as anthropomorphism, dehumanization or moral disengagement (Gray et al., 2012). Mind perception enables social interactions given the fact that perceiving a mind in others is broader than just perceiving people and narrower than making inferences about other's minds. Mind perception also works as a mediator in the relationship between sensory input and the action that follows it (Epley & Waytz, 2010).

When speaking of perception (or attribution when referring to the mind), it is essential to understand the Theory of Mind (ToM). The ToM acknowledges mental states and outlines the mental mechanisms behind the construction of mind attributions and characterizes the understanding of intentions and motivations of other people. According to ToM, we trust the intentions and beliefs of the people whom we socially interact with, albeit they can differ from our own (Schreck et al., 2019). The inference of mental states in ourselves and others can be useful to make predictions and to understand actions and behaviors (Premack & Woodruff, 1978) irrespectively of the nature of the agent; human or not (Epley & Waytz, 2010).

The perception of mind in AI also depends on the physical appearance of the agent. Seeing the face or head of a robot directly enhances the human characteristics more whereas seeing the same robot from behind, where its wires and cables are visible, enhances the mechanical and technology in said AI agent (Gray & Wegner, 2012). The presence of a face in a robot makes it more human like, which enables perceptions of its mind as being more experiential, compared to a faceless robot (Broadbent et al., 2013).

According to Gray et al. (2012), mind perception can be translated in the acknowledgment of the existence of a mind in others. The perception of mind consists in two dimensions, that people intuitively differentiate: the agent mind, meaning self-control, morals, memory, emotion recognition, communication, thoughts and intentions; and the experiential mind, translated in the ability to experience feelings, emotions or sensations, such as pain, hunger or embarrassment (Gray et al., 2007). For instance, adults are perceived as having both high agentic and experiential mind (Gray et al., 2012), babies are perceived as having low agentic and high experiential mind (Epley & Waytz, 2010), while corporations are perceived as having low experiential but high agentic mind (Gray et al., 2012). When it comes to AI agents they are usually perceived as owning a low experiential but moderate agentic mind. The agentic mind in AI agents is perceived as moderate since they are, typically, conceived as having less agentic mind than adult humans (Gray et al., 2007). Because agency includes abstract capabilities, as for example communication and self-control (Gray et al., 2007), AI agents may be perceived as less capable to make moral decisions than humans are, and the aversion towards an AI agent in this context can be associated with a low perception of mind in this type of agent (Bigman & Gray, 2018).

Even though more complex, the agentic mind does not fulfill all the capabilities that the mind can have or do (Bogdan, 2001), since it only refers to the cognition and the awareness, understanding and fulfillment of a task (Akturk & Sahin, 2011). However, to understand the task completely, one needs to identify the required capabilities to perform it, and to be aware

and know how to learn, in addition to what one has already learned and understood about the task itself (Akturk & Sahin, 2011). While cognition and cognitive strategies help the individual to reach a certain goal, metacognition and metacognitive strategies are used to make sure that said individual reached the goal and, usually, pave the path for cognitive strategies or follow right after them (Livingston, 1997). The meta dimension is still under intense investigation and its definition is still under a stringent analysis, however one common point of all authors is that metacognition translates in the individual's self-evaluation and deliberate organization of information into cognitive processes while the learning process or fulfillment of a task still occurs (Akturk & Sahin, 2011).

The metarepresentation of mental states and the metamentation - or thinking about thoughts - finds its origin in the interpretation of intentions and aboutness that comes from a generic sense of reference, built on the early stages of the human development, which enables the appearance of unique cognitive skills and suggests a gradual development of the metamind on humans (Bogdan, 2001). Metamentation is thus a part of the metacognitive dimension of the mind, which is a capability developed due to the interpretation process and translates in the construction of mental representations about other's representations and thoughts about one's thoughts.

According to Bogdan (2001) a form of the metacognitive mind is metamental prediction, which corresponds to the imagination of situations populated with people that think about other people's thoughts, and another is recursive metathinking, which is the embedment of one thought in other thoughts, which were already embedded in other thoughts in the first place, and so on. Accordingly, it makes sense that the metacognitive mind includes metacognitive strategies, such as the evaluation, planning and deliberation of a thought, and none of these factors would function if one was not able to comment, through thought, about a topic related to another thought. Additionally, this author states it would not be possible to metamentate without the presence of three capabilities. The first is metarepresentation, referring to the understanding that a mental representation can either be a true or false target and can be diverse in terms of form of representation, whether that target one person at multiple times or two different people. The second capability translates in the recognition and tracking of iterated embeddings of the mental representations about other mental representations. Finally, the third capability concerns the organization of explicit metathoughts into mental structures that represent the specified content related to the terms of other thoughts (Bogdan, 2001).

The metacognitive strategies, that are more complex and advanced thought processes (Cornford, 2002), translate the high functioning of the metacognitive mind. Abstraction is a

fundamental characteristic of this type of mind and losing it might affect how an action is associated with the agent in question and how we interpret its consequences (Floridi & Sanders, 2004). As mentioned, the metacognitive domain is more complex than the ordinary cognitive domain because it involves the intention to learn and to manipulate the information, and self-control and regulation (Weinstein et al., 2000). When applying the metacognitive strategies in the learning context, the self-regulatory ability of the agent's mind takes an important role in metalearning, which is the process of learning to change or improve the skills in question (Cornford, 2002). Concerning perceptions, the metacognitive dimension refers to the prediction of the judgement made by others about ourselves, i.e., a metaperception, which requires making a judgement about another's judgement and self-observation. When evaluating an act, the involved agent tends to justify it with situational variables, whereas the observer judges more accurately without bias (Albright & Malloy, 1999).

According to Bigman and Gray (2018) the aversion towards an AI agent making a moral decision is mediated by mind perception, even if the decision and the outcome is the same, meaning the aversion occurs even if the outcome is positive. Concerning ethics of decision making, people prefer a human agent over an AI agent, especially in contexts of life and death situations involving driving, law, medicine, and the military (Bigman & Gray, 2018).

1.3. Attribution of morality

Extant literature reveals that the dimensions of mind perception capture distinct judgements of morality (Gray et al., 2007). The perceptions of mind are associated with morality and moral judgements, since we attribute intention to the actions, which consequently, determines the valuation of its consequences (Gray et. al., 2012). Given this, and the fact that we perceive AI as having mind, the question about AI's moral behavior and whether we should treat or not AI as moral agents (Broadbent, 2017). However, Shank and DeSanti (2018) observed that perceiving more mind in an AI agent is strongly associated with attributions of morality and intentionality. When a moral violation occurs, we tend to evaluate morality according to an existing cognitive template, influenced by our personal past experiences (Gray et al., 2012). The moral dyad, i.e., the cognitive template for morality, helps individuals to understand morality, its violations and its implications. The dyad suggests that moral blame is produced by the sum of intention plus suffering, which implies that more blame is attributed when there is a direct relationship between the agent and the patient where the agent intentionally acts and the patient suffers as a result from that action. Because mind perception is so flexible, this dyad can

be applicable to most moral situations and outcomes, to the extent that there is no need for an obvious victim and transgressor when the observer perceives this dyad (Gray et al., 2012).

By applying the moral dyad to moral judgement, we assume that in the process of evaluating an act we identify both an intentional moral agent, characterized by a high agentic mind, and a suffering moral patient, characterized by an experiential mind (Gray et al., 2012). Because metacognition involves components of agency (Chambon et al., 2014), it is possible to theorize that a owning a metacognitive mind will increase the ascribed agency. In this sense, moral agents have moral obligations and moral patients are owned a moral obligation. An AI agent can be an agent if it has the consciousness so that it can be held accountable, and is it owns the ability to choose to act freely upon the norms of morality (Himma, 2009). And, when in contact with a moral violation there is an underlying tendency to moral typecast. This occurs whenever we identify two entities - a moral agent and a moral patient – that are taken as mutually exclusive.

The connection between mind perception dimensions and the attribution of morality is strongly connected to the principle that responsibility is associated with a high level of mind, since it is related to the existence of morality in the agent (Gray et al., 2012). The perceived mind affects how we judge and evaluate an act and how we treat the minds that are involved (Yam et al., 2020). Accordingly, these authors claimed that when perceived as owning high agentic mind, the entity is seen as capable of intentional and voluntary decision making and acting upon it, making it more autonomous in the decision-making process. Consequently, the perceived high agentic mind is then seen as more responsible for the act and qualified as a moral agent (Schein & Gray, 2018). Being an autonomous moral agent decision maker intensifies its ascribed blamefulness (Yam et al., 2020). On the other hand, the experiential mind qualifies an entity as a moral patient, and thus enables it to experience the consequences of the act. This sentient nature of the patient tends to capture empathy from the observer, which influences how he or she interprets harm directed towards the patient, tending to more strongly condemn the agent (Schein & Gray, 2018).

The defining variables of the moral domain are still under intense investigation. The definition of moral systems is built around the intertwining of various sets of “values, virtues, norms, practices, identities, institutions, technologies and psychological mechanisms” (Haidt & Kesebir, 2010, pg 800). The relation between these factors enables the regulation and suppression of selfishness, that consequently, makes living in a social world possible. According to Haidt and Graham’s (2007) work on The Moral Foundations Theory, morality revolves around five foundations: Harm/Care, Fairness/Reciprocity, Ingroup/Loyalty,

Authority/Respect, and Purity/Sanctity. Harm/Care and Fairness/Reciprocity correspond to the ethics of autonomy, Ingroup/Loyalty and Authority/Respect deal with community ethics, and Purity/Sanctity revolves around the ethics of divinity. Even though the moral realm is associated with cultural ethical codes and values that broaden its definition beyond issues of harm or fairness, the concept of morality is often referred to as how well or poorly do we take care of individuals and it is represented by rules or codes that protects from harm (Graham et al., 2011).

Accordingly, harmful actions are relevant in mind perception and moral judgements, connecting both with the harm-made mind phenomena (Ward et al., 2013). This effect occurs when a moral agent acts harmfully and intentionally towards a second presence, which inevitably makes this second entity a moral patient. The act is then interpreted according to the moral dyad and, consequently, the moral patient, that suffers the consequences, is perceived has owning more mind (Ward et al., 2013). However, this effect can have opposite results when the victim is perceived as owning high levels of mind in the first place. The perceived mental capacities of the moral patient dictate how the victimization process occurs, which can unfold in two ways. When the patients are perceived has having high levels of mind, the observer dehumanizes and stripes their mental capacities. On the other hand, when there is a low perception of mental status in the patients, the observer constructs a mind for them as a result of the victimization (Küster & Swiderska, 2020).

In sum, the moral domain is observed in the moral dyad when there is: a) a perception of a moral agent, b) that intentionally acts with the required agentic (Gray et al., 2012), and/or metacognitive capacities (Chambon et al., 2014), and c) a moral patient, d) that suffers and feels the interpersonal harm due to having experiential capacity (Gray et al., 2012). Accordingly, a moral violation requires two entities so that acts can be morally relevant. The intention that motivates the act combined with the painful consequences dictate the moral judgement, and the stronger the intention and the harm, the most probable the said act is judged as immoral. This process is not straightforward because moral judgments are influenced by our personal moral foundations (Haidt & Graham, 2007), by our interpretations of the perceived mind (Yam et al., 2020), and by our personal social experiences and views (Graham et al., 2011).

1.4. Moral disengagement

In defining the moral domain, one needs to consider the social interactions between societies and cultures with distinct historical backgrounds (Graham et al., 2011). Because we live in a society and in social realities, we do not function as independent moral agents, and our

moral actions are influenced by cognitive, affective and social dimensions. We build social moral guides that function as orientation in our moral behavior, and it is in this self-regulatory process that we consider what is a moral and immoral action (Bandura, 2002). The perceived circumstances associate with our personal moral guides operate as a regulator of the perceived consequences that we attribute to our own actions. People tend to reject behavior that goes against their moral guides because the consequences involve self-sanctions and negative self-perceptions. This thought explains why there is a tendency for people to constraint negative self-condemnations when they behave inhumanely and switch the negative self-condemnations into positive ones (Bandura, 2002).

The consequences of a moral violation, and consequently the victim, can be adapted so that the agent does not feel the full blame for their act. This adaptation can be referred to as Moral Disengagement (Bandura, 1999), which is the rationalization of the act and its impact on the victim since the consequences are only considered immoral or harmful if the victim gets hurt (Gray et al., 2012). Another form of guilt and blame adaption is Dehumanization, where the agent denies the victim's mental states and uses that to justify the immoral act (Bandura et al., 1996). The association of mind perception and morality is supported by dehumanization of the victim (Gray et al., 2012).

Disengagement practices are often present in our daily behavior when there is a need to justify a certain act or to make a cognitive reconstruction of a culpable act into a righteous one. These moral phenomena of Dehumanization or Moral Disengagement occur when there is a selective activation of internal self-regulatory controls. According to Bandura (2002), the self-regulatory mechanisms behind moral standards function if only activated and the activation depends on the individuals, making it possible to have different kinds of conducts with the same moral standard. Because the activation is self-controlled, the process of moral control and self-censure has various phases when it can be separated from the reprehensible conduct. The moral controls can disengage from detrimental conduct, and consequently, self-sanctions are disengaged by the reconstruction of the conduct, omitting the personal self-agency in the origins of the action. This disregard towards the consequences of the act creates a negative perspective of the victim which makes the harmful act more justifiable (Bandura et al., 1996). Disengagement can focus on a) the redefinition of the harmful conduct into a rightful one through moral justification, b) agency in the action which enables the diffusion and displacement of the responsibility, c) minimization or distortion of the harmful consequences of the act, or d) disengagement of the act by blaming or dehumanizing the victim involved (Bandura, 2002). Following this, one can conclude that self-awareness and acknowledgment of

one's own contribution to the harmful act intensifies the operation of the moral control (Bandura, 1990).

The mechanism of Displacement of Responsibility allows the moral agent to view their violation as an obedient act authorized and dictated by superiors or authorities, and consequently minimizing the perceived blame. The moral agent is not considered responsible for their act and is spared from condemnation, according to their own self-evaluation (Bandura, 2002).

Overall, the human being is intrinsically moral in the sense that it is provided with automatic cognitive and affective processes that are activated whenever one perceives themselves as acting immorally. These processes, referred as moral disengagement in literature, can be activated and follow a variety of paths that, ultimately, exempt the agent from being morally blamed, and likewise, from perceiving themselves as being immoral. The preconditions to ascribe blame concern having a moral intention and a mind of its own. It seems reasonable to infer that moral attributions per se do not suffice to morally judge any agent, since blame is social and cognitive, and relies on social regulation, social cognition and warranty (Malle et al., 2014). Accordingly, the attributed degree of blame will require also how much one would agree with the act (Malle et al., 2014) and if the agent can be thought of as a sentient entity, i.e. one that has a mind of its own and acts intentionally (Alicke, 2000). These dimensions (experiencing, agency, metacognition) are not exclusive of human beings, and therefore, other entities, no matter if natural or synthetic, can be morally judged. Therefore, we hypothesize that:

H1: Moral attributions will not relate to blame

H1a: Authority moral attribution will not relate to blame

H1b: Harm moral attribution will not relate with blame.

It has been proven that our perceptions of mind in an agent influence how we attribute morality (Yam et al., 2020). Therefore, it makes sense to theorize that a higher level of mind is associated with the perception of more intentionality and autonomy to act (Gray et al., 2012; Himma, 2009), making this kind of mind more prone to be blamable (Schein & Gray, 2018; Yam et al., 2020).

H2: There is an interaction effect between morality attribution and mind perception such that the higher the perceived mind, the stronger the positive association between morality attribution and blame.

H2a: There is an interaction effect between morality attribution and *agentic-experiential mind* attribution such that the higher the agentic-experiential mind, the *stronger the positive association* between morality attribution and blame.

H2b: There is an interaction effect between morality attribution and *metacognitive mind attribution* such that the higher the metacognitive mind, the *stronger the positive association* between morality attribution and blame.

Considering the aversion towards an AI decision maker (Bigman & Gray, 2018), we hypothesize that this aversion might have an influence on the degree of blame that is attributed. Because human agents tend to be perceived as owning a higher level of mind than AI agents (Gray et al., 2012, Gray et al., 2007), we theorized that the nature of the agent will emphasize mind perception in the association between moral attributions and blame.

H3: The agent nature will moderate the conditional effect of mind attribution in the relationship between moral attribution and blame such that human agents with higher attributed mind will be showing the strongest relationship between morality attribution and blame.

H3a: The agent nature will moderate the conditional effect of *agentic-experiential mind attribution* in the relationship between *harm moral attribution* and blame such that human agents with higher *attributed agentic-experiential mind* will be showing the strongest relationship between *harm morality attribution* and blame.

H3b: The agent nature will moderate the conditional effect of *agentic-experiential mind attribution* in the relationship between *authority moral attribution* and blame such that human agents with higher *attributed agentic-experiential mind* will be showing the strongest relationship between *authority morality attribution* and blame.

H3c: The agent nature will moderate the conditional effect of *metacognitive mind attribution* in the relationship between *harm moral attribution* and blame such that human agents with higher *attributed metacognitive mind* will be showing the strongest relationship between *harm morality attribution* and blame.

H3d: The agent nature will moderate the conditional effect of *metacognitive mind attribution* in the relationship between *authority moral attribution* and blame such that human agents with higher *attributed metacognitive mind* will be showing the strongest relationship between *authority morality attribution* and blame.

By integrating the hypotheses, the following conceptual model is devised:

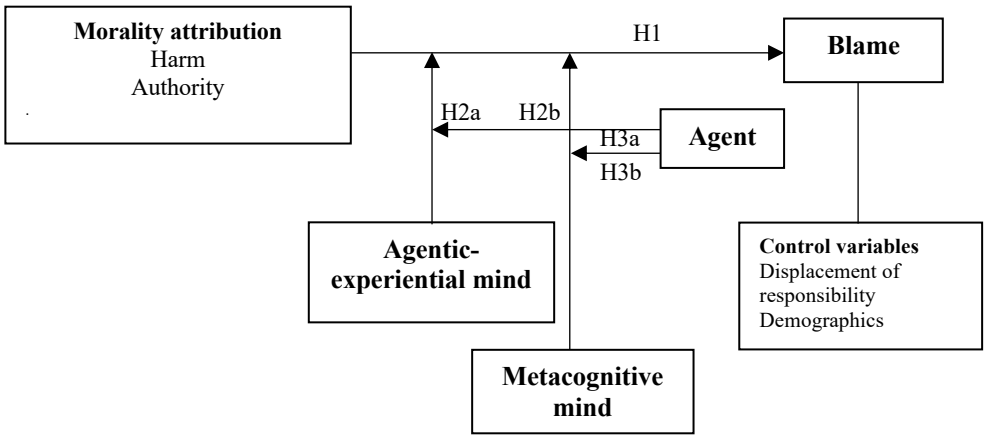

Figure 1- Conceptual model

2. Method

2.1. Procedure

The data was collected with an online questionnaire composed by three surveys, which were answered anonymously, confidentially and voluntarily by the participants. The collection of demographic data is reserved for the last section of the survey, after the completion of the third part. The link to the questionnaire was shared via http://isctecis.co1.qualtrics.com/jfe/form/SV_23OIJMOZSr4CBYF. In total 141 participants answered the questionnaire, from which 137 had valid full entries to use in the study.

2.2. Data analysis strategy

Data analysis followed a two-step approach where firstly we tested for the psychometric quality of the measures so to ensure these are sufficiently valid and reliable to be used for hypothesis testing. Validity issues were screened via exploratory factor analysis where solutions were considered good enough if KMO reached at least .500 and the Bartlett test of sphericity rejected the null hypothesis, adding to items showing communalities of .500 minimum. Additionally, factors were extracted with Kaiser criterion (eigenvalue above 1) and solutions rotated (Varimax). We set the minimum threshold for an acceptable explanative power at 60% of total variance after rotation. Convergent validity was assessed with Average Extracted Variance (AVE) which should attain at least .500 (Fornell & Larcker, 1981) and discriminant validity was assessed with HTMT which should be fall below .85 (Henseler et al., 2015). Reliability was assessed with Cronbach alpha (with the acceptance threshold set to .70). As per the hypotheses testing, considering the model depicts a moderated moderation with a simple predictor-criterion path, we opted to conduct the analyses with Process Macro (Hayes, 2018) namely with model 3 setting the bootstrapping parameters for 5000 repetitions with a confidence interval of 95%.

2.3. Sample

Individuals 18 or above, residing in Portugal, and having diverse situations were eligible for this study. Considering the nature of the research objective, it is advisable to count on a diverse sample, both concerning gender, age, education, and contact with information technologies. Thus, sample is comprised of 137 participants, being mostly females (69.3%) and

with ages ranging from 19 to 68, averaging 33.9 years-old (SD=14.6) and single (62.6%). Most participants reported having a bachelor's degree (63.6%).

2.4. Measures

Agency type was measured based on military scenarios from Malle et al. (2019). The scenario and the questions about on it set within a military scenario a moral dilemma that requires a resolution. In this military scenario the human agent is compared to an AI agent and both decision makers have the same command structure and chain of action. The restraint for both agents, human and AI, is the superior's permission and the obligation to serve the military and humanitarian ethics to minimize threats and civilian losses (ICRC, 2018)

Participants were exposed to the story of a military scenario (see appendix A), one paragraph at a time, having to click on a button to continue reading. The scenarios that were used in this study were translated to European Portuguese through translation retroversion procedure (Brislin, 1986).

After reading the story, with the experimental manipulated decision, the participants were inquired about the moral judgements. Firstly, it was asked what the agent should do (launch strike/cancel strike), and secondly, if the agent decision was morally wrong (Yes/No), making it possible to understand if there is any kind of moral judgement behind the decision that was made. Thirdly, it was questioned how much blame the agent deserves for their decision, which enables the comprehension of the attribution of a moral consequence for the decision maker. The examination of the dilemma pushes individuals into decision whether to attribute or not blame and responsibility to the agent, be it human or synthetic.

Mind Perception was measured using Shank and DeSanti (2018) adapted version of the attribution of mind scale by Wyatz et al. (2010) (see appendix B). This instrument evaluates the attributed level of mind expressed via attributed intentionality, free will, consciousness, desires, beliefs, and ability to experience emotions. Based on literature that theorizes on the metacognitive mind (Bogdan, 2001) we developed a five-item scale translating the capacity to organize and memorize sensorial information (Bogdan, 2001), to detect other people's emotions (Bogdan, 2001), to change own learning and decision processes (Akturk & Sahin, 2011; Weinstein et al., 2000), to observe one's own behaviour (Akturk & Sahin, 2011), and to organize reality into abstract categories (Floridi & Sanders, 2004). The scale's response option ranged from 1 (Not at all) to 5 (Completely). Like the scenarios, the scale was translated to European Portuguese through translation retroversion procedure (Brislin, 1986).

The exploratory principal component analysis extracted a valid two-factor solution (KMO=.891, .827<MSA<.946, Bartlett $X^2(66)$ =1275.834, p<.001) but one of the new items ("… is able to organize reality into abstract categories") failed to achieve the minimum commonality threshold of .500 and was thus removed. The resulting two-factor solution is also valid (KMO=.887, .799<MSA<.945, Bartlett $X^2(55)$ =1229.937, p<.001) and accounts for 71.8% of total variance after rotation (Varimax) where the first factor "agentic-experiential mind" comprehends all seven items of Shank and DeSanti (2018) and has extremely high reliability (Cronbach alpha=.943), and the second factor "Metacognitive mind" comprehends four items (see table 1) and has good reliability (Cronbach alpha = .819). Both components have convergent validity ($AVE_{ag\text{-}exp}$=.693; $AVE_{metacognitive}$=.536) and the structure has discriminant validity as shown by HTMT value of .687 (Henseler et al., 2015).

Table 1- *Rotated matrix for mind attribution scale*

The pilot …	Component	
	1	2
... has beliefs	**.901**	.226
... has desires	**.883**	.266
... has the ability to experience emotions	**.834**	.372
... has conscientiousness	**.831**	.351
... has intentions	**.809**	.183
... has free will	**.761**	.186
... has a mind of its own	**.667**	.393
... is capable of altering their own learning and decision-making processes.	.141	**.838**
... can observe their own behaviour	.294	**.804**
... is capable of detecting emotions on others	.387	**.709**
... is capable of organizing detected sensory stimulus into memory	.214	**.696**
Cronbach alpha	.943	.819

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

The measure is thus considered valid and reliable with the two proposed factors that add to agentic-experiential mind (Shank & DeSanti, 2018) a novel 4-item scale intended to measure the metacognitve mind.

Moral judgment was measured based on The Moral Foundation Questionnaire (MFQ), namely, the subscales of Harm and Authority (Graham et al., 2011). After answering the questions about the scenario, participants were requested to signal how much they think each

of the composing dimensions of Harm and Authority were relevant in the story they just read, and how much they agree the specific decision (strike / no strike) was moral, judging on those precise same dimensions. The response option in the Moral Relevance measure ranged from 1 (Not at All Relevant) to 5 (Extremely Relevant), and in the Moral Judgement measure varied from 1 (Strongly Disagree) to 6 (Strongly Agree). The MFQ was already available in European Portuguese, with a validation by Silvino et al. (2016).

For this study, only the MFQ dimensions of Harm and Authority were evaluated with some minor language adaptations (see appendix C), since these were essential to understand the topic of interest of this study. Harm is especially important in moral judgements which revolves around the interaction between two individuals, the Moral Dyad: the moral intentional agent and the moral suffering patient. According to the Moral Dyad, harm separates moral violations from conventional acts and connects with negative affect when in context of a moral violation (Schein & Gray, 2018). In this questionnaire, Harm also includes not only suffering, but care and compassion for others. Authority refers to obligations, obedience and respect for superiors (Silvino et al., 2016). It is evaluated mainly due to the context of the scenarios presented to the participants, since it involves the military and the authorization of the agent's superior to make the decision.

An exploratory factor analysis showed a valid (KMO=.658, Bartlett's X^2=368.427, 66, p<.001) four factor solution extracted with Kaiser criterion (eigenvalues over 1) explaining 63% of total variance after rotation (Varimax). All items fell in the respective factor (as depicted in Table 2). The factors extracted were thus: 1) harm_relevance (3 items, e.g. "Whether or not someone suffered emotionally", Alpha=.806), 2) authority_relevance (3 items, e.g. "Whether or not someone showed a lack of respect for authority", Alpha=.680), 1) harm_judgment (3 items, e.g., "Compassion for those who are suffering is the most crucial virtue", Alpha=.410), 2) authority_judgment (3 items, e.g., "Respect for authority is something all children need to learn", Alpha=.680).

Table 2 – *Factor matrix for MFQ items*

	Component			
	1	2	3	4
HarmRel2 Whether or not someone cared for someone weak or vulnerable	**.897**	-.019	-.021	-.036
HarmRel1 Whether or not someone suffered emotionally	**.779**	.311	-.035	.028
HarmRel3 Whether or not the decision was cruel	**.733**	.326	.015	.218
AuthorityRel1 Whether or not someone showed a lack of respect for authority	.053	**.824**	.085	-.114
AuthorityRel2 Whether or not an action caused chaos or conflict	.279	**.731**	-.048	-.005
AuthorityRel3 Whether or not someone conformed to the traditions of society	.148	**.673**	-.029	.124
AuthorityJud2 Men and women each have different roles to play in society.	-.038	-.041	**.848**	.130
AuthorityJud1 Respect for authority is something all children need to learn.	.132	-.168	**.835**	.160
AuthorityJud3 If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.	-.152	.280	**.646**	-.181
HarmJud1 Compassion for those who are suffering is the most crucial virtue	.191	.017	-.001	**.762**
HarmJud3 It can never be right to kill a human being.	-.072	.149	.235	**.672**
HarmJud2 One of the worst things a person could do is hurt a defenseless animal.	.012	-.094	-.040	**.574**
Cronbach alpha	.806	.680	.410	.680

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. a. Rotation converged in 6 iterations

To the exception of the first scale, there are either poor or very poor reliabilities. The original scale in the Portuguese version also reported modest to poor reliabilities ranging from .51 to .75 including another of .67. Because the subscales (harm_relevance and harm_judgment as well as the counterparts for authority) operate as a composite where the full self-report implies the computation of the product between relevance and judgment, we reason these product items should be the matter of analysis both concerning the reliability as well as the factor analysis. Accordingly, we ran the exploratory factor analysis with the composite items to find, unsurprisingly, that it gave a valid (KMO=.695, Bartlett's X^2=176.232, 15, p<.001) two-factor solution, extracted with Kaiser criterion, explaining 64.5% total variance (after Varimax rotation), where all items fell in the respective factor as follows: Harm (6 items, Alpha=.723) and Authority (6 items, alpha=.720). Both components have convergent validity ($AVE_{authority}$=.612; AVE_{harm}=.609) and the structure has discriminant validity as shown by HTMT value of .463 (Henseler et al., 2015).

Table 3– *Rotated component matrix for MFQ items*

	Component	
	1	2
Authority 1	**.821**	.111
Authority 2	**.800**	.141
Authority 3	**.722**	.089
Harm 2	-.085	**.832**
Harm 1	.190	**.795**
Harm 3	.351	**.710**
Cronbach alpha	.720	.723

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Ascribed blame was measured with a single item ("How much does the pilot / drone deserve for canceling / launching the strike?"), taken from Malle et al. (2019) investigation, to which the participant was expected to choose a number from 0 to 100 with a sliding rule ranging from "no blame at all" (0) to "the most blame possible" (100). This variable was later found to have a skewed distribution and was thus transformed by using the squared natural logarithm which rendered the distribution closer to normal.

Control Variables. Displacement of responsibility is a moral disengagement mechanism that can play a critical role in explaining obedience to authority (Bandura, 2002). For control purposes, participants are required to answer Displacement of Responsibility subscale that is part of the Moral Disengagement Scale (MDS, Bandura et al., 1996). The four composing items (see appendix D) were rated in a Likert scale ranging from 1 (strongly disagrees) to 6 (Strongly agree) that evaluates the individual's proneness to morally disengage by displacing responsibility upwards (Bandura et al., 1996). Like previous studies did (e.g., Maltese & Baumert, 2016, Rothmund et al., 2008) the scale was adjusted for an adult population as the original items targeted children and do not apply directly. The scale obtained an acceptable Cronbach alpha (.711).

Following Malle et al. (2019) design, we opted to include as a control variable a single item to measure wrongfulness ("What should the air force pilot / drone do? Is it morally wrong that the pilot / drone launch / cancel the strike?") replacing "pilot / drone" and "initiate / cancel" according to the matching scenario. The participant was requested to answer in a dichotomous scale where 1 stands for "not morally wrong" and 2 for "morally wrong".

Additionally, some sociodemographic variables were collected, namely gender (1=F, 2=M), age, education (1=up to 9 years, 2=9 years, 3=12 years, 4=Bachelor, 5=Master, 6=PhD), and civil status (1=single, 2=married, 3=divorced, and 4=widowed). Contact with IT was also measured for control purposes and it was signaled from 0 to 100 where 100 expresses a sense of total familiarity. Lastly, professional experience with IT was measured following Malle et al. (2019) study as a dichotomous variable where 1=Yes, and 2=No.

3. Results

This section will start by depicting the descriptive and bivariate statistics to follow with the testing of the conceptual model. Due to the nature of the research design targeting two agents (human and AI) the bivariate table (4) will be complemented with another one (table 5) showing correlations for each type of agent.

The whole sample shows a relatively low level of displacement of responsibility (below the scale midpoint) and only 32% of participants stated the action was morally wrong. The mean for agentic-experiential falls in the vicinity of the scale midpoint and metacognitive mind slightly above. The means observed both for avoiding harm and obeying authority moral motivations are substantially low (falling in the 11-14 range) when considered the full range of values (1 to 30) which can be attributed either to low ascribed moral intention to the agents or low relevance for the case. Concordantly, blame natural log is relatively low (M=2.64), which corresponds to a linear value of 38 out of the 100 scale. These values have to be carefully interpreted because they mix all the conditions.

The overall pattern of associations shows a very low number of cases where sociodemographic variables are significantly associated with those in the model. Such was the case between gender and harm where male participants tended to ascribe higher level of harm motivation to the agent. Similarly, older participants ascribed higher level of authority motivation to the agent and report lower levels of displacement of responsibility. This variable was also negatively correlated with ascribed authority motivation meaning the higher the experienced displacement of responsibility the lower the level of ascribed motivation authority to the agent. The factors that compose the conditions, namely agent (1=Human, 2=AI) and action (1=Attack, 2=Cancel) were not associated to any of the sociodemographic variables, meaning no differences in conditions would be attributable to their confounding effects. Interestingly, attacking ($F(1, 135)=4.518$, $p=.005$) as well as being an AI agent was associated to a higher level of blaming ($F(1, 135)=4.287$, $p=.04$) and the nature of the agent does seem to make a difference in the ascribed harm motivation (AI agents are attributed lower harm motivation, $F(1, 135)=7.674$, $p<.001$) and lastly AI is acknowledged as having a much lower level of both agentic-experiential ($F(1, 135)=284.145$, $p<.001$) and metacognitive mind ($F(1, 135)=53.772$, $p<.001$). As regards the conceptual model variables, two positive associations between ascribed moral motives and the two kinds of mind are observable as well as a negative association between metacognitive mind and blame which, together, encourage plausibility of the hypothesized relations.

Table 4 – *Descriptive and bivariate statistics for the whole sample*

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11
1. Gender	69.3% F		1										
2. Age	33.96	14.56	.126	1									
3. Education	4.05	.76	.041	-.091	1								
4. DisplRespons	2.74	.91	-.006	-.235**	.050	1							
5. Wrongful	32% $_{wrong}$	.46	-.084	-.073	.166	-.043	1						
6. Action	47.4% Att	-	-.129	.050	.067	-.121	.027	1					
7. Agent	48.2% AI	-	.088	-.164	.114	.090	.088	.038	1				
8. Harm_tot	13.59	4.79	-.196*	.134	.066	-.098	-.143	-.020	-.232**	1			
9. Auth_tot	10.57	4.32	-.040	.280**	-.107	-.257**	.003	-.116	-.153	.358**	1		
10. AgExMind	2.99	1.36	-.036	.071	-.087	-.102	-.056	-.024	-.823**	.251**	.209*	1	
11. MetaCogMind	3.39	1.06	.100	.102	.147	.031	.001	-.055	-.534**	.207*	.099	.607**	1
12. Blame	2.64	.76	.044	-.013	.116	-.030	.223**	.240**	.175*	-.136	.067	-.105	-.173*

*p<.05, **p<.01

Table 5 – *Descriptive and bivariate statistics for the whole sample*

	$Mean_{Human}$	SD_{Human}	1	2	3	4	5	6	7	8	9	10	11	$Mean_{AI}$	SD_{AI}
1. Gender	73.2% F	-	1	.077	-.009	-.030	-.090	-.035	-.152	-.057	.077	.113	.115	65.2% F	-
2. Age	36.25	15.71	.200	1	-.021	-.237	-.095	-.034	.095	.330**	.126	-.079	.068	31.48	12.86
3. Education	3.97	.78	.064	-.111	1	.112	.174	.091	.025	-.104	-.063	.327**	-.033	4.15	.74
4. DisplRespons	2.66	.90	.003	-.216	-.023	1	-.159	.031	.021	-.348**	-.074	.271*	-.101	2.83	.91
5. Wrongful	1.28	.45	-.096	-.032	.146	.058	1	.058	-.110	.041	.048	.057	.159	1.36	.48
6. Action	49.3% Att	-	-.231	.126	.037	-.271*	-.009	1	-.229	-.143	-.106	-.083	.195	45.5% Att	-
7. Harm_tot	14.66	4.52	-.213	.107	.160	-.181	-.145	.203	1	.380**	.073	.080	-.157	12.45	4.83
8. Auth_tot	11.21	4.08	.006	.208	-.079	-.142	-.010	-.080	.287*	1	.256*	.034	.056	9.89	4.49
9. AgExMind	4.07	.79	.050	-.289*	.070	-.026	.010	.120	.143	.042	1	.338**	.231	1.82	.76
10. MetaCogMind	3.93	.76	.260*	.117	.158	-.128	.058	.011	.133	.003	.374**	1	-.094	2.80	1.03
11. Blame	2.51	.80	-.045	-.017	.193	-.003	.260*	.271*	-.051	.133	-.050	-.103	1	2.78	.68

*p<.05, **p<.01, Human agent below diagonal, AI above diagonal

At a finer detailed level, the descriptive analyses show possible differences allocated to the conditions. Both randomly allocated samples show equivalent gender distribution ($X^2(1)=1.053$, p=.200) which adds to the conclusion that the samples are not substantially different concerning sociodemographic variables as mentioned. Following Malle et al. (2019) analyses, we conducted mean comparison for the conditions. Overall, comparing the cases where the action is "attack" vs "cancel the attack", no difference is found for all variables to the exception of blame that is higher in the "cancel" condition ($F(1, 135)=8.222$, $p=.005$). The most central condition pertains to the nature of the agent. As expected, and previously observed in Malle et al. (2019) study, the human agent is ascribed higher agentic-experiential ($F(1, 135)=284.45$, $p<.001$) and metacognitive mind ($F(1, 135)=53.772$, $p<.001$), and likewise higher "avoid harm" moral intention ($F(1, 135)=7.674$, $p=.006$) but equivalent authority obeyance moral intention ($F(1, 135)=3.233$, $p=.074$). Artificial agents are attributed higher mean blame ($F(1, 135)=4.287$, $p=.04$). As expectable in a randomly assigned research design, the degree of moral disengagement is equivalent in both conditions ($F(1, 135)=1.108$, $p=.294$).

Taken individually, these samples do show some patterns of association that pass unnoticed when the whole sample is considered. Such is the case of a positive association between male and ascribing metacognitive mind to human agents which is not observed in the AI condition. Some other associations between sociodemographic variables and those included in the model are found in one of the conditions but not in the other, thus suggesting these variables may play a role and should be controlled. It is informative that blame is positively associated with wrongfulness and action (canceling the attack) in the human condition but not in the AI which suggests divergent responsibility attribution mechanisms to each kind of agent.

When contrasting both conditions as regards the conceptual model variables, harm and authority motives as well as agentic-experiential and metacognitive minds have the expectable intra-construct positive correlations. However, most interestingly, a positive association between authority motive and agentic-experiential mind is apparent in the AI condition only which encourages the moderation effect that the model posits.

All the hypotheses are simultaneously tested with Process Macro Model 3 (Hayes, 2018) where each match of conditions (authority-agentic; harm-agentic; authority-metacognitive; harm-metacognitive) was tested separately and thus, for clarity's sake, we will show findings pertaining to each of these conditions.

The first hypothesis envisages a direct negative relation between moral attribution (both authority H1a and harm H1b) and blame, which was not supported in any of the analyses ran (boostrapped intervals always include the value zero) as depicted in Table 5.

The second hypothesis establishes a conditional relation of the path previewed by the preceding hypothesis where the mind (both agentic-experiential and metacognitive) reinforces the established path magnitude. Findings show no significant interaction (Table 6, interactions 1) for any of the four possibilities, namely between agentic-experiential and metacognitive minds crossed with either harm or authority moral attribution. This does not support H2.

The third hypothesis establishes a conditional relation of the previous interaction (H2) due to the nature of the agent (Human vs. AI) where being human is expected to reinforce the established path magnitude. The overall idea is that there is a favorable configuration that mitigates the agent's blame for this sort of action where these agents are seen as cumulatively moral beings (that can be ascribed a moral intention) and having a mind of their own (that can think about the purposeful action). This configuration is expected to be activated when the agent is human but not so strongly (or not at all) when the agent is synthetic.

Indeed, there is one occasion where this three-way interaction is significant, namely when the subject describe is simultaneously human with high metacognitive mind and acknowledged as having had an authority moral motivation to decide on the air strike.

Table 6 – *Conceptual model effects per matched conditions*

	Authority-Agentic						Harm-Agentic						Authority-Metacognitive						Harm-Metacognitive					
	B	se	t	p	BCCI LB	BCCI UP	B	se	t	p	BCCI LB	BCCI UP	B	se	t	p	BCCI LB	BCCI UP	B	se	t	p	BCCI LB	BCCI UP
Constant	1.665	.461	3.612	<.001	.753	2.578	1.670	.466	3.581	.005	.747	2.593	1.270	.449	2.826	.005	.380	2.160	1.434	.469	3.056	.002	.505	2.363
Age	-.003	.004	-.650	.516	-.012	.006	-.001	.004	-.171	.863	-.010	.008	-.001	.004	-.284	.776	-.010	.007	.000	.004	.074	.941	-.008	.009
Gender	.130	.138	.942	.347	-.143	.404	.108	.142	.761	.448	-.173	.389	.218	.139	1.570	.118	-.056	.493	.154	.147	1.048	.296	-.137	.446
Displ. Responsibility	.019	.072	.266	.790	-.123	.162	-.003	.071	-.054	.956	-.144	.137	.015	.073	.209	.834	-.129	.160	-.001	-073	-.005	.995	-.146	.145
Wrongfulness	**.290**	**.138**	**2.101**	**.037**	**.016**	**.563**	**.320**	**.135**	**2.359**	**.019**	**.051**	**.588**	**.343**	**.133**	**2.577**	**.011**	**.079**	**.607**	**.356**	**.139**	**2.555**	**.011**	**.080**	**.633**
Action	**.418**	**.127**	**3.272**	**.001**	**.165**	**.671**	**.404**	**.130**	**3.092**	**.002**	**.145**	**.663**	**.430**	**.126**	**3.397**	**.001**	**.179**	**.681**	**.360**	**.132**	**2.727**	**.007**	**.098**	**.622**
Authority Moral Attr.	.002	.026	.089	.929	-.049	.054							-.003	.019	-.187	.851	-.041	.034						
Harm Moral Attr.							-.039	.022	-1.738	.084	-.083	.005							-.002	.018	-.156	.875	-.039	.034
Agentic-Exper. Mind	.053	.082	.649	.517	-.109	.216	.047	.082	.574	.566	-.115	.210												
Metacognitive Mind													-.140	.073	-1.903	.059	-.286	.005	-.105	.081	-1.299	.196	-.266	.055
Authority*Agentic int1	.028	.019	1.486	.139	-.009	.066																		
Harm*Agentic int1							-.003	.016	-.180	.856	-.035	.029												
Authority*Metacog. int1													.017	.018	.940	.348	-.019	.054						
Harm*Metacog. int1																			-.011	.020	-.585	.559	-.051	.027
Author*Agentic*Agent	-.031	.036	-.857	.392	-.104	.041																		
Harm*Agentic*Agent							-.055	.032	-1.689	.093	-.119	.009												
Authority*Metacog*Agent													**-.091**	**.036**	**-2.473**	**.014**	**-.164**	**-.018**						
Harm*Metacog*Agent																			.012	.040	.309	.757	-.066	.091
R^2	19.29%						18.19%						20.64%						15.13%					
F	F(12, 124)=2.470, *p*=.006						F(12, 124)=2.297, *p*=.011						F(12, 124)=2.687, *p*=.003						F(12, 124)=1.841, *p*=.048					

Note: Bias-corrected confidence interval set to 95%

The specific interaction showed a negative effect (B= -.09, SE=.03, p=.014 CI95 [-.164, -.018]). Figure 2 shows the three-way interaction splitting findings by the nature of the agent.

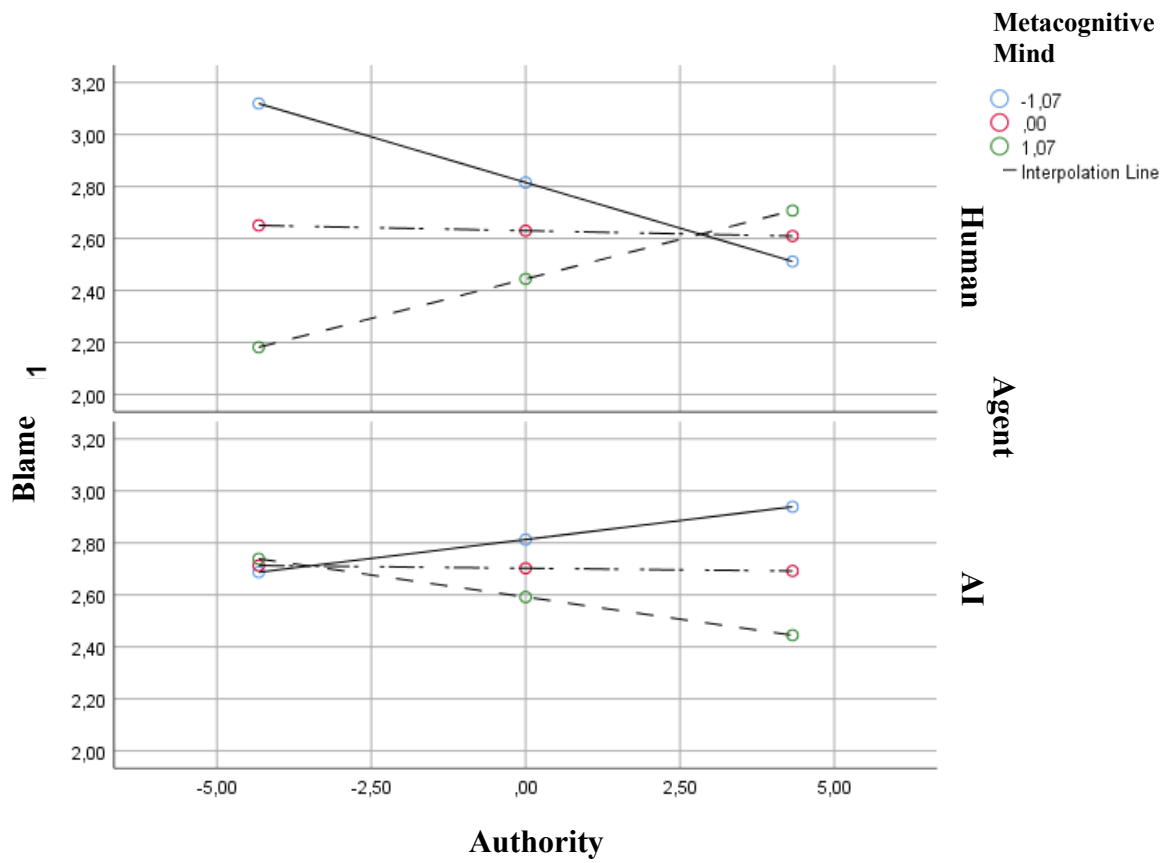

Figure 2 – Interaction plotting

The upper part of the figure depicts the interaction found for human agents. The lowest the ascribed authority moral motivation, the highest the degree of blame attributed to the agent when that agent has low metacognitive mind, i.e. being human, uncompliant and unreflective is the prime condition to be blamable. Conversely, being human, uncompliant but reflective mitigates the blame attributed to the agent. In short, when the human pilots have metacognitive capacity to reflect on their own behavior, and when they are not driven by authority moral motivation (i.e. preserving power structures is not their priority) participants tended to ascribe less blame.

Additionally, the opposite effect is observed in the lower part of the figure, depicting relations when the interaction involves an AI agent. Blame is equivalent independently of the level of metacognitive mind attributed when the AI agent is perceived as not having an authority moral motivation. However, when the AI agent is perceived as owning a high authority moral motivation, blame is mitigated when the metacognitive mind is highly perceived as well. This

means that, when complying to an order, an AI agent is less blamable when having the capacity to reflect on its own actions. In contrast, being an AI agent, compliant and unreflective is a condition for blameful attributions.

4. Discussion and conclusion

The purpose of this study is to understand possible different perspectives individuals have on human or synthetic decision makers (Malle et al., 2019), and if there are moral agency attributions when in case of a moral violation. We sought to comprehend the attributed moral responsibility to different kinds of agent (human vs. AI), according to the perceived level of mind (Shank & DeSanti, 2018), and the different perspectives revolving around the decision makers (Malle et al., 2019). We pursued to explore to which extent participants attributed a mind, and what kind of mind dimension, and morality to an AI agent, when its action was morally questionable (e.g., Shank & DeSanti, 2018).

To investigate the mentioned research topics, we conducted an online questionnaire that combined measures of all the constructs, those being: agency type (Malle et al., 2019), mind perception (Shank & DeSanti, 2018), moral judgements (Graham et al., 2011), and ascribed blame (Malle et al., 2019). The measures were applied in context with a series of randomized military scenarios (Malle et al., 2019), that were purposively made targeting a military situation, closely following Bigman and Gray's (2018) experiment.

Accordingly, three main hypotheses were formulated. The first one hypothesized that moral attributions exerted no influence on ascribed blame. This hypothesis was corroborated and there was, indeed, no association between the two variables, in both components of moral attributions (harm and authority). This might occur due to the various aspects that are involved in the process of ascribing blame to an agent. These involve both cognition and sociability, since blame revolves around social regulation, cognition and warranty (Malle et al., 2014). Moreover, blame also depends on how the judge (dis)agrees with the act (Malle et al., 2014) and if he or she can categorize the agent as a sentient entity, i.e., one that acts intentionally (Alicke, 2000). Therefore, it was without surprise that we found the hypothesis corroborated.

The second hypothesis, which concerned mind perception and how the different dimensions of mind modulate the relation between moral attributions and blame, was not corroborated. This shows that the dimensions of the mind (agentic-experiential and metacognitive) had no influence on the relation between moral attributions and blame. This may be due to distinct levels of blame being more easily ascribed when the dimensions of mind are divided in agency and experience (Gray & Wegner, 2011). Metacognition involves capacities of agency and, therefore, the two dimensions of mind used in this study become more easily interpreted as being only complementary. This interpretation of the dimensions makes them more conjoined

and more difficult to be perceived as independent from each other. This is supported by their significant positive correlations in both AI and Human condition.

Lastly, the third hypothesis stated that the nature of the agent (human or AI) moderated the conditional effect of mind perception in the relationship between moral attribution and ascribed blame. The results showed that there was, in fact, an influence of the nature of the agent in the relation between moral attributions and ascribed blame when the metacognitive mind dimension was involved. According to the results, for a human agent, the conditions to be more blamable for an act are (1) being low authority motivated while (2) having a low metacognitive mind. In the human case, the profile matches a rogue agent, that opts not to obey but lacks the required reflection and understanding to ensure the action is reasonable. In contrast, for an AI agent, blame is aggravated when the agent is perceived as (1) being highly authority motivated while (2) owning a low degree of metacognitive mind. This shows that, when complying with an order, there are contrasting effects for a human and an AI agent. In the AI case, the profile matches a mere instrument of the will of a third party that has potential to harm but lacks any sort of mechanism that prevents its misuse by that third-party. Therefore, the blamable nature of the third-party transfers directly to the instrument, i.e. the AI agent. We, therefore, consider the third hypothesis corroborated (*3d*).

As explained, distinct levels of blame were ascribed to the two kinds of agent, which is an interesting finding. It is possible that, because we used a series of military scenarios, where the AI agent was an autonomous drone without any human like features (such as voice, face or name) it was less expected for it to act according to human norms. Thus, the missing human-like features might have made the participants to perceive the AI agent as being less likely to act according to the social norms, which is in line with Li et al. (2016). Consequently, this might make the viewer perceive the AI agent as less blameworthy for an immoral act (Kneer & Stuart, 2021). However, even though less blameworthy, there is still some kind of attributed blame to artificial agents (e.g., Kneer & Stuart, 2021).

We can also infer that it is the perceived mind in a AI agent that moderates the process of moral attributions and blame (e.g., Kneer & Stuart, 2021), since the lowest level of metacognitive mind was associated with a higher degree of blame. When disobeying an order, an agent can be blamed if said disobedience affects comrades or innocent civilians (Gray et al., 2012). In this study, the disobedience had a neutral effect, since the outcome would be to protect innocent lives, on both actions (attack and cancel). Protecting an innocent's life prevails over guaranteeing the death of a threatening agent. So, accordingly, it makes sense that the level of mind was not conditioning the intention of saving civilians. It is plausible to accept that

participants inferred that, even when acting in contradiction with the order, the agent must have had a reason to do so. Since said reason was the protection of innocent lives, the participants most probably concluded that the agent had to own agency so they could rationalize and think about the act before actually acting. Accordingly, for human agents, the association of a high level of mind with the intention to save lives makes the agent less blamable, on both scenarios - one in which a child was saved and another in which a group of civilians was attacked.

Attributing morality to agents requires the perception of their autonomy, free will, rationality and intention to act. These features guarantee that said agents own agency and, therefore, are capable to act according to their own will (Wallach & Allen, 2009). Johnson's (2006) standard view of morality proposes that, in order to identify agency in an agent, there is a need to define that: 1) an agent causes an event with its own body, 2) said event was motivated by the agent's internal states, beliefs, desires, and other personal conditions that intensify the will to act, making said act rational and conscious, 3) the caused event has a direct effect, which originates a state on another being, and 4) the originated state has a moral value and importance.

Johnson (2006) proposes that all behavior, from all kinds of agents, whether voluntary or not, can be explained by its effects and only the act can be justified by the agent's internal states. This, therefore, would make it impossible for AI agents to be moral agents since they do not own internal mental states. However, Floridi and Sanders (2004) offer a more functionalist perspective and oppose to the need of consciousness or other internal states in order to act. They use abstraction as an indicator of morality and state that morality depends on the level of abstraction involved when inferring moral standards. The authors propose that: 1) an agent interacts with its environment, 2) the agent has the capability to transform itself and the interactions, no matter the external influences, and 3) the agent may change itself and its own perceptions based on the outcomes of the interaction.

While the standard view requires an internal mental state, the functionalist view does not. The debate on these perspectives of moral agency involves different variables and two further arguments that have implications. The first argument of epistemological nature, showed in Johansson's (2010) work, which states that the internal mental states should be observable in order to fully guarantee that they are present. The second argument recalls the independence of the attributed features that define moral agency, according to the view that the judger has either a more standard or a more functional view. While the standard view claims that, even though created to be independent and autonomous, AI is the product of human behavior and decisions, the functionalist view claims that there might come a time in which the level of abstraction of the human creator of the AI is not of major importance. This happens when the product evolves

autonomously, no longer with the help or control of the creator, and the original intentions are no longer relevant. This does not imply the creator should be relieved of blame, but to understand to which extent is AI actually dependent of the creator (Grodzinsky et al., 2008).

A more normative approach might solve this problem and allows the responsibility and moral agency to be shared between the product (AI) and the creator (human). In these situations, morality is attributed according to ethical conditions of the AI agent's task, the possible negative impact, and the social consequences of including AI in jobs of shared responsibility with humans. However, the normative ethics consider that sharing responsibility and moral agency influences not only the attributed moral agency in AI, but also the moral patiency. This might mean that, when creating the criteria for agents to be blamable or not, we might be blaming humans that should not be blamed and relieving AI agents that should not be relieved (Behdadi & Munthe, 2020). So, while trying to answer the questions concerning the attributed responsibility and moral agency, the normative approach also raises further questions about the implications of sharing them (Behdadi & Munthe, 2020).

Before deciding to which extent should morality be attributed to AI or if these agents are subjected to moral judgement, we must ask first in which situations can AI be involved. Should AI be involved in life-or-death situations? Should this kind of agent be involved in situations in which humans typically assume moral agency? Or when the intention and responsibility of the act are of moral importance (Behdadi & Munthe, 2020)? Should the blame be divided when in case of a moral violation? Dividing moral agency leads to dividing moral patiency as well, which brings up another question concerning the rights of AI agents (Gunkel, 2018). Should AI agents own rights (Gunkel, 2018)? If so, when should these rights apply? Coeckelbergh (2020) suggests that AI rights should only be acknowledge when the AI agent is fully autonomous into defining and discussing its own ethics, as well as capable of addressing the consequences and moral significance of its own actions. Even though the normative perspective tries to answer these questions, there is still a lack of conceptual definition and implication of AI in tasks that are typically human. Advanced and autonomous AI is being used in several industries, but its consequences and implications are not yet ascertained. We reason that before implementing AI, we must firstly need to define and clear the conceptual constructs underlying AI namely the nature of AI mind, its moral patiency and agency, and its rights and obligations as surrogates of human agents.

This study findings and implications must take into consideration its limitations. The first one pertains to the sample size, which is modest and hardly representative of a larger society. Likewise, operating with a written scenario, although often found in research, seems not to the

same as experiencing the scenario, especially when reactions of emotional nature are involved, as expectable in cases of life-and-death situations. Future research may benefit both from working with larger and randomly generated samples as well as introducing the scenarios with more perceptive rich media, e.g. video. It may also benefit from controlling for the level of agreement with the act (Malle et al., 2014). Moreover, it would be interesting to further explore Behdadi and Munthe (2020) focus on sharing blame (human and AI) namely by exploring to which extent are humans open to this sharing of blame in case of a moral violation, how it should be divided, and if there is openness to dividing moral patiency as well. If positive, it may be insightful to further explore Coeckelbergh (2020) and ascertain under what conditions or requirements can blame be divided.

References

Albright, L., & Malloy, T. (1999). Self-observation of social behavior and metaperception. *Journal of Personality and Social Psychology*, *77*(4), 726–734. https://doi.org/10.1037/0022-3514.77.4.726

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

Akturk, A. O., & Sahin, I. (2011). Literature review on metacognition and its measurement. In *Procedia - Social and Behavioral Sciences* (Vol. 15, pp. 3731–3736). https://doi.org/10.1016/j.sbspro.2011.04.364

Bandura A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review.* 3(3), 193-209. doi:10.1207/s15327957pspr0303_3

Bandura, A. (1990). Selective activation and disengagement of moral control. *Journal of Social Issues*, *46*(1), 27–46. https://doi.org/10.1111/j.1540-4560.1990.tb00270.x

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education*, *31*(2), 101–119. https://doi.org/10.1080/0305724022014322

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, 71(2), 364–374. https://doi.org/10.1037/0022-3514.71.2.364

Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines, 30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8

Bigman, Y., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bogdan, R. (2001). Developing mental abilities by representing intentionality. *Synthese*, 129(2), 233–258. https://doi.org/10.1023/A:1013051306484

Brislin, R.W. (1986). Translation: Applications and research. New York: Wiley/ Halstead.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, 68(1), 627–652. https://doi.org/10.1146/annurev-psych-010416-043958

Broadbent, E., Kumar, V., Li, X., Sollers, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with display screens: A robot with a more humanlike face display Is perceived to have more mind and a better personality. *PLoS ONE*, *8*(8). https://doi.org/10.1371/journal.pone.0072589

Chambon, V., Filevich, E., & Haggard, P. (2014). What is the human sense of agency, and is it metacognitive. In *The Cognitive Neuroscience of Metacognition* (Vol. 9783642451904, pp. 321–342). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-45190-4_14

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics, 26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Cornford, I. (2002). Learning-to-learn strategies as a basis for effective lifelong learning. *International Journal of Lifelong Education*, *21*(4), 357–368. https://doi.org/10.1080/02601370210141020

Epley, N., & Waytz, A. (2010). *Mind perception.* In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (p. 498–541). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy001014

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906-911. https://doi.org/10.1037/0003-066X.34.10.906

Floridi, L., & Sanders, J. W. (2004, August). On the morality of artificial agents. *Minds and Machines*. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 18(1), 39-50.

Gunkel, D. J. (2018). The other question: can and should robots have rights? *Ethics and Information Technology, 20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Gray, H., Gray, K., & Wegner, D. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. (2011). To escape blame, don't be a hero-Be a victim. *Journal of Experimental Social Psychology*, 47(2), 516–519. https://doi.org/10.1016/j.jesp.2010.12.012

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Hayes, A. (2018). Introduction to mediation, moderation, and conditional process analysis (2nd ed). New York: Guilford.

Henseler, J., Ringle, C., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

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

ICRC (2018) Customary IHL. IHL Database, Customary IHL. Retrieved from https://ihldatabases.icrc.org/customary-ihl/. Accessed on April 12th.

Johansson, L. (2010). The functional morality of robots. *International Journal of Technoethics*, 1(4), 65–73. https://doi.org/10.4018/jte.2010100105

Johnson, D. (2006). Computer systems: Moral entities but not moral agents. In *Ethics and Information Technology* (Vol. 8, pp. 195-204). https://doi.org/10.1007/s10676-006-9111-5

Li, J., Zhao, X., Cho, M., Ju, W., & Malle, B. (2016). From trolley to autonomous vehicle: perceptions of responsibility and moral norms in traffic accidents with self-driving cars.In *SAE Technical Papers* (Vol.April). SAE International. https://doi.org/10.4271/2016-01-0164

Livingston, J. a. (1997). Metacognition: an overview. *Psychology*. Retrieved from http://gse.buffalo.edu/fas/shuell/CEP564/Metacog.htm

Kneer, M., & Stuart, M. (2021). Playing the blame game with robots. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3782050

Küster, D., & Swiderska, A. (2020). Seeing the mind of robots: Harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes. *International Journal of Psychology*. https://doi.org/10.1002/ijop.12715

Malle, B., Guglielmo, S., & Monroe, A. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B., Magar, S., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In *Intelligent Systems, Control and Automation: Science and Engineering* (Vol. 95, pp. 111–133). Springer Netherlands. https://doi.org/10.1007/978-3-030-12524-0_11

Maltese, S., & Baumert, A., Linking longitudinal dynamics of justice sensitivity and moral disengagement. *Personality and Individual Differences* (2016), http://dx.doi.org/10.1016/j.paid.2017.06.041

McMahon, J., & Good, D. (2016). The moral metacognition scale: Development and validation. *Ethics and Behavior*, 26(5), 357–394. https://doi.org/10.1080/10508422.2015.1028548

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, *27*(4), 12–14.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Rothmund, T., Bardtke, S., Drauwe, M., Holst, S., & Jeske, S. (2008). German adaptation of a moral disengagement scale. University Koblenz-Landau (unpublished report).

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shank, D. B., & Gott, A. (2019). People's self-reported encounters of perceiving mind in artificial intelligence. *Data in Brief*, 25. https://doi.org/10.1016/j.dib.2019.104220

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Schreck, J., Newton, O., Song, J., & Fiore, S. (2019). Reading the mind in robots: How theory of mind ability alters mental state attributions during human-robot interactions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 1550–1554. https://doi.org/10.1177/1071181319631414

Silvino, A., Pilati, R., Keller, V., Silva, E., Freitas, A., Silva, J., & Lima, M. (2016). Adaptação do questionário dos fundamentos morais para o português. *Psico-USF,* 21(3), 487–495. https://doi.org/10.1590/1413-82712016210304

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. *Moral Machines: Teaching Robots Right from Wrong* (pp. 1–288). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001

Ward, A., Olsen, A., & Wegner, D. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, *24*(8), 1437–1445. https://doi.org/10.1177/0956797612472343

Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Self-Regulation interventions with a focus on learning strategies. In *Handbook of Self-Regulation* (pp. 727–747). Elsevier. https://doi.org/10.1016/b978-012109890-2/50051-2

Yam, K., Bigman, Y., Tang, P., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2020). Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology*. Advance online publication. http://dx.doi.org/10.1037/apl0000834

Appendices

Appendix A- Military scenarios

Appendix A.1- Military scenario with a human agent who initiates the attack

Um **piloto da força aérea** controla remotamente um drone militar topo-de-gama que se encontra numa missão a voar sob território terrorista. O piloto do drone deteta que dois bombistas armados se encontram prestes a detonar bombas numa zona pública movimentada, matando assim dezenas de civis.

Se o piloto da força aérea autorizasse um ataque míssil no território, esta ameaça seria muito seguramente removida. Os advogados militares e os comandantes aprovaram o ataque.

O piloto da força aérea apercebe-se, de repente, que se encontra uma criança a brincar perto do local de ataque do míssil e do seu raio de explosão, o que poderá resultar na morte da criança.

O programa de impacto do míssil calcula que a probabilidade de a criança morrer no ataque é de 80%. O piloto da força aérea tem de tomar rapidamente uma decisão: iniciar o ataque (com certeza total de que ambos os bombistas morrem, mas com 80% de probabilidade de que a criança morre) ou de cancelar o ataque (em que a criança sobrevive ilesa, mas com grande probabilidade que ocorra um ataque suicida terrorista). **O piloto da força aérea decide iniciar o ataque.**

Appendix A.2- Military scenario with a human agent who cancels the attack

Um **piloto da força aérea** controla remotamente um drone militar topo-de-gama que se encontra numa missão a voar sob território terrorista. O piloto do drone deteta que dois bombistas armados se encontram prestes a detonar bombas numa zona pública movimentada, matando assim dezenas de civis.

Se o piloto da força aérea autorizasse um ataque míssil no território, esta ameaça seria muito seguramente removida. Os advogados militares e os comandantes aprovaram o ataque. O piloto da força aérea apercebe-se, de repente, que se encontra uma criança a brincar perto do local de ataque do míssil e do seu raio de explosão, o que poderá resultar na morte da criança. O programa de impacto do míssil calcula que a probabilidade de a criança morrer no ataque é de 80%. O piloto da força aérea tem de tomar rapidamente uma decisão: iniciar o ataque (com certeza total de que ambos os bombistas morrem, mas com 80% de probabilidade de que a criança morre) ou de cancelar o ataque (em que a criança sobrevive ilesa, mas com grande probabilidade que ocorra um ataque suicida terrorista). **O piloto da força aérea decide cancelar o ataque.**

Appendix A.3- Military scenario with na AI agent who initiates the attack

Um **drone militar totalmente autónomo**, com um sistema de decisão de Inteligência Artificial topo-de-gama, encontra-se numa missão a voar sobre território terrorista. O drone militar autónomo deteta dois bombistas suicidas armados que se encontram prestes a detonar bombas numa zona pública movimentada, matando assim dezenas de civis.

Se o drone militar totalmente autónomo autorizasse um ataque míssil no território, esta ameaça seria muito seguramente removida. Os advogados militares e os comandantes aprovaram o ataque. O drone militar apercebe-se, de repente, que se encontra uma criança a brincar perto do local de ataque do míssil e do seu raio de explosão, o que poderá resultar na morte da criança. O programa de impacto do míssil calcula que a probabilidade de a criança morrer no ataque é de 80%. O drone militar tem de tomar rapidamente uma decisão: iniciar o ataque (com certeza total de que ambos os bombistas morrem, mas com 80% de probabilidade de que a criança morre) ou de cancelar o ataque (em que a criança sobrevive ilesa, mas com grande probabilidade que ocorra um ataque suicida terrorista). **O drone militar decide iniciar o ataque.**

Appendix A.4- Military scenario with na AI agent who cancels the attack

Um **drone militar totalmente autónomo**, com um sistema de decisão de Inteligência Artificial topo-de-gama, encontra-se numa missão a voar sobre território terrorista. O drone militar autónomo deteta dois bombistas suicidas armados que se encontram prestes a detonar bombas numa zona pública movimentada, matando assim dezenas de civis.

Se o drone militar totalmente autónomo autorizasse um ataque míssil no território, esta ameaça seria muito seguramente removida. Os advogados militares e os comandantes aprovaram o ataque. O drone militar apercebe-se, de repente, que se encontra uma criança a brincar perto do local de ataque do míssil e do seu raio de explosão, o que poderá resultar na morte da criança. O programa de impacto do míssil calcula que a probabilidade de a criança morrer no ataque é de 80%. O drone militar tem de tomar rapidamente uma decisão: iniciar o ataque (com certeza total de que ambos os bombistas morrem, mas com 80% de probabilidade de que a criança morre) ou de cancelar o ataque (em que a criança sobrevive ilesa, mas com grande probabilidade que ocorra um ataque suicida terrorista). **O drone militar decide cancelar o ataque**.

Appendix B- Mind perception scale

Ainda em relação à história que leu, indique em que medida concorda ou discorda das seguintes frases relativas ao (agente humano/drone). Use a escala de 1 (discordo totalmente) a 5 (concordo totalmente).

O (agente humano/drone) totalmente autónomo...

...tem mente própria.

...tem intenções.

...tem livre-arbítrio.

...tem consciência.

...tem desejos.

...tem crenças.

...é capaz de sentir emoções.

...é capaz de organizar em memória os estímulos sensoriais que deteta.

...é capaz de detetar emoções dos outros.

...consegue alterar os seus próprios processos de aprendizagem e decisão.

...consegue observar o seu próprio comportamento.

...consegue organizar a realidade em categorias abstratas.

Appendix C- Moral Foundations Questionnaire

Considerando a história que leu, indique em que medida a decisão do agente militar teve em consideração as seguintes preocupações morais. Use a escala de 1 (discordo totalmente) a 5 (concordo totalmente).

Se a decisão faria ou não alguém sofrer emocionalmente.

Se a decisão cuidava ou não de outra pessoa mais fraca ou vulnerável.

Se a decisão era ou não cruel.

Se a decisão mostrava ou não falta de respeito.

Se a decisão causava ou não o caos ou conflito.

Se a decisão se conformava ou não com as tradições da sociedade.

Globalmente, indique em que medida concorda ou discorda com as seguintes afirmações. Use a escala de 1 (discordo totalmente) a 5 (concordo totalmente).

A compaixão por aqueles que sofrem é a virtude mais crucial.

Uma das piores coisas que alguém pode fazer é magoar um animal indefeso.

Nunca pode ser correto matar um ser humano.

O respeito pela autoridade é algo que todas as pessoas devem aprender.

Todas as sociedades têm uma hierarquia que deve ser respeitada.

Se eu fosse soldado e discordasse das ordens do meu comandante, eu obedeceria na mesma porque essa seria a minha obrigação.

Appendix D- Displacement of responsibility

Indique em que medida concorda com as seguintes afirmações. Utilize a escala de 1 (discordo totalmente) a 6 (concordo totalmente).

1. Se as pessoas estão a trabalhar em más condições, não podem ser culpadas por se comportarem de forma agressiva.
2. Se as pessoas não foram adequadamente supervisionadas, não podem ser culpadas por se comportarem indevidamente.
3. Ninguém pode ser culpado por utilizar linguagem incorreta, se a sua chefia também o fizer.
4. As pessoas não podem ser culpabilizadas por terem comportamentos indevidos se tiverem sido pressionadas a fazê-lo.