# A psychology and game theory approach to human–robot cooperation

Te-Yi Hsieh (BSc, MSc)

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

December 2021

School of Psychology
University of Glasgow
62 Hillhead Street Glasgow
G12 8QB

# Abstract

Social robots have great practical potentials to be applied to, for example, education, autism therapy, and commercial settings. However, currently, few commercially available social robots meet our expectations of ‘social agents’ due to their limited social skills and the abilities to maintain smooth and sophisticated rea-life social interactions. Psychological and human-centred perspectives are therefore crucial to be incorporated in for better understanding and development of social robots that can be deployed as assistants and companions to enhance human life quality. In this thesis, I present a research approach that draws together psychological literature, Open Science initiatives, and game theory paradigms, aiming to systemically and structurally investigate the cooperative and social aspects of human–robot interactions.

In Chapter 1, the three components of this research approach are illustrated, with the main focus on their relevance and value in more rigorously researching human–robot interactions. Chapter 2 to 4 describe the three empirical studies that I adopted this research approach to examine the roles of contextual factors, personal factors, and robotic factors in human–robot interactions. Specifically, findings in Chapter 2 revealed that people’s cooperative decisions in prisoner’s dilemma games played with the embodied Cozmo robot were not influenced by the incentive structures of the games, contrary to the evidence from interpersonal prisoner’s dilemma games, but their decisions demonstrated a reciprocal (tit-for-tat) pattern in response to the robot opponent. In Chapter 3, we verified that this Cozmo robotic platform can displays highly recognisable emotional expressions to people, and people’s affective empathic might be counterintuitively associated with the emotion contagion effects of Cozmo’s emotional displays. Chapter 4 presents a study that examined the effects of Cozmo’s negative emotional displays on shaping people’s cooperative tendencies in prisoner’s dilemma games. We did not find evidence supporting an interaction between the effects of the robots’ emotions and people’s cooperative predispositions, which was inconsistent with our predictions informed by psychological emotion theories. However, exploratory analyses suggested that people who correctly recognised the Cozmo robots’ sad and angry expressions were less cooperative to the robots in games. Throughout the two studies on

prisoner’s dilemma games played with the embodied Cozmo robots, we revealed consistent cooperative tendencies by people that cooperative willingness was the highest at the start of games and gradually decreased as more game rounds were played.

In Chapter 5, I summarised the current findings and identified some limitations of these studies. Also, I outlined the future directions in relation to these topics, including further investigations into the generalisability of different robotic platforms and incorporating neurocognitive and qualitative methods for in-depth understanding of mechanisms supporting people’s cooperative willingness towards social robots. Social interactions with robots are highly dynamic and complex, which have brought about some unique challenges to robotic designers and researchers in the relevant fields. The thesis provides a point of departure for understanding cooperative willingness towards small-size social robots at a behavioural level. The research approach and empirical findings presented in the thesis could help enhance reproducibility in human–robot interaction research and more importantly, have practical implications of real-life human-robot cooperation.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgement

The outbreak of the COVID-19 pandemic, declared a public health emergency in March 2020, occurred about halfway through my PhD degree. This pandemic cast a shadow over almost everyone in the world, including myself. The sudden halt of all in-person testing experiments and uncertainty of all future research plans, accompanied by some unexpected incidents that happened in my family and personal life, greatly impacted my mental and physical heath. There were times when I had great doubt in myself, and seriously questioned whether I would and could continue this research work. However, it was because of this difficult time that I realised how many kind and selfless people are around to support me. As I look back now, I'm grateful that I decided to persist in this work, with the help from some people I would like to thank below.

First of all, I would like to express my utmost gratitude to my supervisor Prof Emily Cross, whom I had worked with since my master's degree. Her passion and expertise in research had inspired me to undertake this degree and peruse my curiosity in psychology and science. Without her, I would probably not have known and worked in this fascinating social robotics field. In addition to the invaluable academic guidance on my scientific work, what I felt the most grateful was that she is a supervisor who cares not only PhD students' academic performance but also their well-being. Research work is mostly full of challenges. Though it is essential for researchers to learn to deal with difficulties and obstacles, and think of challenges as learning opportunities, eventually, nothing is more important than personal health. I'm glad I learned this lesson at an early stage and had someone to guide me through this PhD journey.

As an international student studying abroad, I felt the luckiest to be able to work with #TeamSoBots in the SoBA lab where diversity is celebrated. Creating a social network alone in a new place is not easy. However, I found settling in much easier thanks to the kindness and friendliness of everyone in #TeamSoBots. Especially I want to thank Bishakha Chaudhury for her superb programming support for my experiments. This thesis would not be possible without Bish's assistance. Also, I'm grateful for the support from Dr. Maki Rooksby in research, public engagement, and career guidance. Maki has very generously shared her experience, patiently answering and discussing any questions I have had. The things I had learned from #TeamSoBots are not limited to the knowledge aspects but also the kindness and care people showed to one another.

One of the most important things I learned during the pandemic and the PhD journey was to cherish the relationships with my family, friends and everyone I care about. The emotional support I got from my partner, family and friends in Taiwan was one of the reasons why I could manage to accomplish things I never thought possible before. Also, I want to thank my parents for raising me to become a person who believes that women, or more precisely people regardless of their genders and races (etc.), can define their own destiny and should have equal opportunities for all fields, including science. Without this belief, I would not have had the courage to go on this journey on my own and to explore my potential.

I dedicate this thesis to my family and to my grandmother, who gave me the best childhood and, despite growing in a generation that women received little education, she paid great attention to me and my sister’s education. I hope she would be proud of what I have accomplished so far.

# Author's declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Printed name: Te-Yi Hsieh

Signature: ______________

# Research Output

## Publications and open-access research materials related to the empirical chapters of this thesis

### *Chapter 2*

- Open Science Framework (OSF) study link: https://osf.io/res67/

- **Hsieh, TY.**, Chaudhury, B. & Cross, E. S. (2020). Human-robot cooperation in prisoner dilemma games. *HRI '20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 257-259.

- **Hsieh, TY.**, Chaudhury, B. & Cross, E. S. (2020, July 8). Human-robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots. https://psyarxiv.com/q6pv7/

### *Chapter 3*

- OSF study link: https://osf.io/p49jv/?view_only=3148c25ace084c6db5d2760778a2d8b9

- **Hsieh, TY**. & Cross, E. S. (2021) The Role of Empathic Traits in Emotion Recognition and Emotion Contagion of Cozmo Robots. *In Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 802-806).

### *Chapter 4 (Introduction and Methods)*

- OSF study link: https://osf.io/tjs8m/?view_only=d52ffba154ed4236b07c663291a5b053

- **Hsieh, TY**. & Cross, E. S. (2022). People's dispositional cooperative tendencies towards robots are unaffected by robots' negative emotional displays in prisoner's dilemma games, *Cognition and Emotion*, DOI: 10.1080/02699931.2022.2054781

## Other work

Timmerman, R.*, **Hsieh, TY.***, Henschel, A., Hortensius, R. & Cross, E. S. (2021). Individuals expend more effort to compete against robots than humans after observing competitive human-robot interactions. *The 2021 International Conference on Social Robotics (ICSR'21)*, 044, v3.

* co-first authors

de Jong, D., Hortensius, R., **Hsieh, TY.**, & Cross, E. S. (2021). Empathy and schadenfreude in human-robot teams. *Journal of Cognition*, 4(1): 35, 1-19.

# Author Contributions

### *Chapter 2:*

**TYH**: Conceptualization, Methodology, Investigation, Data Analysis and Curation, Writing, Visualization; **BC**: Programming, Data Curation, Editing; **ESC**: Conceptualization, Writing, Visualization, Supervision

### *Chapter 3:*

**TYH**: Conceptualization, Methodology, Investigation, Data Analysis and Curation, Writing, Visualization; **ESC**: Conceptualization, Writing, Supervision

### *Chapter 4:*

**TYH**: Conceptualization, Methodology, Investigation, Data Analysis and Curation, Writing, Visualization; **ESC**: Conceptualization, Writing, Supervision

### *Key:*

**TYH**: Te-Yi Hsieh; **ESC**: Emily S. Cross; **BC**: Bishakha Chaudhury

# Chapter 1 General Introduction

> "We are at the beginning of a revolution that is fundamentally changing the way we live, work, and relate to one another. In its scale, scope and complexity, what I consider to be the fourth industrial revolution is unlike anything humankind has experienced before." (Schwab, 2016)

As Schwab (2016) notes in the quote above, technology has been changing our world in an exponential and unprecedented way. The innovation of artificial intelligence (AI), robotics, three-dimensional (3D) printing, and machine learning have not only fundamentally redefined the technical parts of our lives, but have also redefined the scope of our physical and social spheres (Schwab, 2016; Xu et al., 2018). Particularly, the headway made in AI and robotics has stimulated great public interest, as well as concern, among many people (Makridakis, 2017). Science fiction has long played a role in shaping public imagination and fear towards artificial agents (Cross & Ramsey, 2021; Henschel et al., 2021), dating back at least to Karel Čapek's (1920) play 'Rossum's Universal Robots (R.U.R.)', which is the first time the world was introduced to the word 'robot'. In this theatre piece, Čapek depicts a scenario where artificial workers — what he has called robots, and defined as machines originally created to replace human labour — eventually cause the extinction of the human species. Although the kinds of AI technology dreamed up in science fiction usually deviate far from current development of autonomous artificial agents (Henschel et al., 2021; Makridakis, 2017), many science fiction works delve into humanity's concern over AI technology overpowering humanity. This underscores the importance of taking humanity into consideration during the rapid technological advancement (Xu et al., 2018), as well as keeping humans and the front and centre of new social technological developments (Broadbent, 2017a; Cross & Ramsey, 2021; Eyssel, 2017). In the field of social robotics in particular, it is vital to incorporate psychological perspectives to the designs of these machines, and into research exploring human—robot interactions (HRIs), in order to create the robots that can successfully assist people in our society (Broadbent, 2017a; Cross, Hortensius, et al., 2019; Henschel et al., 2020; Kompatsiari et al., 2018). In order to establish the foundations for this thesis, in the following section, I begin by defining what social robots are, as well as an overview of some of the key findings we have learned about them so far from psychological perspectives.

## 1.1. The definition of social robots

In the current literature, a universal definition of social robots is not agreed upon (Dautenhahn, 2007). Although the definitions adopted by individual researchers are similar to some extent, there is not a consensus in terms of what *specific* social characteristics or abilities social robots should equipped with. For example, in Fong et al.'s (2003) study, they proposed that ‘socially interactive robots’ should be able to perceive and display emotions, communicate smoothly, use natural gestures, have personalities, and form social relationships with others. On the other hand, Breazeal (2003) defined social robots by four sub-classes: (1) *socially evocative robots* (i.e., robots that engage people in social interactions); (2) *social interfaces* (i.e., robots that can foster communications through natural social cues and social intelligence); (3) *socially receptive robots* (i.e., robots that learn from social interactions with people but do not actively engage others in interactions); and (4) *sociable robots* (i.e., robots that have anthropomorphic cognition processes in order to form their own motivations and goals in social interactions) (Breazeal, 2003).

Other researchers have stressed the physical embodied characteristics of social robots and include a more human-centred perspective in their definitions. For instance, Bartneck and Forlizzi (2004) proposed that social robots are semi- or fully autonomous embodied robots that interact with users in a way that follow human social norms. Also, Duffy (2003) underscored social robots’ general purposes to reach their own and societal goals. Finally, Dautenhahn (2007) defined social robots by their different purposes and settings in which they might be deployed in our society (e.g., cleaning, entertainment, or healthcare). A robot companion at home is expected to have sophisticated enough social skills and intelligence to engage in day-to-day social interactions with users, as well as physical abilities to assist in household tasks (Dautenhahn, 2007).

Overall, previous researchers defined social robots with diverse focuses on robots’ functionality, artificial intelligence, social skills, and the value to human users and to the society. They also coined various terms to describe the subtypes of social robots as exemplified above (Breazeal, 2003; Fong et al., 2003). In this thesis, I am most interested in exploring how people interact with social robots that are physical embodied (as opposed to appearing as images or videos on a screen). Therefore, I define social robots as embodied artificial agents that are capable of engaging in social interactions with others (Dautenhahn, 2007; Hegel et al., 2009; K. M. Lee et al., 2006). In psychology, interpersonal social interaction refers to the process where two or more individuals exchanging verbal information and/or nonverbal social cues (e.g., affective signals, gestures) (U. Frith & Frith, 2001). Here in the context of HRI, I define social interactions between human users and social robots from a human-centred perspective. Such human—robot social interactions entail social robots responding to and interacting with people in a humanly readable way (for example, through verbal communication, displays of emotional cues, or bodily gestures) so people can perceive and understand social signals sent from robots and act accordingly.

## 1.2. A psychological perspective of human—robot interaction

The emergence of social robots has provoked interest among multiple disciplines, including robotics, engineering, and psychology (Broadbent, 2017a; Cross, Hortensius, et al., 2019; Cross & Ramsey, 2021; Hortensius & Cross, 2018). Research into the design of and human engagement with social robots has brought about brand-new challenges and questions to each of these fields, since these devices exist somewhere between the categories of objects and real social beings, and most people still have extremely limited real-life experience with them (Alves-Oliveira et al., 2016; Cross, Hortensius, et al., 2019; Cross & Ramsey, 2021; Hortensius & Cross, 2018). Particularly, psychology (the science of human perception, cognition, emotions, and behaviours) has been considered one of the most relevant fields through which we can gain insights into human-centred robotic design and development (Broadbent, 2017a; Cross, Hortensius, et al., 2019; Henschel et al., 2020; Kompatsiari et al., 2018). Some of the bigger theoretical questions psychologists are grappling with through investigations into HRI include: could social robots ever become authentic social beings to us?

(Gasser, 2021; Hortensius & Cross, 2018; Pender, 2018); how do people's expectations and attitudes towards robots shape their behaviours in real-life HRI (Nomura et al., 2008; Syrdal et al., 2009)?; and will social robots' human-like characteristics (e.g., anthropomorphic appearances and emotional displays) make us interact with them in a way that is similar to interactions with other people (Hoegen et al., 2018; Kayukawa et al., 2017; Krach et al., 2008; Leite et al., 2008)?.

So far, a large number of empirical studies have investigated the research topics that are conventionally examined in interpersonal settings in the context of HRI. These include, for example, people's empathic responses to robots (Cross, Riddoch, et al., 2019; Riddoch & Cross, 2021; Rosenthal-von der Pütten et al., 2014), trust relationships with robots (Correia et al., 2016; Hamacher et al., 2016; Hancock et al., 2011; Schniter et al., 2020), and personalities factors that shape HRI (Robert, 2018; Walters et al., 2005). However, some unique challenges and limitations have also been identified in HRI as a field (Alves-Oliveira et al., 2016; Baillie et al., 2019; Cross & Ramsey, 2021; Jost et al., 2020; Złotowski et al., 2015). I have selected what I believe to be some of the more major challenges in this space, and outline each of these in the following sections.

### *1.2.1 'Wizard of Oz' experiments versus autonomous robot experiments*

The 'Wizard of Oz (WoZ)' research approach, in the context of HRI research, refers to the experimental designs that a social robot participants interact with is controlled by an experimenter behind the scenes (Belpaeme, 2020). Currently, the WoZ design has been adopted by many studies since not many social robots are yet capable of fully autonomous operation in natural social interactions with naïve human users (Broadbent, 2017; Cross & Ramsey, 2021; Riek, 2012). Although this approach might ostensibly iron out the present technical limitations of social robots (e.g., in natural dialog generation) and could provide insights into what HRI might look like in the future when social robots are more advanced, it remains questionable whether the interactions happening via WoZ can be regarded as real HRIs or whether we should interpret the interactions as interpersonal interactions between a participant and an experimenter (or confederate) embodied as a robot (Belpaeme, 2020; Broadbent, 2017; Riek, 2012).

Given the possible difference between interacting with a fully autonomous robot and with a robot controlled via WoZ means, researchers who adopt the WoZ approach are advised to report explicitly how robots are controlled manipulated in WoZ experiments (Belpaeme, 2020; Riek, 2012). Also, Riek (2012) suggested that it would be helpful for experimenters who play the role of a ‘wizard’ to receive some experimental training beforehand, to minimise the possibility of procedures and manipulations being biased by personal factors. Furthermore, Innes and Morrison (2020) suggest that researchers should consider conducting double-blind experiments where neither participants nor experimenters who control the robots are aware of the experimental manipulations and research purposes. By double-blinding experimental processes, researchers could prevent the undesirable situation where experimenters’ expectations and beliefs either consciously or unconsciously influence the research processes, thus helping to avoid biased results and conclusions (Innes & Morrison, 2020).

On the other hand, fully autonomous robots mean that robots are equipped with the capacities to interact with people without the assistance of human operators. Though HRI research using autonomous robots might be limited to the social and physical abilities that currently available robots can perform, such research provides a more precise and realistic snapshot of how people might interact with social robots “in the wild”, given the present development of robotic platforms. Moreover, since experimenters don’t necessarily need to be present in the same room with participants when they interact with robots, any undesirable “experimenter bias” can be avoided. Experimenter bias, or observer-expectancy effect, means that experimenters who have the knowledge of experimental manipulation might either consciously or unconsciously influence participants’ behaviours and responses during experiments (Bierman & Jolij, 2020; Phillips et al., 2000). Even the mere presence of experimenters could have social facilitation effects where people’s behaviours when they feel the presence or observation of experimenters might be different from when they are on their own (Guerin, 1986). For example, people might behave in a more socially desirable way, or even just pay more attention to the task, when others are around. The adoption of autonomous robots in HRI research therefore has the potential to sidestep some of these potential biases, and ensure that participants' behaviours measured in HRI experiments truly reflect the ways people respond to social robots instead of the ways participants want to present themselves in front of experimenters, or the ways in which participants interact with a human-controlled robot.

### *1.2.2 Physically embodied robots versus virtual robots on-screen*

As mentioned in the Section 1.1, not all definitions of social robots suggest robots should have physically tangible bodies. For example, in Lee et al.'s (2006) study, they suggested that social robots can be embodied either physically in real-life or virtually on-screen, as long as they are able to interact socially with others. A significant amount of studies have investigated people's attitudes and reactions towards robots through online experiments where robots are two-dimensionally presented as images or in videos (e.g., De Jong et al., 2021; Stroessner & Benitez, 2018; Tulk & Wiese, 2018). Although online research or research that displays videos or images of robots via screens provides a valuable point of departure for understanding real-life HRIs, we must still be mindful that interactions with robots on-screen are in no way equivalent to interactions with physically embodied robots (Grossman et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018; K. M. Lee et al., 2006; Wykowska et al., 2016). A growing amount of empirical evidence supports this very point. For instance, research has demonstrated that people empathise an physically embodied robot more than a virtual robot on-screen (Kwak et al., 2013; Seo et al., 2015). Also, people evaluate an physically embodied robot more positively than the virtual one (K. M. Lee et al., 2006; Li, 2015).

Due to the impact of the COVID-19 global pandemic, which shut down in-person experimentation for most researchers around the world in early 2020 (just about halfway through this thesis), online experimentation emerged as a safer and more feasible alternative for continuing to conduct HRI research. However, we should be more careful when thinking about the extent to which findings from online experiments apply to investigations of embodied or real-life HRI (Jost et al., 2020). Ultimately, given the current evidence revealing people's differential responses induced by robots' physical embodiment (Kwak et al., 2013; K. M. Lee et al., 2006; Seo et al., 2015; Wykowska et al., 2016), lab-based experiments and field studies remain irreplaceable by online experiments or screen-mediated interactions (Grossman et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018; K. M. Lee et al., 2006; Wykowska et al., 2016).

### *1.2.3 Generalisability of empirical HRI studies*

The robotic platforms researchers have used for HRI studies vary dramatically in terms of robots' physical appearances and functions (Henschel et al., 2021; Stock-Homburg, 2021). There are humanoid robots that inspired by human shapes (e.g., Pepper, iCub, NAO) as well as robots that resemble animals (e.g., Paro, Miro, KAROTZ). Also, there is an example of Cozmo robots whose designs are shaped and informed by the media (i.e., the titular character in the animated film 'Wall-E'). Researchers have found that robots' physicality can elicit differential expectations towards them, thereby shaping people's attitudes and responses (Duffy, 2003; Goetz et al., 2003). As such, it is essential to acknowledge that just because researchers report people engage with a particular robot in a particular manner, these findings might not be replicated by other robots of different types (Henschel et al., 2020; Hortensius et al., 2018; Hortensius & Cross, 2018; Innes & Morrison, 2020; Jost et al., 2020).

In addition to the diverse robotic platforms researchers have adopted so far for HRI research, different study designs (e.g., lab-based vs online; short-term interaction vs long-term interaction; lab study vs field study) also make generalisability of research findings challenging. For example, as mentioned in the Section 1.2.2, different embodiments of social robots used in research — either physically embodied in the real world or virtually embodied on-screen — can profoundly shape people's responses and attitudes to these agents (Kwak et al., 2013; K. M. Lee et al., 2006; Li, 2015; Seo et al., 2015). Also, in Section 1.2.1, differences between WoZ studies and studies using autonomous robots were illustrated. It is therefore imperative to use caution when interpreting findings from different study designs, using different agents, or different stimulus presentations, and to be realistic about the scope findings from individual studies can be generalised to (c.f., Cross & Ramsey, 2021).

Another issue related to generalisability of HRI studies is the concern of ecological validity of lab-based HRI research (Belpaeme, 2020). For experimental psychologists, it is important to minimise the impact of confounding variables and to perform well-controlled experiments to, for example, clarify causal relationships between variables. However, this approach is usually criticised by others for the lack of ecological validity since the real world is full of noise and all manner of unpredictable, dynamic change (Belpaeme, 2020; Holleman et al., 2020). Similarly, in lab-based HRI studies, researchers have questioned the extent to which people's attitudes and behaviours measured in artificial and controlled lab environments represent the kinds of social interactions people might actually engage in with robots encountered in real-life (Belpaeme, 2020; Henschel et al., 2020). One way to address this issue is to conduct field studies where participants interact with robots in natural settings, such as their homes, workplaces, or hospitals (Agrawal & Williams, 2018; Stubbs et al., 2007; Van der Putte et al., 2019). Though data collected in the wild may involve more noises and confounding factors, and it is not always feasible to move expensive robots out of laboratories, field studies on HRI could provide valuable exploratory insights and are sometimes more suitable for specific research questions (e.g., the application of social robots in autistic therapy and in hospitals) (Belpaeme, 2020; Van der Putte et al., 2019).

In addition to consider the feasibility of field studies, one of the solutions to increase ecological validity in laboratory environments is to prolong the duration of HRI (Belpaeme, 2020; Cross, Riddoch, et al., 2019; Leite et al., 2013). Compared to one-off and short-term interactions, long-term HRI studies prevent potential false conclusions due to the novelty effect (Sung et al., 2009). Several studies of long-term HRI have been carried out (see a review by Leite et al., 2013). For example, in Cross, Riddoch, et al.'s (2019) study, participants took a small Cozmo robot home for a five-day interaction period. Using functional neuroimaging techniques, participants' brain activity when seeing the Cozmo robot and a human actor displaying pain were measured both before and after the five-day interactions. Although the results did not provide evidence supporting that the five-day socialisation with Cozmo led to human-like empathic brain responses when viewing a robot “in pain”, the study demonstrated the feasibility, importance and value of investigating long-term HRI (Cross, Riddoch, et al., 2019).

Another potential solution to improve ecological validity of experimental research is by adopting virtual reality techniques (Parsons, 2015). Parsons (2015) proposed that virtual reality can provide more dynamic and real-life stimuli while keeping experimental environments well-controlled in psychological and neuroscientific research. HRI researchers have also started to use virtual reality for studying human interactions with robots, and have verified the utility of virtual reality as useful research tool in this space (in a way, serving as a mix between embodied and screen-based investigations of HRI; Liu et al., 2017; Villani et al., 2018). Though virtual reality enables more realistic depth perception than screen presentation, the heavy headsets used for virtual reality are not always tolerable or preferable to users (Liu et al., 2017; Villani et al., 2018).

In addition to ecological generalisability of experimental HRI research, it is equally important to discuss whether the results of a specific sample can be generalised to and replicated by participants with other cultural backgrounds, in different experimental settings (Lim et al., 2020). As this issue is not unique to HRI but instead relates to the whole of psychological research in general, I will discuss this in the next section. To sum up this section, I outlined three challenges in HRI research, including Wizard of Oz, online experiments using two-dimensional robots, and generalisability across robotic platforms and from the lab to real life. While there are undoubtedly many more challenges to consider when carrying out investigations of HRI (c.f., Cross & Ramsey, 2021), I have focused on the main areas that are particularly relevant to this thesis. More comprehensive discussions and guidelines can be found in the literature (e.g., Innes & Morrison, 2020; Jost et al., 2020, Cross & RAmsey 2021). In the following section, I turn my attention to another challenge, which, again, is not unique to HRI, but is of serious concern to researchers across many fields, including medicine, biology, and artificial intelligence (Hutson, 2018; Munafò, 2016; Open Science Collaboration, 2015). As one of the meta-goals of the experimental work in this thesis was to ensure the research was conducted as responsibly and reproducibly as possible, in the next section I turn my attention to the reproducibility crisis (as it relates to the field of psychology in particular), and why this is relevant to my thesis work and the field of HRI in general.

## 1.3. Reproducibility crisis in psychology

Recently, the research validity and reliability of psychological studies has been questioned on a field-wide level via large-scale replication studies (Klein et al., 2018; Open Science Collaboration, 2015). Through this initiative, it has been found that the findings from a large proportion of published studies failed to replicate, and the effect sizes found in the replications were mostly smaller than the original reported values. For example, in Open Science Collaboration's (2015) large-scale replications, the average effect size found in the results of the replications was 0.197, which was much smaller than the average of the original published studies (mean effect size = 0.403). Also, in another collaborative replication project 'Many Labs 2', 75% (21 out of the 28 replication studies) of the effect sizes were smaller than the original reported values, and only half of the study results (14 out of 28) found supporting evidence the original findings at the criterion of $p < .0001$ (Klein et al., 2018).

The issue of sample representative was also examined in Klein et al.'s (2018) 'Many Labs 2' project. Commonly, psychological studies have been criticised for carrying out experiments mostly on undergraduate university students in Europe and North America, a population that has been described as 'WEIRD' (an acronym that stands for Western, Educated, Industrialised, Rich, Democratic) (Henrich et al., 2010). In 'Many Labs 2', Klein et al. (2018) looked into the variation between study samples and also between settings. Although they found only three effects (out of the 28 being studied) that showed differences between typical WEIRD and "less WEIRD" samples, their results of heterogeneity tests revealed that the effects that showed the most variation across samples and setting are usually the larger and reliable effects (Klein et al., 2018). In other words, a null effect would show little variation across samples and settings (Klein et al., 2018). This all highlights the importance of replication and the essentiality of taking participants' demographic information into account, especially when investigating effects of interest that are supposed to be larger.

Although the reproducibility crisis is not unique to psychology (Hutson, 2018; Munafò, 2016), this does not mean that we can ignore it or hope other researchers will find solutions to this crisis without as many researchers as possible getting involved. Immediate and visible changes are needed. In the next section, I summarise the Open Science practices that researchers are advised to adopt in order to improve the reproducibility and robustness of our scientific work, and why I believe these practices will be particularly beneficial for the fledgling field of social robotics.

### *1.3.1 Strategies of tackling reproducibility crisis*

The reproducibility crisis not only calls for immediate alteration to publication bias (the fact that statistically significant results are more likely than null results to be published in journals), but also for researchers to adopt more transparent and rigorous research approaches (Munafò, 2016; Open Science Collaboration, 2015, 2017). Some good practices that are conducive to research reproducibility and credibility include:

**(1) Explicitly report the experimental procedures and analysis plans in preregistration or registered reports.**

Practices like preregistered a study plan before data collection on a repository like Open Science Framework (OSF): http://osf.io/ or choosing to the submission option of registered reports could, at least to some extent, prevent p-hacking, selective analysing and reporting (Nosek et al., 2019; Open Science Collaboration, 2017). Registered report is a form of journal articles where an article will undergo two peer-review processes (one before data collection and one after a complete paper written up). The additional first peer-review is to ensure study plans (including hypothesis formation, sample size calculation, and sampling and analysis plans) are legitimate and rigorously designed before collecting data, and if a study plan is approved (which is called "in-principle accepted"), the study results will be publish regardless of null or significant results (Nosek & Lakens, 2014). There are an increasing number of journals that accept articles to publish in the format of registered report (see a list here: https://www.cos.io/initiatives/registered-reports) and evidence has revealed a marked contrast between the results of studies published in conventional ways (the positive result rate = 96%) and results from registered reports (the positive result rate = 44%) (Scheel et al., 2021).

Preregistration and registered reports push researchers to think more rigorously about their hypotheses, analysis, and sample sizes before data collection. This point is strengthened further via registered reports, where research and analysis plans are peer-reviewed and data collection is only carried out *after* the plans are approved by field experts (Nosek & Lakens, 2014). Another advantage of this practice is to distinguish differences between confirmatory hypothesis testing and exploratory analyses. By differentiating the two approaches, researchers should be able to more precisely interpret statistical results (especially of *p*-values) (Nosek & Lakens, 2014; Open Science Collaboration, 2017).

### (2) Perform high-powered research:

In null hypothesis significance testing (NHST), higher statistical power enables us to more accurately falsify a null hypothesis (thereby providing evidence for an effect), when that effect truly exists. As power is determined by the alpha level, sample sizes, and effect sizes of interest, the most straightforward way to increase power is by increasing sample sizes. Given the status quo that low-powered designs are prevalent in literature (Klein et al., 2018; Open Science Collaboration, 2015, 2017), it is important to plan to test a sufficient sample size to achieve high power (ideally higher than 0.9; Lakens, 2014) and thereby prevent false conclusions (Open Science Collaboration, 2017). Unfortunately, recruiting a large sample size is not always feasible for individual research teams, and several useful strategies have been mentioned. Lakens (2014) proposed that sequential analyses — conducting planned interim analyses with a stricter alpha level — allow researchers to have an earlier stop of data collection while controlling for Type I error rates. Furthermore, the Open Science Collaboration (2017) suggests that within-subject designs and stronger experimental manipulations which reduce confounding effects could also help detect effects of interest. Finally, the strength of multi-lab collaboration to integrate resources of individual teams has been perfectly demonstrated in the examples of Open Science Collaboration (2015) and Many Labs 2 (Klein et al., 2018).

#### (3) Ensure materials and data are made available to the science community

Making the research procedures, materials, and analyses transparent can help alleviate the difficulties of replication studies and meta-analyses (Open Science Collaboration, 2017). There are already platforms where researchers can transparently report or share their study materials, for example, the Open Science Framework (OSF): http://osf.io/. Additionally, although most journal editors and reviewers have been aware of the publication bias issue and are more open to considering paper submissions that report null results, it is still useful to make research findings and reports available as preprints (on platforms like PsyArXiv: https://psyarxiv.com/) so colleagues from the same field can keep up with the cutting-edge advancements in a timely manner (Bourne et al., 2017), or, as the COVID-19 crisis has shown us, so the world can access breaking research findings before they have gone through the peer review process (with the usual caveats in place that preprints have not yet undergone the same kind of peer review that published papers have).

### *1.3.2 The implications of reproducibility crisis to HRI research*

As a considerable number of HRI studies share the same methodology with psychological research, the issue of reproducibility is crucial to consider in the field of HRI (Belpaeme, 2020). As Belpaeme (2020) states in the quote below, the recent reproducibility crisis in psychology is not only relevant but also extremely informative to the investigations of HRI:

> “HRI is lucky to mature during one of the biggest revolutions in experimental psychology: the “replication crisis” was one of the most seminal moments in psychology and its repercussions are felt far and wide, including in HRI.” (Belpaeme, 2020, p.355)

A call for more rigorous experimental designs, manipulations, and reporting in the HRI field is urgently needed, since Innes and Morrison's (2020) recent demonstration of the prevalent experimental artefacts and biases in HRI after reviewing articles in the two journals — *ACM Transactions on Human-Robot Interaction* and the *International Journal of Social Robotics*) — as well as papers from the proceedings of the *Annual ACM/IEEE International Conference on Human Robot Interaction*. Some of the biases identified by Innes and Morrison (2020) relate to the points I discussed in Section 1.2 above, for example, most experimenters who played the role of 'wizards' in studies using WoZ designs had complete understanding of the experiments, including manipulations and hypotheses, which can easily bias results. Similarly, there were a number of studies where experimenters were both coders of participants' behaviours and the persons who made predictions of the results (Innes & Morrison, 2020). In both cases, few of them discussed explicitly how these procedures might be biased (Innes & Morrison, 2020). Another limitation brought up by Innes and Morrison (2020) was the outdated theoretical references of psychological theories cited in the HRI studies that sought to examine psychological phenomena — for example, obedience by Milgram's study design, cognitive dissonance, etc. — in the context of HRI. In these examples, the authors did not appear to be aware of recent developments and theoretical updates in the topics, nor did they mention the potential methodological limitations that have been discussed at length in the psychological literature, and how these limitations could impact the result interpretations (Innes & Morrison, 2020).

In a multi-disciplinary field like HRI, it might be unrealistic to expect all researchers to excel in every domain of knowledge and methodological practice that is relevant to a topic. However, since the investigations of HRI will nevertheless involve human participants, it is crucial to incorporate psychological perspectives and be aware of some caveats in experimental designs as this may significantly impact the research validity and development of the field (Belpaeme, 2020; Cross, Hortensius, et al., 2019; Eyssel, 2017; Henschel et al., 2020). In addition to the room for improvement in HRI experimental designs, researchers in HRI have been advocating the good research practices in line with Open Science initiatives, for example, being transparent in results reporting, acquiring diverse samples, and valuing replication studies (Belpaeme, 2020; Innes & Morrison, 2020). Although psychology itself is still evolving, HRI and related fields could already gain considerable insights from the knowledge psychologists built up on experimental control and measurements. Moreover, the recent reproducibility crisis in psychology should inspire HRI researchers to adopt a more rigorous research approach at this early stage.

## 1.4. Game theory for understanding HRI

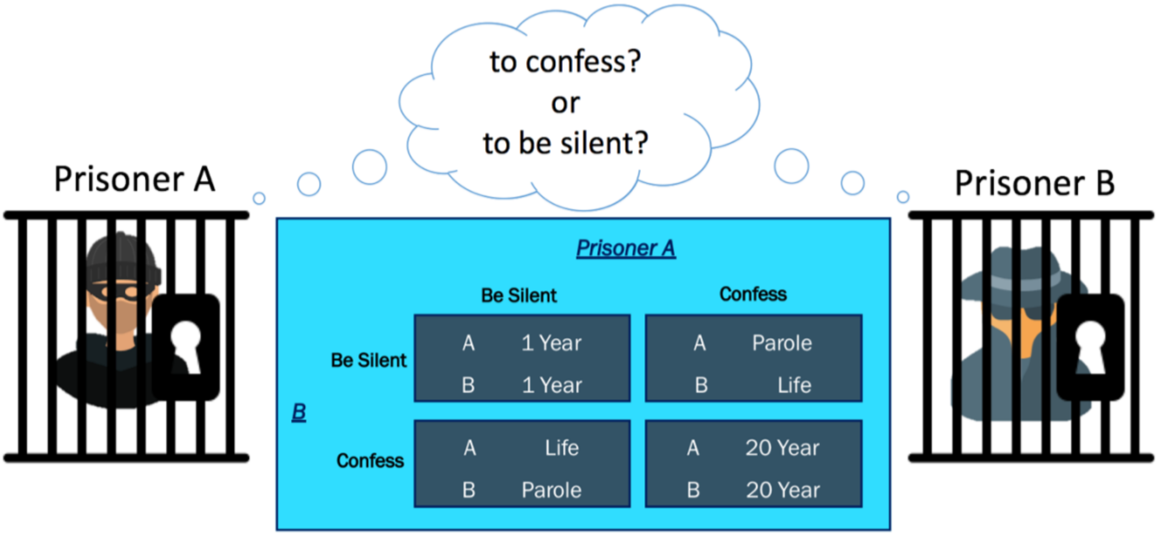
As stated by Innes and Morrison (2020), HRI experiments that involve human subjects require structural interactions and sophisticated designs that allow researchers to clarify effects and the relationships between variables of interest. Therefore, in this thesis, I chose to adopt game theory and the designs of economic games for investigating people’s decision-making processes in HRI.

Game theory — the theoretical framework that seeks to characterise human decision-making processes via structural games and mathematical models — is traditionally a subject in mathematics and economics. Over the past several decades, however, game theory has been applied to an increasing number of fields that extend beyond mathematics and economics, including politics, computer science, biology, and the social sciences, including psychology (Osborne, 2004; Sanfey, 2007). The reason why game theory paradigms hold great psychological interest is due to abundant empirical evidence suggesting that people’s decisions in interactive economic games are not entirely ‘rational’ (i.e., always maximising individual profits) but are often more cooperative, altruistic, and reciprocal (Colman, 2003; Fehr & Fischbacher, 2004; Rapoport & Chammah, 1967; Sanfey, 2007), in contrast to the propositions of classical economic theories (Swanson, 1996). Specifically, our social behaviours in interpersonal settings are mostly in line with social norms (Fehr & Fischbacher, 2004). A famous example of people’s ‘irrational’ social behaviours has been reported in ultimatum games. In an ultimatum game, two players are involved, with one being a ‘proposer’ and the other being a ‘receiver’. The proposer can freely decide the amount of money they are willing to share with the other. The receiver, on the other hand, can only choose to accept or reject the offer. By accepting, both players receive the amounts of money according to the proposer's allocation, whereas by rejecting, both players receive noting. The frequently found result in this game is that approximate half of receivers reject offers that are less than 30% of the total amount of money (Bland et al., 2017). This contradicts the prediction of rational economic decisions, which should be to accept all the possible offers so long as an offer is greater than zero. Moreover, the finding demonstrates how people take interpersonal factors into account when making decisions. Even in situations where they should play strategically to earn money, they would rather forgo individual profits in order to 'punish' the greedy players.

Other empirical research has demonstrated effects of numerous social factors — including trust, attitudes, and expectations towards the game opponent(s) — on social decision-making processes in interpersonal economic games (Berg et al., 1995; Charness et al., 2016; Chaudhuri et al., 2002; Murphy & Ackermann, 2015; Peshkovskaya et al., 2017). Given the valuable insights into human social and economic behaviours revealed via game theory approaches, more and more researchers interested in HRI and in human—computer interaction (HCI) have started to examine how people behave when playing economic games with artificial agents (Correia et al., 2016; de Melo, Carnevale, et al., 2014; de Melo, Gratch, et al., 2014b; Hoegen et al., 2018; Tulk & Wiese, 2018). The two types of games theory that are most frequently used by HRI and HCI researchers are prisoner's dilemma games and ultimatum games. In the following, I detail each of these paradigms in turn.

### *1.4.1 Prisoner's dilemma games*

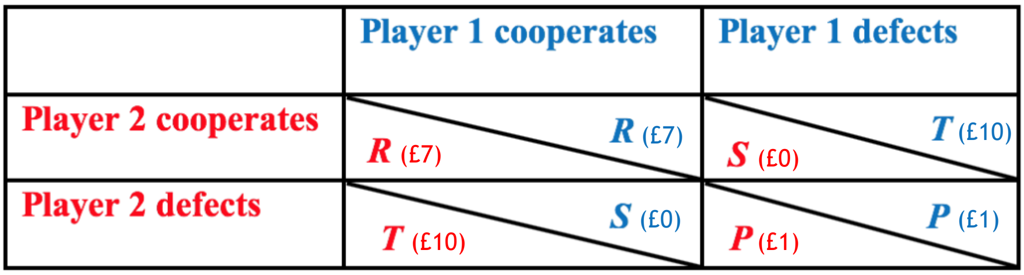
In a classical prisoner's dilemma scenario (**Figure 1-1**), two prisoners ("prisoner A" and "prisoner B" in **Figure 1-1**) are imprisoned in individual jails without any communicational channel, and each of them can choose either to confess their crime or to remain silent (i.e., not to confess). The consequences of individual prisoners' decisions will depend on both of their choices. First, if both prisoner A and B want to cooperate with each other and remain silent. Each of them will serve only one year in jail. However, choosing to confess is tempting since, if one of them chooses to confess, the one confessing will be set free immediately, and the one remains silent will be sentenced to life-time imprisonment, which is the worst possible outcome. In this sense, choosing to confess is regarded as betraying the other. Finally, if both of them are tempted by the potential freedom after confessing, they will all serve 20 years in jail, which is much longer than the 1-year imprisonment (i.e., the outcome of mutual cooperation). This scenario of prisoner’s dilemma games depicts the situation when individual profit is at odds with collective profit, and the games are usually used to measure people’s cooperative tendencies in social interactions (Pothos et al., 2011b; Rapoport & Chammah, 1967).



| | | Prisoner A | |
|---|---|---|---|
| | | Be Silent | Confess |
| B | Be Silent | A 1 Year<br>B 1 Year | A Parole<br>B Life |
| | Confess | A Life<br>B Parole | A 20 Year<br>B 20 Year |

**Figure 1-1.** A classical scenario of a prisoner's dilemma.

Researchers usually adopt a monetary version of prisoner’s dilemmas where the consequences of decisions become different amount of monetary rewards or game points proportional the eventual monetary payoffs (Moisan et al., 2018; Rapoport & Chammah, 1967). **Figure 1-2** present an example of such games. In the situation of one player cooperating while the other does not, the player who chooses not to cooperate earns the highest amount of reward (**Figure 1-2**; e.g., £10), whereas the player who cooperates receives the worst outcome (e.g., £0). When both players choose to cooperate with the other, each of them will receive a moderate reward (e.g., £7). When both players choose not to cooperate with the other, both are given a minimal payoff (e.g., £1).



| | Player 1 cooperates | Player 1 defects |
|---|---|---|
| Player 2 cooperates | *R* (£7)<br>*R* (£7) | *T* (£10)<br>*S* (£0) |
| Player 2 defects | *S* (£0)<br>*T* (£10) | *P* (£1)<br>*P* (£1) |

**Figure 1-2.** An exemplified payoff matrix in prisoner's dilemma games. R = rewards; T = temptation; S = sucker's payoff; P = punishment. The dilemma is defined by two rules: T > R > P > S, and 2R > T + S.  Adapted from Hsieh, TY., Chaudhury, B. & Cross, E. S. (2020). Human-robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots. PsyArXiv. https://psyarxiv.com/q6pv7/

Numerous studies in HCI have adopted the game design to examine the extent to which people would cooperate with virtual agents and to explore the relevant factors (such as agents' emotional displays) shaping the cooperative tendencies (de Melo, Carnevale, et al., 2014; de Melo, Gratch, et al., 2014b; de Melo & Terada, 2019; Hoegen et al., 2018). However, much less is known about people's cooperative tendencies when they play these kinds of games against physically embodied robots. Some researchers have probed reciprocity in games played with a humanoid NAO robot (Sandoval et al., 2016), and the extent to which human-like dialog with a robot influences people's cooperative tendencies towards them (Maggioni & Rossignoli, 2021). The current evidence of human behaviours in human—robot prisoner's dilemma games remains limited, however, in terms of the topics examined and the robotic platforms used. It is important for investigate questions in this area further because people's cooperative willingness towards social robots in prisoner's dilemma games could shed light on future human—robot cooperation in the real world where choosing to work with robot assistants (e.g., business, healthcare, or educational settings) might denote forgoing short-term individual profit (e.g., monetary expenses, time investment in learning how to cooperate robots) in exchange for better collective payoffs (e.g., better and more efficient work performance, as robots don't need to rest or sleep). Cooperative tendencies towards embodied robots in the context of prisoner's dilemma games therefore could be regarded as approximations of the extent to which people would be willing to cooperative with embodied social robots in real life (Kayukawa et al., 2017; Sandoval et al., 2016).

### 1.4.2 Ultimatum games

As introduced above, an ultimatum game is played by two players: one ‘proposer’ who decides the allocation of money or rewards, and one ‘receiver’ who accepts or rejects the offer. If a participant plays the role of a proposer, their allocation of money/rewards can be regarded as the extent to which they wish to be fair to the other participant. One the other hand, a receiver’s decision is a measure of their preference of fairness and how the perception of fairness influences their treatment of the proposer (Bland et al., 2017; Osborne, 2004). HRI researchers have adopted ultimatum games to examine how people make social decisions in games played with robot players, compared to human players. For example, Sandoval et al. (2016) reported that participants who played the proposer role made fairer decisions to a human confederate than a NAO robot. In Torta et al.'s (2013) online study, participants played ultimatum games with a human, a humanoid robot, and a computer. They found a marginally significant effect that participants (playing the receiver’s role) rejected a computer more than they did to a humanoid robot and a human. Also, participants took more time to decide whether to accept or reject a computer’s offer, compared to a robot’s or a human’s offer. Torta et al. (2013) explained the findings by the level of anthropomorphic characteristics of the three agents, which made participants respond to a human and a humanoid robot opponent in a more similar way. However, in Nishio et al.'s (2018) study, the authors did not find any evidence for the main effect of agents’ human-like appearance on participants’ decisions as receivers. In other words, they found that after engaging in short verbal dialogs with game opponents, participants’ game responses to highly human-like android, but not a mechanical robot or a computer, would become more similar to the responses to a human opponent. Finally, Terada and Takeuchi (2017) found that a robot’s emoji-like facial expressions displayed on a monitor “head” could induce more generous offers from human participants when these human participants played the role of proposers. While these findings are beginning to inform our understanding of how people behave in reciprocal economic interactions with embodied robots, much remains to be explored in this space. For example, it remains unclear the extent to which people’s social decisions in ultimatum games are shaped by agent types or human likeness of agents, due to conflicting findings in this space (Nishio et al., 2018; Torta et al., 2013). However, these studies demonstrate that investigations of HRI or HCI in the context of ultimatum games holds great potential to provide important new insights into how people's sense of fairness and willingness to offer might be shaped by the type of player they interact with (e.g., human vs. artificial, human-like vs. mechanoid, embodied vs. screen-based, etc.).

Research into social decision-making processes in economic games carries further implications for real-life social interactions (Chaudhuri et al., 2002; Rand & Nowak, 2013). Interpersonal research on economic game behaviours has demonstrated that people's social decisions measured in experiments are linked with their real-life cooperative and charitable behaviours (Capraro et al., 2019; Capraro & Perc, 2021). Similarly in HRI and HCI research, researchers are becoming increasingly interested in people's cooperative behaviours towards artificial agents (de Melo & Terada, 2019; Hoegen et al., 2018; Kayukawa et al., 2017), since gaining in-depth understanding in cooperation with robots or with virtual agents could yield profound practical values in, for example, industry, education, and healthcare settings. In contrast to self-report data ,which might be biased by social desirability (Fisher & Katz, 2000), behavioural measures in structural games can offer a more robust and well-controlled approach to provide insights into the factors and contexts that shape real-life HRI.

## 1.5. The current research approach

In this thesis, I adopt a research approach that incorporates both psychology and game theory perspectives in an attempt to gain better understanding of how people make social decisions during interactions with a physically embodied robot. More importantly, I aim to identify relevant factors that shape our cooperative tendencies towards robots. The reason why the research focus is on human—robot cooperation is that social robots have considerable practical applications in healthcare, education, and industrial settings (Broadbent, 2017a; Dautenhahn, 2007; Fong et al., 2003). As such, it is easy to foresee a future where humans will need to work with robots in these highly complex social settings. This near-future prospect highlights the importance and urgency of acquiring better understanding in the factors shaping people's tendencies to cooperate with robots, to maximise the utility of robots (Schrempf et al., 2005). In order to address the overarching goal, the research approach I use in this thesis has the following components:

- **Psychological perspective of HRI**: To predict how people would behave in economic games played with robots, the literature on interpersonal games was extensively reviewed. Largely informed by interpersonal literature, the thesis explores the roles of personal factors — including social value orientation, predisposition to anthropomorphism, negative attitudes towards robots, and individual differences in emotion recognition — in human—robot cooperation.

- **Open Science practices**: Two empirical studies (Chapters 2 and 4) were preregistered, and the study in Chapter 4 was submitted and conducted as a registered report (in-principle-acceptance granted by *Cognition and Emotion*). The study materials, anonymous data, and analysis codes for all studies (Chapter 2-4) are fully reported on the Open Science Framework (OSF; see individual chapters for study-specific OSF links).

- **Physically embodied, autonomous robot**: Except for the online study reported in Chapter 3, most of the HRI investigations that compose this thesis were conducted with physically embodied robots that operated autonomously as they played games with participants. Specifically, the robot platform used in the thesis was the Cozmo robot (manufactured by Anki Inc.). These palm-size edutainment robots have high flexibility to be programmed and feature extremely expressive LED screen faces. Researchers have also proposed that they can be suitable research tools for HRI experiments (Chaudhury et al., 2020; Cross, Riddoch, et al., 2019).

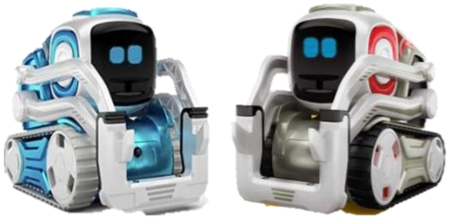


**Figure 1-3.** Cozmo robots used throughout the thesis. Cozmo robots are portable (5 x 7.2 x 10 inches) and highly expressive with an LED screen (128 x 64 resolution).

- **Game theory and economic games**: The experimental designs of the main embodied HRI studies (Chapters 2 and 4) are based on the principles of game theory (prisoner's dilemma games), in order to systemically and structurally examine the social dynamics in the interaction processes. The choice of iterated prisoner's dilemma games is mainly because we can observe the changes of people's cooperative tendencies during iterated game rounds. For example, by plotting people's decisions made in a series of game rounds, we can identify when their cooperative willingness start to decrease. This could provide richer insights into people's social decisions-making process compared to other one-off decision-making tasks (e.g., ultimatum games).

- **Mixed effects modelling**: Given the strengths of mixed effects statistical approach to model both fixed and random effects (Debruine & Barr, 2019; Field & Wright, 2011), most analyses in the thesis were conducted with mixed effects models. Furthermore, mixed effects models enable us to carry out analyses at a trial-by-trial level, instead of doing it only on aggregate data.

### *1.5.1 Topics and variables examined in the thesis*

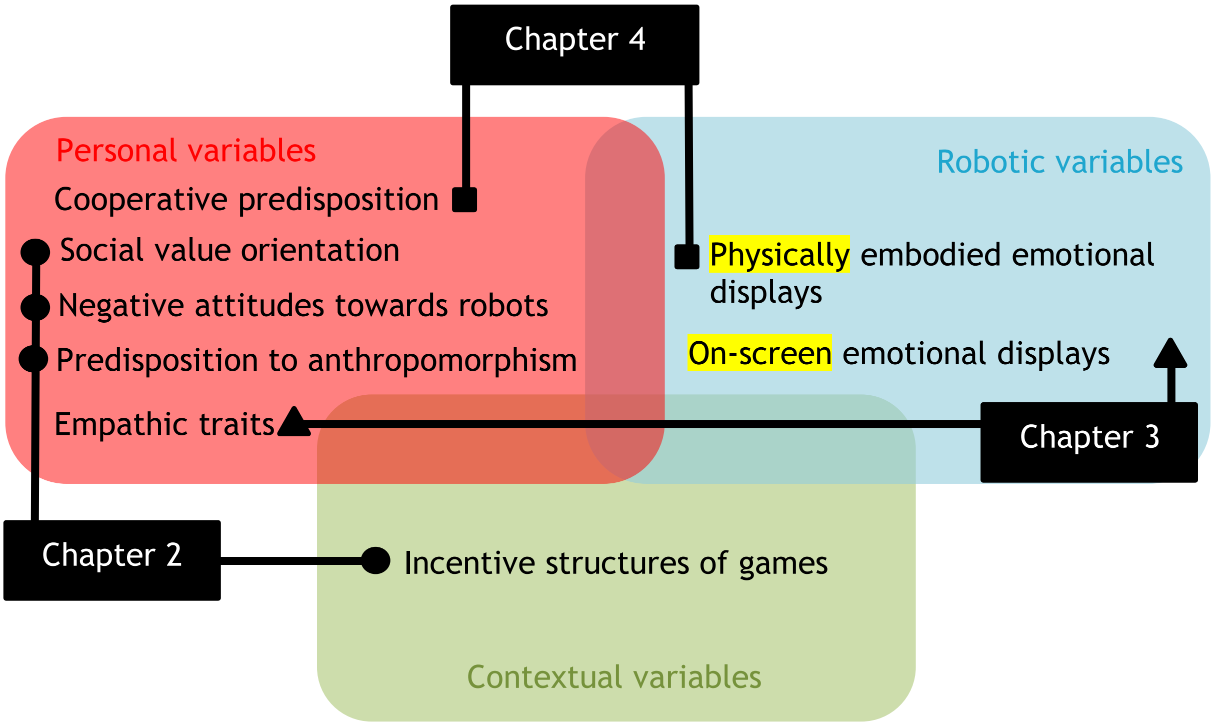

Chapter 4
Personal variables
Cooperative predisposition
Social value orientation
Negative attitudes towards robots
Predisposition to anthropomorphism
Empathic traits
Robotic variables
Physically embodied emotional displays
On-screen emotional displays
Chapter 3
Chapter 2
Incentive structures of games
Contextual variables

**Figure 1-4.** The variables examined in the three empirical studies (in Chapter 2-4). Mainly three categories of variables pertain to human—robot cooperation are investigated, including contextual, personal, and robotic variables.

**Figure 1-4** summarises the variables examined in the three empirical studies that compose the body of this thesis (Chapters 2-4). Overall, the variables can be categorised into three groups: contextual variables, personal variables, and robotic variables. Contextual variables refer to the manipulation made in experimental settings, which in this case, the designs of economic games. Personal variables mean the factors relates to individuals' personality traits and temperaments. Specifically, I explored the impact of social value orientation (i.e., temperamental orientation of altruism) (Chapter 2), negative attitude towards robots (Syrdal et al., 2009) (Chapter 2), predisposition to anthropomorphism (Ruijten et al., 2019) (Chapter 2), temperamental empathic traits (measured by Interpersonal Reactivity Index; Davis, 1983a) (Chapter 3), and cooperative predisposition (i.e., people's default cooperative level when facing a prisoner's dilemma). As we cannot manipulate personal variables, most of the personal variables we examined in this thesis are necessarily exploratory.

Robotic variables involve characteristics of robots, and in this thesis, the focus is particularly on robots' displays of emotional expressions.

### *1.5.2 An overview of the three empirical studies*

In this section, I present an overview of the main research question asked in each chapter and summarise the empirical studies below.

**Chapter 2 — Are people's cooperative tendencies towards a robot opponent shaped by incentive structures of prisoner's dilemma games?**

Prisoner's dilemma games have been used to investigate human cooperative behaviours for a long history (Rapoport & Chammah, 1967) and currently, the games have also been adopted for examining people's willingness to cooperate with artificial agents (de Melo, Carnevale, et al., 2014; Hoegen et al., 2018; Kayukawa et al., 2017). However, the experimental set-up and game designs vary dramatically in the HRI literature, which makes it challenging to form a conclusive perspective about how people cooperate with artificial agents in such economic games. Furthermore, interpersonal evidence has revealed that people's cooperative intentions are shaped by the levels of incentives provided for cooperation in human—human games (Moisan et al., 2018; Rapoport, 1967). Therefore, we developed a lab-based human—robot prisoner's dilemma games played with Cozmo and examined whether different incentive structures of the games led to different cooperative tendencies among participants. In the results, we did not find evidence supporting the effects of contextual incentives on people's cooperative decisions, which stands in contrast to the findings from interpersonal games (Moisan et al., 2018). However, we found significant evidence for a reciprocal pattern in participants' game decisions via exploratory analyses of this dataset.

**Chapter 3 — How well can people recognise Cozmo's emotional displays and to what extent the emotion recognition is shaped by individuals' empathic traits?**

In the online investigation, we acquired high recognition rates for Cozmo's happy, sad, and angry emotion displays, which were higher than the average recognition rates of the previous 43 studies on emotion recognition of other robots (e.g., NAO, Pepper, Keepon robot, and KOBIAN robot; Stock-Homburg, 2021). Contrary to our predictions, we found the empathy subtype ‘empathic concern’ (Davis, 1980) was negatively associated with participants’ (n = 103) recognition accuracy.

**Chapter 4 — Are people’s cooperative tendencies towards a robot influenced by the robot’s emotional displays and do the influences of emotions differ by individuals’ cooperative predispositions?**

After validating the recognisability of Cozmo’s emotional displays, in Chapter 4, the influence of Cozmo’s sad and angry expressions on people’s cooperative tendencies was examined in prisoner’s dilemma games. We also investigated the interaction between the robots’ emotions and people’s cooperative predispositions on cooperative decisions. In the results of 60 datasets, we found significant impact from people’s cooperative predispositions and emotion recognition accuracy on their cooperative decisions in the games. However, contrary to our preregistered predictions, no significant difference was found between the effects of the robots’ sad and angry emotions.

To conclude, the thesis incorporates the psychological and game theory approaches to investigate people’s cooperative tendencies in economic games with robots. The investigation includes the factors of contextual incentives (Chapter 2), personal temperament and traits (Chapters 3 and 4), and agent-related factors (Chapter 4) in human—robot cooperation. The overall findings are conclusively discussed in Chapter 5. The studies outlined here provide a point of departure for rigorous investigations of human—robot cooperation and identify relevant factors shaping people’s cooperative willingness to robots.

## 1.6. Summary

Considering the utilities and advantages social robots can generate for human society (e.g., across clinical, commercial, and educational settings; Broadbent, 2017; Duffy, 2003; Fong et al., 2003), it is important to gain better understanding in how people interact with robots designed to engage us on a social level, and how we can improve the social quality of HRI. In this chapter, I reviewed the unique challenges faced by those conducting HRI research, and the current movement that is gaining momentum among psychologists and other researchers, including those interested in HRI, for conducting more reproducible science. In response to the challenges highlighted by the Open Science movement, the approach I adopt in the thesis is to adhere to Open Science best practices as much as is possible while incorporating the perspectives of psychology and game theory into my empirical work. There is a long history of psychologists building knowledge of the human mind and behaviour via well-controlled experiments and valid measures. On the other hand, game theory and economic game designs provide a structural way to investigate human behaviour and decision-making process in a relatively engaging and entertaining context. Importantly, using tasks derived from game theory for HRI research allows us to base evidence upon the well-developed literature on economic behaviours and social decision-making studies. With this approach, three empirical studies were conducted to build a more complete understanding in people's cooperative tendencies towards robot and to identified relevant factors shaping such tendencies. The factors being explored in the thesis include contextual, personal, and robotic factors. Through the approach I have used and the findings presented via the following chapters, I am hopeful this work highlights the value and utility of using rigorous research methods grounded in well-defined theory from the social sciences for understanding and improving human—robot interactions.

# Chapter 2 Human-robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots

This chapter is a more recent, updated version of the following preprint:

Hsieh, TY., Chaudhury, B. & Cross, E. S. (2020, July 8). Human-robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots. https://psyarxiv.com/q6pv7/

## 2.1. Abstract

Understanding how people interact socially with robots will be important for designing robots to work on social tasks. Here, we investigate undergraduate participants' situational cooperation tendencies towards a robot opponent using prisoner's dilemma games. With two conditions where incentives for cooperative decisions were manipulated to be high or low, we predicted that people would cooperate more often with the robot in high-incentive conditions. Our results showed incentive structure did not predict human cooperation overall, but did influence cooperation in early rounds, where participants cooperated significantly more in the high-incentive condition. Exploratory analyses revealed other two behavioural tendencies: (1) participants played a tit-for-tat strategy against the robot (whose decisions were random); and (2) participants only behaved prosocially toward the robot when they had achieved high scores themselves. Our findings highlight ways in which social behaviour toward robots might differ from social behaviour toward humans, and inform future work on human–robot interactions in collaborative contexts.

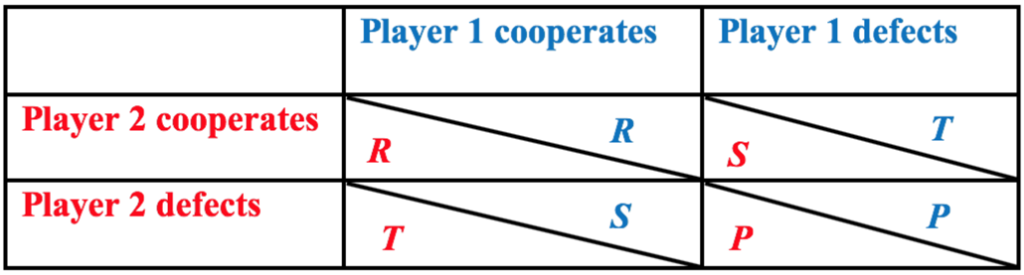
**KEYWORDS:**
Human-robot interaction, human-robot cooperation, prisoner's dilemma games, Rapoport's K-index, reciprocity

## 2.2. Introduction

Social robots are becoming valuable tools for assisting people with daily life, as they take on new roles in healthcare, education, and therapy (Broadbent, 2017a). However, many commercially available social robots suffer from the criticism of not fitting users' expectations, especially in terms of the richness or appropriateness of their social responses, which in turn diminishes people's ability to build long-term, enduring social relationships with these machines (Frennert & Östlund, 2014; Graaf et al., 2016). On one hand, robot designers and engineers are endeavouring to build more socially-sophisticated robots, mostly by increasing robots' human-likeness in terms of physical features, motion, and behaviours (Dautenhahn et al., 2009; Ishiguro, 2006; Yu Ogura et al., 2006). On the other hand, however, others have argued that it is equally, if not more, imperative to gain deeper understanding into the psychological mechanisms and factors that underpin and shape the quality of human-robot interaction, which might not necessarily or solely be the level of human-likeness (Broadbent, 2017a; Cross, Hortensius, et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018; van Straten et al., 2020).

One important aspect of HRI that calls for further psychological investigation is human–robot cooperation (Sandoval et al., 2016; Schrempf et al., 2005; Wu et al., 2016). Cooperation is a pivotal theme in human social behaviours and is key to building mutual and group interests (Axelrod, 1984; Fehr & Fischbacher, 2004). Forming amiable and cooperative relationships with robots should also maximize the utility of robots (Schrempf et al., 2005). Taking eldercare robots as an example, an ideal healthcare robot might take care of various aspects of an elderly individual's everyday life, such as administering medicine, updating family on health status, and providing social interaction to combat loneliness. If elderly individuals do not comply with a robot's health instructions, engage with a robot socially, or accept a robot as a collaborator, the robot's utility is diminished and human users miss out on the potential benefits the robot can offer. A clearer understanding of humans' willingness to cooperate with robots, and the possible factors that shape such willingness, are thus required to reap the social and economic benefits socially assistive robots could offer.

In literature examining human-human and human-robot cooperation, prisoner's dilemma (PD) games are often used to explore collaborative behaviour between individuals (or agents) (Axelrod, 1984; Van Lange et al., 2013). In a classic PD game, two players make simultaneous decisions – to cooperate or defect – with their individual payoff determined by both players' decision on any given trial. If both players choose to cooperate, they each earn a moderate amount but not the highest rewards (***R*** in **Figure 2-1**; e.g., £7 each). If only one chooses to cooperate, the defecting player receives the most rewarding payoff (***T***; e.g., £10), while the cooperating player gets the worst outcome (***S***; e.g., £0). Finally, if both players choose to defect, both receive a minimal payoff (***P***; e.g., £1 each). Thus, while defection might be a profitable choice in terms of individual gain, cooperation brings about better chances of forming cooperative social relationships and of higher mutual gain in the longer term.



| | Player 1 cooperates | Player 1 defects |
|---|---|---|
| Player 2 cooperates | *R*<br>*R* | *T*<br>*S* |
| Player 2 defects | *S*<br>*T* | *P*<br>*P* |

**Figure 2-1.** Payoff matrix of prisoner's dilemma games. R = rewards; T = temptation; S = sucker's payoff; P = punishment. Designs of payoff matrix should follow the two rules: T > R > P > S; 2R > T + S.

Different designs of payoff matrices in PD games significantly influence people's cooperative tendency (Moisan et al., 2018). To standardize PD game incentive structures, Rapoport (1967) (Rapoport, 1967) proposed the K-index as a measure of anticipated cooperation, which is calculated as follows:

$$\frac{(R - P)}{(T - S)}$$

Simply put, the K-index represents the incentives for cooperation provided by a PD game's payoff matrix (Rapoport, 1967). A higher K-index means more incentives for cooperation are provided by the game context, leading to higher cooperation rates among human players (Moisan et al., 2018; Rapoport, 1967). The propositions of Rapoport's K-index are in line with several social behaviour models, such as preferences for social efficiency (Charness & Rabin, 2002) and the cooperative equilibrium model (Capraro, 2013). These models, coupled with empirical evidence from interpersonal PD games (Capraro et al., 2015; Moisan et al., 2018), suggest that people's cooperative tendency is shaped by payoff structures in PD games. This stands in contrast to the neoclassical economic theory's prediction (Swanson, 1996) that people should act rationally to maximise self-gain and therefore defect all along.

Prior work suggests that people employ similar social behaviours in human-robot and human-human economic games. For example, participants in previous studies were equally cooperative with human or artificial opponents (de Melo et al., 2010; Krach et al., 2008; Wu et al., 2016); and have demonstrated the same reciprocal responses to a Nao robot (a child-sized humanoid robot) as to a human confederate (Sandoval et al., 2016). Moreover, other research reports human cooperative behaviours to be impacted by emotions displayed by artificial agents, in line with the appraisal theory of emotion (de Melo, Carnevale, et al., 2014; de Melo et al., 2010, 2012). However, the experimental set-up and designs of these studies varied considerably, making it difficult to assess the role played by contextual factors or draw conclusions about how people behave and cooperate with artificial agents in such economic games. Moreover, most work on understanding human cooperation with artificial agents has been conducted using online economic games (de Melo, Carnevale, et al., 2014; de Melo et al., 2012; de Melo & Terada, 2019; Hoegen et al., 2018; Moisan et al., 2018). As such, we have a limited understanding of human cooperative and competitive behaviours toward a physically present robot (Kayukawa et al., 2017; Sandoval et al., 2016); a gap in knowledge that is becoming increasingly important to fill (K. M. Lee et al., 2006; Seo et al., 2015; van Straten et al., 2020). Therefore, in this study, we examined people's willingness to cooperate with a physically embodied social robot in PD games' where the incentive structures (i.e., K-index) are manipulated. In line with previous research findings (Moisan et al., 2018), we predict that participants who play a high K-index PD game against a robot will make more cooperative decisions than those who play a low K-index game, regardless of a robot opponent's random ordered game decisions. Given that defection is always a preferable option in terms of individual payoff in a single PD game, people's willingness to cooperate with a robot might suggest that we confer some manner of social status to the robot, since cooperation in this context requires a mindset of focusing on collective payoff, and accepting possible betrayal from a robot.

## 2.3. Methods

### *2.3.1. Open Science Statement*

Prior to data collection, all manipulations, measures, and the sample size justification and main hypotheses were pre-registered on the Open Science Framework (OSF): https://osf.io/res67/. Consistent with recent proposals (Galak et al., 2012), we report all manipulations and all measures in the study. In addition, following open science initiatives (Munafò, 2016), the data, stimuli, and analysis code associated with this study are freely available on the Open Science Framework. By making the data available, we enable others to pursue tests of alternative hypotheses, as well as more exploratory analyses. All study procedures were approved by the College of Science and Engineering Ethics Committee (University of Glasgow, Scotland) – approval number: 300180201.

### *2.3.2. Participants*

We recruited seventy participants ($M_{age}$ = 23.6, *SD* = 3.62; 50 females), who had normal or corrected to normal vision and no history of neurological or psychiatric disorders, from the University of Glasgow's psychology subject pool system. The sample was composed of people from diverse national backgrounds, but all currently living in the UK -- 25 (35.71%) of them report being from the UK, 8 (11.43%) from China, 6 (8.57%) from the US, 4 (5.71%) from India, and the other 27 (38.57%) from the rest of 20 different countries **(Appendix A, Table S1)**. The pre-registered sample size was determined by a simulation-based power analysis for generalised mixed-effects models, and the parameters used for simulation were based on Moisan and colleagues' study (Moisan et al., 2018). In order to make sure that our sample was naïve to robots, we measured their daily exposure to robots and also to robot-relevant films or series they had seen (e.g., Westworld, Star Wars, Wall-E) (Riek et al., 2011) before taking part in the PD games. On a scale from 1 (never) to 7 (daily), the median of daily engagement with robots for our sample was 2, with an interquartile range (IQR) of 2. The median number of robot films seen by participants is 3 (IQR = 3) out of 14 films. Two Wilcoxon rank sum tests were performed to test whether the participants in high K-index and low K-index conditions differed in their daily engagement with robots or in the number of films featuring robots seen. We found no difference between the two samples' scores for either of the scales (daily engagement with robots: $W = 730$, $p = .15$; numbers of robotic films seen: $W = 759$, $p = .083$), which verified that the two samples had a similar level of prior exposure and were generally naïve to robots. Participants' informed consent was obtained prior to the experiment beginning, and participants were reimbursed with £6 (per hour) or 4 course credits at the end of the study.

### *2.3.3. Game Design*

Participants played one practice game and one formal PD game with a commercially available Cozmo robot (manufactured by Anki Inc.— **Figure 2-2**). Equipped with four motors, Cozmo fork-lift style arm and head can move in the vertical plane, and its steering wheels can drive in all directions. The Cozmo robot also has a well-developed software development kit (SDK) platform, which users can use to customize its programming using Python language and which we used to develop our human-robot PD game.

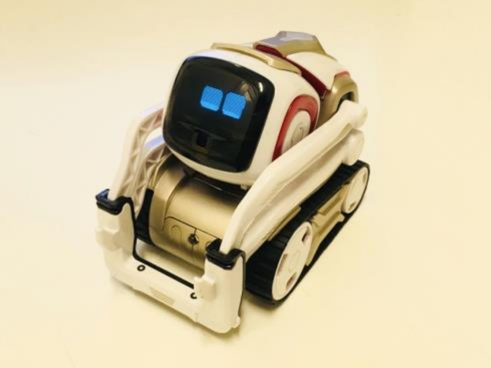
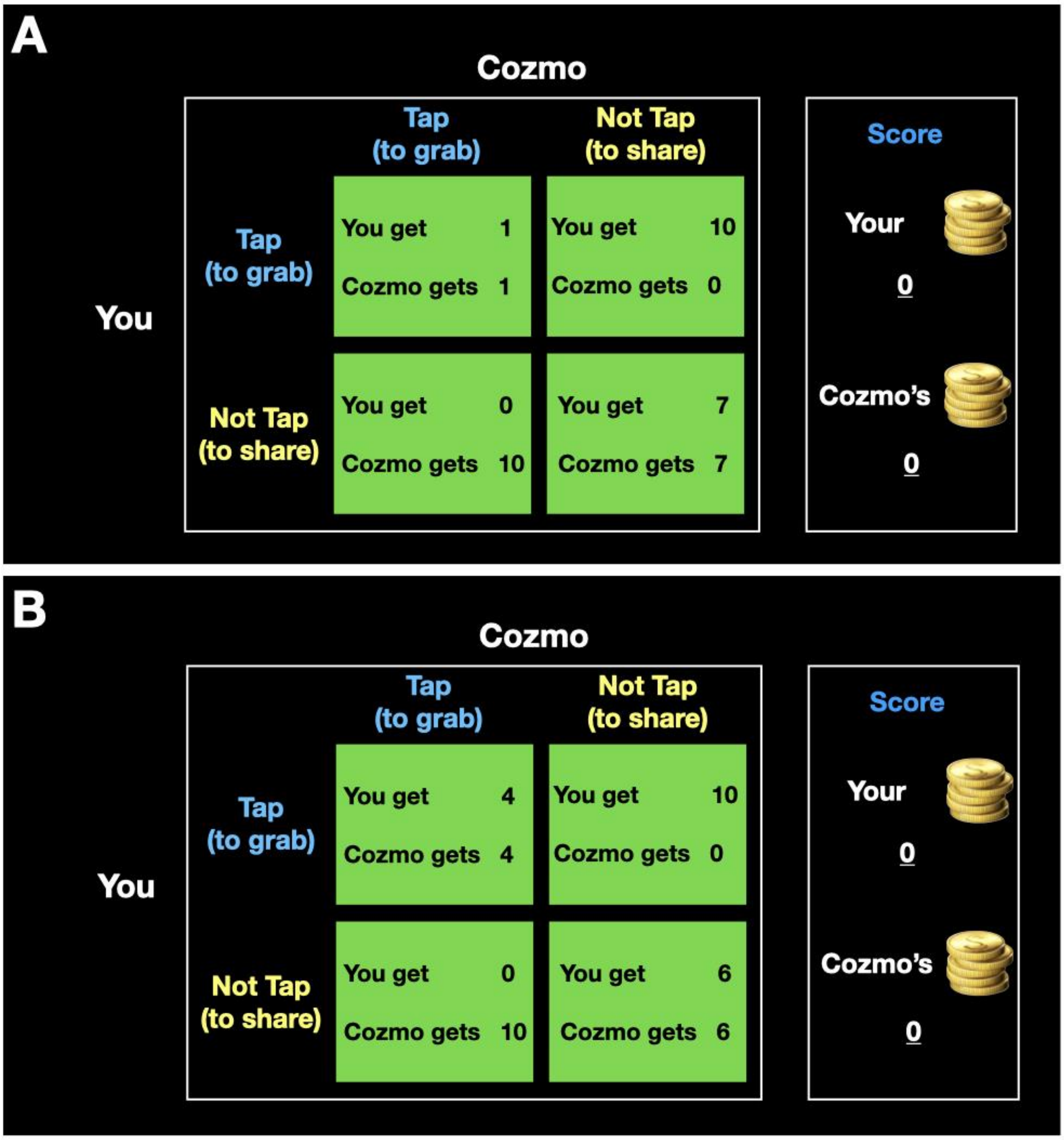


**Figure 2-2.** The Cozmo robot used in this study. Cozmo is palm-sized (5 x 7.2 x 10 inches), with an LED screen (128 x 64 resolution) as a face, which allows it to produce variable and expressive facial expressions, such as happiness, anger, sadness, and surprise. Along with its emotionally expressive face, Cozmo also produces robotic vocal interjections, and can be programmed to speak simple words and phrases with a mechanical sounding voice. However, in the current study, Cozmo's emotionality remained neutral across two conditions.

Before the games started, the experimenter presented a short introductory video to participants about the PD game rules and verbally explained the cover story of the experiment with the following text: "*In this study, we are running a robot competition and aim to know which Cozmo is the best economic game player* (showing participants five other Cozmo robots on the shelf). *In each game round, a certain amount of coins will be available to you and Cozmo, and both players will make simultaneous decisions either to keep all the coins or to share coins with the other. Your individual payoff will depend on both of your decisions. The more coins you get the higher possibility you'll win a shopping voucher in the end, and the Cozmo that wins will be used in our following study, but if Cozmo loses the game, its memory and data will be entirely erased.*"

We used the script of erasing Cozmo's memory as its punishment for losing because prior work has demonstrated that such a prompt is useful in eliciting people's real concerns and empathy towards a robot (Seo et al., 2015), and in the case of this study, should further convince participants that the game is meaningful to Cozmo with real consequences. Participants were randomly assigned to either the high K-index (K=(7-1)/(10-0) = 0.6) game or the low K-index (K=(6-4)/(10-0) = 0.2) game (**Figure 2-3A**, **B**, respectively). The experimenter also answered participants' questions and made sure that they fully understood how to play the game before it started.



A

Cozmo

| You | | Tap (to grab) | Not Tap (to share) |
|---|---|---|---|
| | Tap (to grab) | You get 1<br>Cozmo gets 1 | You get 10<br>Cozmo gets 0 |
| | Not Tap (to share) | You get 0<br>Cozmo gets 10 | You get 7<br>Cozmo gets 7 |

Score

Your 0

Cozmo's 0

B

Cozmo

| You | | Tap (to grab) | Not Tap (to share) |
|---|---|---|---|
| | Tap (to grab) | You get 4<br>Cozmo gets 4 | You get 10<br>Cozmo gets 0 |
| | Not Tap (to share) | You get 0<br>Cozmo gets 10 | You get 6<br>Cozmo gets 6 |

Score

Your 0

Cozmo's 0

**Figure 2-3.** The schematic of game screens. Panel A illustrates a high K-index PD game (K = 0.6). Panel B illustrates a low K-index PD game (K = 0.2).

### *2.3.4. Setup and Apparatus*

We developed the human-robot PD game via Python 3.5.3 to examine people's cooperative tendency in different game contexts (technical details and programmes can be found on the Github page: https://github.com/CozmoGame4Sobot/Prisonner-s-Dilemma). The setup of the experiment is shown in **Figure 2-4**. Participants faced a screen demonstrating the payoff matrix, real-time outcomes, and game scores during the game. Cozmo was placed on the right side of the screen, on a custom-built 4.3 cm thick paper box with an overhang on the side between the two players to prevent participants from seeing Cozmo's interactive cube (see below). This design was to prevent participants from cheating, as some might try to observe Cozmo's decision first before deciding which cube to choose for themselves to maximise payoff. However, the setup still allowed participants to see the whole body of Cozmo since Cozmo would drive backwards to a point where its entire body was visible by participants (panel **B** of **Figure 2-4**), and where it could "watch" the screen until it made a choice for how to respond. This ensured that the robot was within participants' sight for all of the experiment except when it made its choice to keep or share.

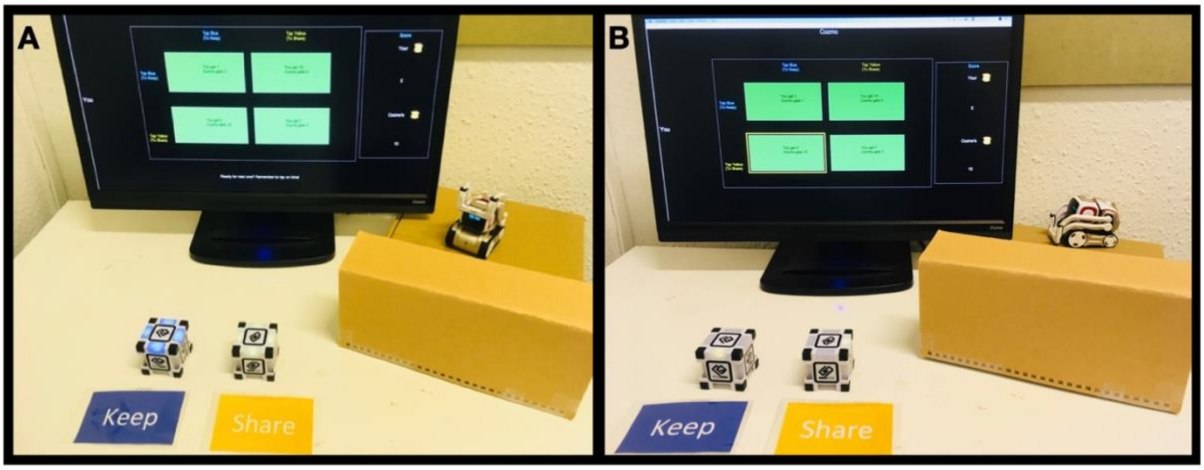


**Figure 2-4.** The experimental setup: (A) the PD game environment from participants' perspective. Participants faced a game screen and the Cozmo robot, and made responses via tapping two interactive cubes, which represented "to keep" and "to share" decisions (B) Cozmo turning to face the screen to 'see' the updated game scores after it had made a decision.

Players used interactive cubes equipped with LED lights inside to make decisions in each game. Each participant was given two interactive cubes, illuminated in different colours to reflect their different choices (participants tapped the blue cube to keep the coins and the yellow cube to share the coins). Cozmo used only one cube to respond, in order to prevent participants from anticipating Cozmo's choices from the direction it drove towards. We designed practice games to familiarise participants with the ways of responding and with the payoff matrices. When practising, participants were asked only to respond to specific goals on the screen (e.g., tap the yellow cube to get 7 coins), to avoid their gaining actual PD game experience before the formal game started.  In formal PD games, we manipulated Cozmo's game decisions to share for 10 trials and to keep for 10 trials, randomized across participants. This decision structure was chosen to control Cozmo's behavioural competitiveness. Both human players and Cozmo made their responses by tapping the top of the cubes, which were connected to a controlling laptop via WiFi, and the players' responses were recorded by Python log files.

### *2.3.5. Measures*

Participants also completed several questionnaires, which were used to explore the role of different human factors in human-robot cooperation, and to measure participants' evaluation of Cozmo after the PD games. First, a social value orientation (SVO) (Murphy et al., 2011) questionnaire was used to measure people's temperamental pro-sociality. The SVO scale has a significant relationship with cooperative decisions in interpersonal social dilemmas (Andrighetto et al., 2020; Murphy & Ackermann, 2015). Participants are asked a series of questions regarding how much endowment a person was willing to ascribe to themselves and to an unknown other, to evaluate the main drive of their social decisions —— whether it was self-profit, collective profit, or relative profit (Murphy et al., 2011). Second, the negative attitudes toward robots scale (NARS) (Syrdal et al., 2009) was included to understand people's prior attitudes to robots in HRI research. Although no study has yet directly tested the relationship between negative attitudes and cooperative behaviours toward robots, the general correlation between such attitudes and people's social behaviours toward robots is suggestive of a possible relationship. Third, we measured participants' predisposition to anthropomorphism (Ruijten et al., 2019), to explore whether an individual's temperamental tendency to humanize non-living things influenced the decision-making process in the current game environment. These three scales were administered before the PD games were performed. Upon completion of these games, participants were asked to evaluate Cozmo's game performance and strategy. Both pre-game and post-game questionnaires were pre-registered and administered via the FormR survey framework (Arslan et al., 2020) (https://formr.org).

### 2.3.6. Procedure

The experiment comprised three main sections. First, participants were given instructions and asked to provide written informed consent. Cozmo would then introduce itself by saying "Hello participants, I'm Cozmo." Afterwards, participants completed a series of PC-based questionnaires, including prior experience with robots scale, NARS, SVO, and the predisposition to anthropomorphism scale. Second, participants completed one practice and one formal PD game with Cozmo in a lab booth. Third, participants completed a final set of questionnaires, including subjective evaluation of Cozmo's performance and strategies, and their demographics. Following all procedures, participants were debriefed, paid, and thanked for their participation.

### 2.3.7. Data Analysis

We pre-registered the use of a mixed effects logistic regression model to examine the main research question: the extent to which people's decisions to cooperate with a robot would be impacted by the different incentive structures of PD games. Additionally, we used a multiple regression model to explore the role of several additional factors on human players' cooperation rates in the human–robot PD games. These factors were assessed via questionnaire and included negative attitudes toward robots, social value orientation traits, and predisposition to anthropomorphism. Finally, for exploratory purposes, we employed two additional mixed effects models to investigate the impact of (1) Cozmo's prior game decisions; and (2) the presentation of players' game scores on individual human decision. Findings from the final two exploratory models can offer insights for future experimental designs on related questions and can help to identify additional factors that shape human cooperative behaviours in the current context.

## 2.4. Results

To investigate our main research question – whether participants' cooperative/non-cooperative game responses in the iterated PD games were influenced by the incentive structure of the PD games – we adopted a mixed effects logistic regression model as our main pre-registered analysis. We followed Barr et al.'s suggestion (Barr et al., 2013) and started with the maximum random effects structures -- see Equation (1) below. The model successfully converged with a fixed effect of incentive structure, subject-level random intercepts, round-level random intercepts, and random slopes for the conditional effects on game rounds. Results of the analysis are shown in **Table 2-1**.

*decision ~ incentive structure + (1 | subject) + (1 + incentive structure | round)* (1)

**Table 2-1**

Results of the mixed effects logistic regression model that examined the effects of incentive structures on human cooperative decisions towards a robot

| | Main model | | | | | |
|---|---|---|---|---|---|---|
| | decision ~ incentive structure + (1 \| subject) + (1 + incentive structure \| round) | | | | | |
| | *Estimate* | *SE* | *z* | *p-value* | *Low CI* | *High CI* |
| intercept | -0.467 | 0.190 | -2.46 | 0.014* | -0.838 | -0.095 |
| incentive structure | -0.301 | 0.232 | -1.30 | 0.194 | -0.756 | 0.153 |
| AIC | 1756.8 | | | | | |
| BIC | 1788.2 | | | | | |
| Log-likelihood | -872.4 | | | | | |

*CI* = 95% confidence interval. **p* < .05; ***p* < .01; ****p* < .001
Abbreviations: *SE* = standard error; *CI* = confidence interval.

The overall incentive structure of PD games was not found to be predictive of participants' game decisions ($\beta$ = -0.301, *p* = .194, 95%*CI* = [-0.756, 0.153]) across 20 game rounds, which means participants in the high K-index game did not share coins more frequently than those in the low K-index game did, in contrast to our prediction.

### *2.4.1. Impact of Incentive Structures on Cooperative Decisions in PD Games*

For descriptive statistics, we calculated cooperation rates by dividing the number of cooperative decisions participants made by the number of total game rounds they performed. The mean cooperation rate of participants playing in the high K-index condition was 0.40, while that of participants in the low K-index condition was 0.34. We also visualized the binary game data (see **Figure 2-5**) to assess the distribution of the participants' decisions (in the two conditions – high and low K-index) across 20 game rounds. The tendency difference between these two conditions was salient especially at the start of games (**Figure 2-5**). When playing the high K-index game, participants began with a high tendency to cooperate, but this tendency declined rapidly after the first 5 game rounds. Conversely, the curve in the low K-index condition remained relatively flat throughout the 20 rounds.

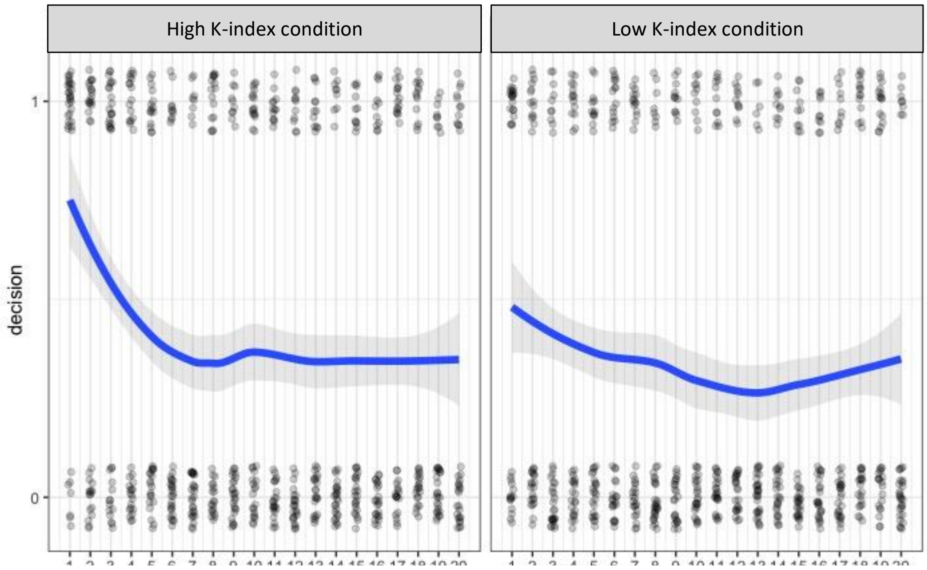


**Figure 2-5.** Distribution of game decisions (sharing coded as 1; keeping coded as 0) across 20 game rounds. A nonparametric smoothed curve was added to provide a clearer view of the cooperative trends. Cooperative decisions were notably more frequent in the high K-index condition than in the low K-index condition, especially in the first few game rounds.

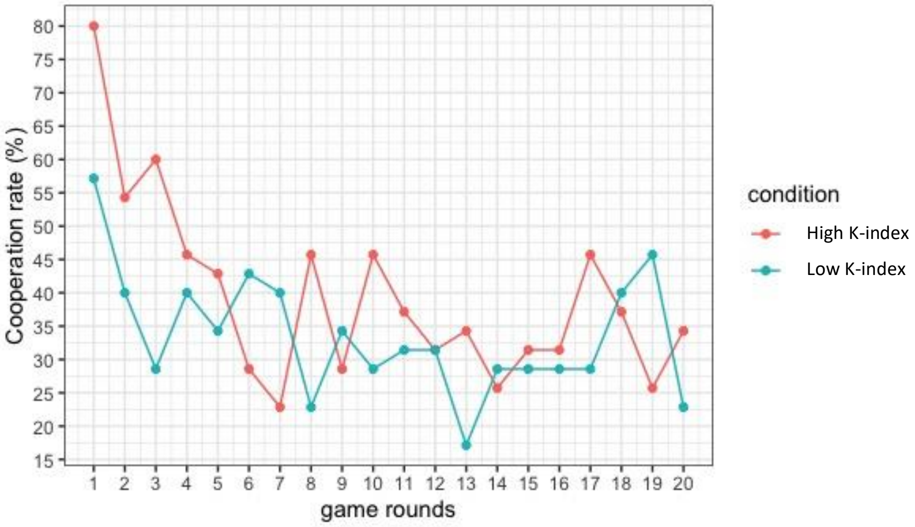


**Figure 2-6.** Changes of cooperation rates (N of subjects who shared / N of total subjects) across 20 game rounds. A higher percentage of participants in high K-index game chose to cooperate (compared to those in the low-K-index condition), but people in both conditions showed decrease and fluctuation in cooperation rates after the initial rounds.

We calculated the average cooperation rates (N of subjects who shared / N of total subjects) per game round and per K-index condition, and further observed that the cooperation tendency declined and fluctuated across both conditions (**Figure 2-6** and **Appendix A, Table S2**). In the first game round, 80% of people in the high K-index game chose to share coins with Cozmo, but only 57.1% of participants in the low K-index condition did so. Similarly, cooperation rates in both game conditions dropped after the first few rounds and fluctuated till the end. We then examined statistically if the participants in these two conditions had different cooperation tendencies, especially in the first game round. Such an analysis can be meaningful because it extracts the possible impact of incentive structure on cooperation from other potentially influencing factors, such as quality of HRI, the order of Cozmo's presented decisions, and all the relevant experiences during the game. In this analysis, we treated decisions made by participants in their first game as one-shot PD games and used a logistic regression model, which revealed that participants' first-game decisions were significantly affected by the game structure ($\beta$ = -1.10, p = .043); participants shared coins (cooperated) more often in the high K-index game than did those in the low K-index game. Odds ratio calculations also suggested that the odds of cooperation in condition one (high K-index, 28/20 = 1.4) was three times more likely than that in condition two (low K-index, 7/15 = 0.47). Below we present exploratory analyses conducted in order to identify possible factors driving the fluctuations in participants' cooperation tendencies.

### *2.4.2. Exploratory analyses*

**Reciprocity in HRI**

Reciprocity is an important theme in human social behaviour and plays a major role in the decision-making process of cooperation (Axelrod, 1984; Fehr & Fischbacher, 2004; Sandoval et al., 2016; Van Lange et al., 2013). Evidence shows that people can behave reciprocally toward social robots in certain contexts (Sandoval et al., 2016). We were therefore also interested to know whether our participant samples responded reciprocally to Cozmo' game decisions (i.e., chose to share coins after Cozmo shared or chose to keep coins after Cozmo kept) in our specific experimental context. To probe this possibility, every game decision made by participants was paired with Cozmo's decision from the previous round, and the data were examined by a mixed effects logistic regression model. Again, we started with a maximal model in terms of random structures (Barr et al., 2013). We then trimmed the complexity to arrive at a model that converged by removing random slopes for incentive structure (given that the focus of this analysis is more on Cozmo's decisions). The final model is provided in equation (2), which included Cozmo's decision and incentive structure as the fixed effects and controlled subject-level and round-level random effects.

*decision ~ Cozmo's decision*incentive structure + (1+Cozmo's decision | subject) + (1 | round)* (2)

The results of exploratory model 1 are presented in **Table 2-2**. This analysis yielded a significant fixed effect of Cozmo's decision ($\beta$ = 0.516, *p* = .046, 95%*CI* = [0.010, 1.020]), suggesting that participants were more likely to share coins if Cozmo shared in the previous round, and more likely to keep if Cozmo did so previously. However, neither the incentive structure (*p* = .544) nor the interaction between Cozmo's decision and the incentive structure (*p* = .110) were predictive of the participants' game decisions.

**Table 2-2**

Results of exploratory analysis 1: mixed effects logistic regression model that examines reciprocity in human-robot interactions

| | Exploratory model 1 | | | | | |
|---|---|---|---|---|---|---|
| | decision ~ Cozmo's decision*incentive structure + (1+Cozmo's decision \| subject) + (1 \| round) | | | | | |
| | *Estimate* | *SE* | *z* | *p-value* | *Low CI* | *High CI* |
| intercept | -0.850 | 0.195 | -4.36 | 0.000*** | -1.230 | -0.468 |
| Cozmo's decision | 0.516 | 0.256 | 2.00 | 0.046* | 0.010 | 1.020 |
| incentive structure | 0.011 | 0.272 | 0.04 | 0.968 | -0.522 | 0.544 |
| Cozmo's decision* incentive structure | -0.609 | 0.367 | -1.66 | 0.097 | -1.330 | 0.110 |
| AIC | 1624.5 | | | | | |
| BIC | 1666.0 | | | | | |
| Log-likelihood | -804.3 | | | | | |

*CI* = 95% confidence interval. **p* < .05; ***p* < .01; ****p* < .001

### The influence of presenting real-time game scores to participants

The designs of PD games that probe human-agent (social robots or virtual agents) interactions differ considerably in the literature (de Melo, Carnevale, et al., 2014; de Melo & Terada, 2019; Hoegen et al., 2018; Kayukawa et al., 2017; Sandoval et al., 2016). One variable among many published studies was the revealing of real-time game scores or not to participants during iterated PD games. In some studies, real-time game statistics (i.e., the players' scores after each round has been played) were shown to participants (de Melo, Carnevale, et al., 2014; Hoegen et al., 2018; Kayukawa et al., 2017), but not in other studies (de Melo & Terada, 2019; Sandoval et al., 2016). In the current study, we presented each player's game scores on the game screen to create a sense of competitiveness and to increase the entertainment value of the game. However, little is known about the extent to which such score presentation drives people's cooperative decisions in games, and to what extent it might affect their decisions. In order to clarify this, we ran a second exploratory mixed effects model – as shown in equation (4) – using subjects' scores and Cozmo's scores as the fixed effects, with subject-level, round-level, and condition-level random effects included. The equation (3) represents the model that converged after removing random slopes for subject's score, and random slopes for Cozmo's score from the maximal model.

*dcision ~ Cozmo's score*subject's score + (1 | subject) + (1 | round) + (1 | incentive structure condition)* (3)

The results of this second exploratory model 2 (see **Table 2-3**) revealed a significant main effect from Cozmo's score ($\beta$ = -0.023, *p* = .009, 95%*CI* = [-0.041, -0.006]). In other words, participants were less likely to make cooperative decisions when Cozmo's scores were higher. Additionally, the interaction between Cozmo's score and the participant's own score ($\beta$ = 0.000, *p* = .001, 95%*CI* = [0.000, 0.000]) was a significant predictor of a subject's cooperative decisions, which is visualized by the R package "effects" (Fox & Weisberg, 2018) in **Figure 2-7**. From this analysis, we observed that as subjects' scores increased incrementally, the relationship between Cozmo's score and the probability of making cooperative decisions changes from a negative correlation to a positive correlation. In other words, if players earned very little, they were less likely to cooperate with or be generous to Cozmo. However, when players had a considerable endowment, they were more willing to share, especially if Cozmo also achieved high scores.

**Table 2-3**

Results of exploratory analysis 2: mixed effects logistic regression model that examines the impact of real-time game scores on cooperative decisions

| | Exploratory model 2 | | | | | |
|---|---|---|---|---|---|---|
| | decision ~ Cozmo's score*subject's score + (1 \| subject) + (1 \| round) + (1 \| incentive structure condition) | | | | | |
| | *Estimate* | *SE* | *z* | *p-value* | *Low CI* | *High CI* |
| intercept | 0.018 | 0.216 | 0.08 | 0.933 | -0.406 | 0.442 |
| Cozmo's score | -0.023 | 0.009 | -2.61 | 0.009** | -0.041 | -0.006 |
| subject's score | -0.010 | 0.006 | -1.67 | 0.095 | -0.022 | 0.002 |
| Cozmo's score* subject's score | 0.000 | 0.000 | 3.26 | 0.001** | 0.000 | 0.000 |
| AIC | 1645.1 | | | | | |
| BIC | 1681.4 | | | | | |
| Log-likelihood | -815.6 | | | | | |

*CI* = 95% confidence interval. **p* < .05; ***p* < .01; ****p* < .001

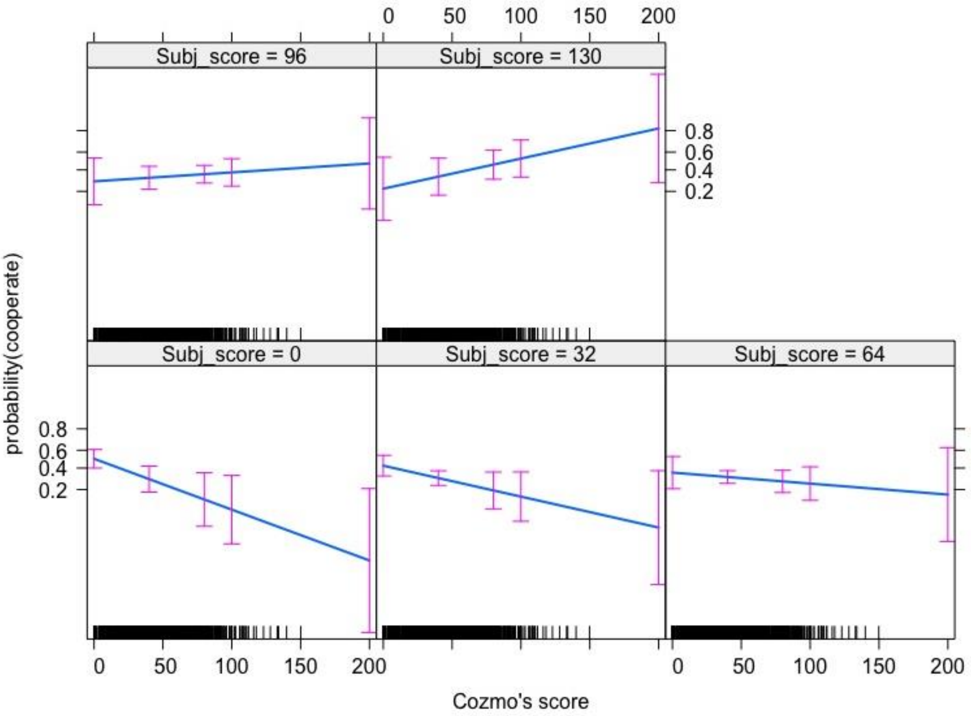


**Figure 2-7.** Interaction between Cozmo's and subjects' scores on probability of cooperation. Although both participants' scores and Cozmo's scores were continuous variables, we used Cozmo's score to define the x-axis as it is a more influential factor (Fox & Weisberg, 2018). The figure demonstrates that, if participants earned low scores (e.g., subj_score = 0), the probability of cooperation with Cozmo decreased as Cozmo won more, but if participants already had earned high scores (e.g., subj_score = 96), the probability of cooperation increased as Cozmo earned more. Pink vertical lines represent standard errors of each value.

### Human factors

Three pre-game scales – NARS, SVO, and predisposition to anthropomorphism – were selected to explore the relationships between human factors and cooperative decisions in PD games, and to inform future research into relevant human factors that shape cooperative and competitive behaviour toward robots.

Results of a multiple regression model ($F(3, 65) = 4.05$, $p = .011$, $R^2 = .119$) showed that only the predisposition to anthropomorphism scale ($\beta = .01$, $p = .046$) had significant impact on the participants' overall cooperation rates (i.e., dividing the sum of times people shared by the total game rounds played). This result suggests that participants who anthropomorphized Cozmo also tended to cooperate with it more. In our further pre-registered and exploratory analyses, we account for the impact of dispositional anthropomorphism by including subject-level random effects. Apart from anthropomorphism scale, neither SVO ($\beta$ = .01, *p* = .137) nor NARS ($\beta$ = -.00, *p* = .145) were found to have a relationship with cooperation rates.

### Subjective Evaluation of Cozmo's performance and game strategy

After participants played PD games against Cozmo, we asked them to guess Cozmo's cooperation rate (i.e., what percentage of Cozmo's decisions were cooperative — choosing to share) and to report Cozmo's and their own game strategies, for the purpose of a manipulation check and exploration. The mean cooperation rate participants guessed was 49.6% (*SD* = 19.64), which suggested that generally, participants thought Cozmo was neither too cooperative nor too competitive. A two-sample t-test further validated that both groups' estimates of Cozmo's cooperation rates did not significantly differ ($M_{high\text{-}K}$ = 49.39, $M_{low\text{-}K}$ = 49.429, *t*(60.3) = -0.007, *p* = .994). This was in line with our manipulation of Cozmo's cooperation rate – 50% in each game – which was set to control its behavioural competitiveness.

Regarding the open-ended question of whether Cozmo adopted any strategy in games, 80% (56 out of 70) participants said yes: 24 participants indicated that Cozmo was reciprocal or responsive to their decisions in games; 18 participants thought Cozmo adopted intentional strategies, such as being cooperative at first to gain participants' trust and then betraying them to win the most coins, or mostly sharing so both players could win the maximum coins. The subjective evaluation of Cozmo's game strategy varied tremendously among participants, but generally showed that participants attributed considerable intelligence and agency to Cozmo, which was not grounded in the reality of Cozmo's programming/behaviour.

## 2.5. Discussion

In the current study, we examined whether people's willingness to cooperate with a social robot is impacted by different incentive structures of prisoner's dilemma games, as has been shown to be the case in when these types of games are played between human competitors (Moisan et al., 2018). We developed a computer-mediated human-robot PD game and examined the frequencies of participants sharing coins (cooperating) with a Cozmo robot in high and low K-index conditions. We hypothesized that people in the high K-index condition (when cooperation is a relatively more rewarding choice) would share coins more often. Our findings suggest that the game's incentive structure did not exert any general influence on people's cooperative decisions across 20 rounds of gameplay. Instead, only in initial game rounds, participants in the high K-index condition cooperated significantly more than those in the low K-index condition. This unexpected result highlights the differential responses people make to embodied robots compared to the screen-mediated human agents in Moisan et al.'s (2018) study. However, the quick decay of cooperation rates and people's reciprocal tendencies were consistent with prior evidence from interpersonal economic games showing that people are less likely to cooperate or make public contributions after experiencing others' uncooperativeness (Gunnthorsdottir et al., 2007; Houser & Kurzban, 2002). Future studies will need to replicate the current findings and further explore the extent to which the gradually diminishing effect of incentive structures is a unique phenomenon to embodied HRI.

Exploratory analyses revealed two other influential factors underpinning participants' cooperative decision making. First, people showed a strong tendency to respond reciprocally toward Cozmo - a tit-for-tat strategy - regardless of the game condition they were assigned to. Reciprocity is regarded as a fundamental feature of human social behaviours (Chaudhuri et al., 2002; Fehr et al., 2002; Gintis, 2000) and has also been reported in studies examining interactions between humans and robots (Kahn et al., 2004; S. A. Lee & Liang, 2016; Sandoval et al., 2016). In our experiment, not only did participants react reciprocally toward Cozmo, but they also regarded Cozmo as behaving reciprocally toward them, while in reality, Cozmo carried out randomly ordered decisions. This observation ties in to the three factor theory of anthropomorphism proposed by Epley et al (Epley et al., 2007). According to this theory, when people have limited understanding about an agent, and when they are motivated to interact effectively with an agent to clarify a situation, they are more inclined to anthropomorphize the agent and to apply rules for interacting with other humans. This account fits our experimental context well, where players did not have extensive prior experience with robots in general, or the Cozmo robot specifically, and were attempting to anticipate Cozmo's next decisions in order to win a bigger payoff. It is thus understandable that participants tended to overinterpret cues from Cozmo's action and regard them as meaningful and intentional.

Additionally, our findings show that score presentation significantly affected participants' game decisions, especially for the presentation of the robot opponent's scores. Overall, participants were less likely to share coins when Cozmo's scores were high. However, such impact was more intricately shaped by participants' own scores (**Figure 2-7**). Participants behaved prosocially toward the robot (i.e., were more willing to share coins) only when they had personally achieved high scores. This seemingly counter-intuitive benevolent behaviour might be explained by two possible scenarios: first, participants were motivated to win more coins to beat other (human) participants'(and not Cozmo's) game records to win a shopping voucher, which means their chance of winning a prize did not have a direct relationship with the relative performance against Cozmo. This consequently allowed for the possibility of a win-win situation, in which participants were satisfied with their coin earnings, and could also help Cozmo escape punishment (i.e., by not having its data wiped) after losing. Second, feeling powerful and competent can increase individuals' sense of control and empathy toward others, which further leads people to engage in more prosocial behaviours and activities (Bhargava & Chakravarti, 2009; Côté et al., 2011; Magee & Langner, 2008). Our participants generally displayed a prosocial temperament, as evidenced by their SVO scores, which might have led them to act prosocially toward Cozmo as long as their self-interests were fulfilled. This point is also supported by the self-reported data participants gave when asked to identify the strategies used in games *(e.g., "I tried to keep 10 coins advantage. When I had 20 coins more than the robot, I shared.", "I aimed to have a certain gain by going for safe decisions (keeping coins for myself), accumulating some wealth, and only then I felt comfortable to take the risk of cooperating.")*.

Nevertheless, an alternative explanation could be that the interaction between Cozmo's and participants' scores on cooperation tendency was an outcome of participants' reciprocal behaviours in games. Specifically, we observed that participants, when earning low scores, were less likely to cooperate with Cozmo, and especially when Cozmo's score was much higher. This was likely the case because participants perceived that Cozmo had taken advantage of them (i.e., participants cooperated while Cozmo defected) previously for multiple times. It is thus conceivable that people would be unwilling to cooperate after the robot gained high scores by being uncooperative toward them. On the other hand, we found that participants, when already earning high scores, were more likely to cooperate with Cozmo, and this effect was even more pronounced when Cozmo's scores were also high. This could be explained by previous mutual cooperation and therefore mutual benefit (in terms of score). After such win-win cooperative experiences, participants would presumably keep cooperating and reciprocate Cozmo's prior cooperation. Granted, in this study we are not able to provide a decisive answer as the underlying social and psychological motives underpinning participants' game play decisions. Nevertheless, our study advances our understanding of human-robot cooperation, as well as human social behaviour in general, by providing several factors for researchers to consider when using economic games to exploring human–robot cooperation, including incentive structures, reciprocity, and the presentation of game status. Researchers should be aware of the impact of incentive structure when interpreting the results in one-shot PD games and when comparing human–robot cooperation rates between different game designs. Our findings also highlight how personal factors— such as predisposition to anthropomorphism—influenced human behaviours during HRI, and demonstrate the power of mixed effects model to control such subject-level random effects.

However, our findings also raise several questions and limitations for future research to address. First, although the vignette of erasure of Cozmo's memory (adapted from Seo et al.'s study) was found effective in convincing participants of the real and meaningful consequences happening to Cozmo if it lost games (as evidenced by the self-reported data). We acknowledge the possible confounding impact caused by individuals' empathetic responses and therefore adopted mixed effects models to better control for possible subject-level random effects. Future studies could use more structured quantitative measures to assess how meaningful each participant thinks an economic game is to a robot or any other non-human agent, to ensure the validity of this kind of paradigm. For example, researchers could manipulate (e.g., increase or decrease) the extent of punishment and rewards a robot receives during human–robot PD games, and measure how these manipulations impact participants' perceptions and cooperative willingness.

Secondly, previous work has highlighted the risks of generalising findings from one robotic platform to HRI overall (Henschel et al., 2020; Hortensius et al., 2018; Hortensius & Cross, 2018), underscoring the need to clarify the extent to which different robot manifestations (in terms of size, function, sophistication, human-likeness, etc) influence human cooperation. The Cozmo robot we used in the study is rather toy-like, and small in size. Future research will need to replicate this work with larger (or even human-sized) robots if such findings are to be generalised to real-life situations where the robots we cooperate with and look to for assistance and companionship might be larger, more advanced, and more capable of assisting people in daily life scenarios. Another aspect of generalisability concern is related to the sample diversity. Although our sample was comprised of 24 different nationalities, a majority of participants came from a western cultural background. Future research could investigate human–robot cooperation in more diverse cultural contexts as it is important to take cultural influences into consideration when designing and studying HRI (Lim et al., 2020).

Thirdly, we did not directly compare here cooperation with a robot to cooperation and with a human confederate, but instead borrowed the insights from human-human interaction to predict human behaviours in HRI. The main aim of our study was to investigate the impact of situational incentives on human-robot cooperation, rather than to examine possible differential responses to robot and human competitors in economic games. However, future studies might wish to include a human confederate as well, to examine in more detail the extent to which the effects of incentive structures depend on the agents that people interact with. Finally, in the current study we only examined the difference between K-indices of 0.6 and of 0.2. Future research could include more levels of K-indices to acquire a fuller understanding of how our willingness to cooperate with a robot changes according to different incentive structures of human–robot PD games.

To conclude, our findings show that the incentive structure of a human–robot PD game influenced human cooperation only at the beginning of the game. Throughout the whole game, participants' cooperative/non-cooperative decisions were driven more by the robot's decision (following a tit-for-tat strategy) and by the presentation of game scores in each round.

**Acknowledgements:** We thank Kohinoor Darda, Ruud Hortensius, Laura Jastrzab, and Richard Ramsey for thoughtful feedback on earlier manuscript versions and the ERC (Grant agreement number 677270 to ESC) and Leverhulme Trust (PLP-2018-152 to ESC) for funding.

# Chapter 3 The Role of Empathic Traits in Emotion Recognition and Emotion Contagion of Cozmo Robots

This chapter is an exact copy of the manuscript submitted as a Late-Breaking Report to the 2022 ACM/IEEE International Conference on Human-Robot Interaction (under review):

Hsieh, TY. & Cross, E. S. (2021) The Role of Empathic Traits in Emotion Recognition and Emotion Contagion of Cozmo Robots. Manuscript submitted for publication in *HRI '22: Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction.*

## 3.1. Abstract

In this online study, we investigated how well people could recognise emotions displayed by video recordings of a Cozmo robot, and the extent to which emotion recognition is shaped by individuals' empathic traits. We also explored whether participants who report more empathic tendencies experienced more emotional contagion when watching Cozmo's emotional displays, since emotion contagion is a core aspect of empathy. We tested participants' perceptions of Cozmo's happiness, anger, sadness, surprise, and neutral displays. Across 103 participants, we report high recognition rates for most emotion categories except neutral animations. Furthermore, the mixed effects modelling revealed that an empathy subtype (the empathic concern subscale from the Interpersonal Reactivity Index) significantly impacted emotional contagion. Contrary to predictions, participants with high empathic concern subscale scores were *less* likely to find the robot's videos emotionally contagious. The study validates the utility of Cozmo robots to display recognisable emotional cues, and further suggests that empathic traits could shape our affective interactions with robots, though perhaps in a counterintuitive way.

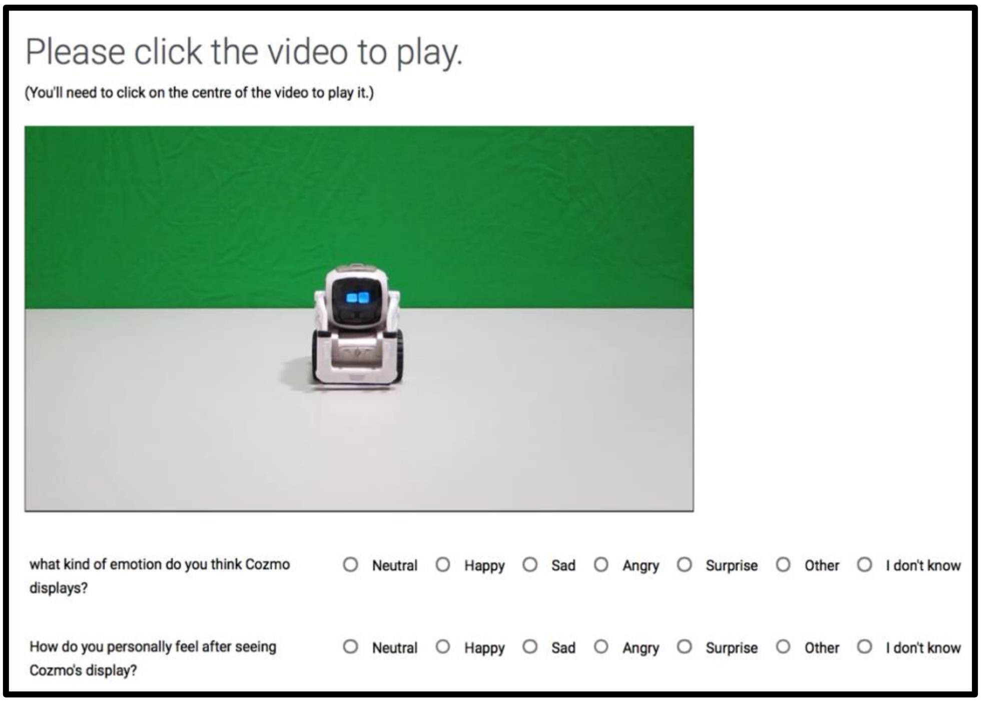
**Keywords**— Human-robot interaction, Dispositional empathy, Emotion recognition, Emotion contagion

## 3.2. Introduction

Accurate recognition of others' emotional cues is a crucial factor that contributes to effective and smooth interpersonal interactions (Barrett et al., 2019; Hess et al., 2016; Van Kleef, 2009). Similarly in human—robot interaction (HRI), the capacities for social robots to display appropriate and recognisable emotion cues can be conducive for forming meaningful and socially sophisticated relationships with users (Hortensius et al., 2018; Tsiourti et al., 2019). On the other hand, emotion recognition abilities for people to recognise robots' emotional cues might differ by individuals, by robotic platforms, and by emotion types (Stock-Homburg, 2021). The current psychology literature has well documented the individual differences in recognising human facial expressions (Barrett et al., 2019; Besel & Yuille, 2010). In particular, individual differences in empathic traits have been linked with differential performances in emotion recognition. For example, empathic people have been found to perform better in facial expression recognition tasks (Konrath et al., 2014);  emotional empathy (i.e., the ability to feel the emotions others experience) is related to better recognition of facial expressions within a short period of time (Besel & Yuille, 2010); and people with autism spectrum conditions who have difficulties with emotion recognition tasks also record low scores in self-report empathy scales (Martin et al., 2019; Sucksmith et al., 2013). Given the relationship between empathic traits and people's recognition abilities for human emotional expressions, it is important to examine whether similar links exist between dispositional empathy and accurate recognition of robots' emotional displays. Current evidence has suggested that people could correctly recognise about 50% to 60% of embodied robots' emotional displays (based on 43 HRI studies reviewed in Stock-Homburg, 2021), the research here could help explain the individual differences in emotion recognition of robots and set the foundations of bespoke social robots based on users' personality traits.

Empathy, as a multidimensional construct, refers to not only a person's ability to cognitively understand others' perspectives, but also the tendencies of being affectively connected to another person's inner experience (Davis, 1983b). The affective component of empathy is therefore associated with emotional contagion, which is a phenomenon that occurs when we automatically synchronize our own emotional states with others' (Hatfield et al., 1993; Prochazkova & Kret, 2017). Previous studies have found that highly empathic people are more likely to experience emotions from non-human targets like art (Stavrova & Meckel, 2017) and music (Vuoskoski & Eerola, 2012). It is therefore of interest to determine whether people's baseline empathic tendencies might also make them more likely to experience vicarious feelings from robots' emotional displays.

In this study, we used Cozmo entertainment robots (manufactured by Anki Inc., **Figure 3-1**) as the robotic platform to display emotional expressions. Cozmo robots' affordability, portability and flexibility to be programmed have made them suitable tools for HRI research (Chaudhury et al., 2020; Cross, Riddoch, et al., 2019). Consequently, a better understanding of people's emotion recognition of Cozmo's simple emotional displays stands to benefit future studies aiming to investigate the display of human readable emotions by embodied robots. Additionally, the current research could help bridge the gap between psychology and HRI research by raising awareness of a personal factor – empathic traits – in social and affective interactions with robots. Specifically, based on human psychological evidence (Besel & Yuille, 2010; Konrath et al., 2014; Stavrova & Meckel, 2017; Vuoskoski & Eerola, 2012), we predicted that people who reported high empathic traits could more accurately recognise Cozmo's emotional displays, and would be more likely to feel the vicarious feelings from the robot's emotional expressions.



Please click the video to play.

(You'll need to click on the centre of the video to play it.)

what kind of emotion do you think Cozmo displays? ○ Neutral ○ Happy ○ Sad ○ Angry ○ Surprise ○ Other ○ I don't know

How do you personally feel after seeing Cozmo's display? ○ Neutral ○ Happy ○ Sad ○ Angry ○ Surprise ○ Other ○ I don't know

**Figure 3-1.** Screenshot of the online emotion rating task. Participants would report the emotion(s) they recognised from a short video of Cozmo's emotional displays and also the feeling(s) they experienced after watching the video.

## 3.3. Methods

We devised an online experiment via formR (Arslan et al., 2020) to explore the relationships between people's empathic traits and emotion recognition and emotion contagion of the Cozmo robot's emotional displays. The online experiment involved three sections: (1) participants filled out the Interpersonal Reactivity Index (IRI) (Davis, 1983a) as a measure of their empathic traits; (2) they watched and rated a series of videos showing Cozmo's different emotional displays (each approximately 10 seconds long) (**Figure 3-1**); (3) they answered demographic questions of their age and gender. The details of the first two sections are explained below.

### *3.3.1. Empathy Measures*

The Interpersonal Reactivity Index (IRI) is a widely used empathic trait measure from the psychological literature (Besel & Yuille, 2010; Davis, 1983a). The IRI involves four subscales: *perspective taking, fantasy, empathic concern,* and *personal distress*. *Perspective taking (PT)* focuses on the cognitive component of empathy, which is the readiness to see things from others' points of view. *Fantasy (FT)* scale measures whether people tend to imagine themselves as characters in novels or movies, and how easily they become emotionally engage with fictional characters. *Empathic concern (EC)*, ascribed to emotional aspect of empathy, is about how often people experience feelings of others' sufferings. Lastly, *personal distress (PD)* subscale assesses whether observing others' misfortunes usually results in their own anguish. Each subscale contains seven items and items are rated on a five-point Likert scale from 0 (does not describe me well) to 4 (describes me very well).

### *3.3.2. Cozmo Emotion Rating Task*

Three experimenters watched all the 348 Cozmo animations from the Github repository – https://github.com/cozmo4hri/animations (Chaudhury et al., 2020)– and selected five emotion categories that were most salient. The final set of videos displayed happy (animation numbers: 92, 94, 100, 193), angry (73, 74, 136, 137), sad (63, 134, 152, 190), surprising (24, 65, 91, 200), and neutral (25, 99, 160, 208) emotions. In the emotion rating session (**Figure 3-1**), participants watched a video of Cozmo displaying a specific emotion type (around 10 seconds long) and answered what they recognised from the video: "neutral", "happy", "sad", "angry", "surprise", "other" (with a text space for more details), or "I don't know". Furthermore, participants also reported their subjective feeling(s) after watching each video, using the same options provided. The first question was a measure of participants' emotion recognition accuracy (i.e., that the emotion a person recognised from a Cozmo's video was in line with the emotion the experimenters intended the robot to display). The second question about their personal feelings was to know whether participants' emotional states were influenced by Cozmo's emotional displays (emotion contagion). Participants rated a total of 20 videos (four videos for each category and five emotion categories) and video order was randomized across participants.

## 3.4. Results

We report the relevant research materials, anonymous data, and analysis codes on the Open Science Framework (OSF) project page - https://osf.io/p49jv/?view_only=3148c25ace084c6db5d2760778a2d8b9 - following open science initiatives (Munafò, 2016). All analyses were done with R v4.0.1 (R Core Team, 2020). In total, one hundred and three valid samples (average age = 32.3 years old; 43 females, 57 males, one non-binary, and 2 preferring not to report) were collected for the online experiment.

### *3.4.1. Emotion Recognition and Subjective Feelings for Cozmo's emotional displays*

We calculated the recognition rates of Cozmo's five emotions and the report rates of subjective feelings after watching the robot's videos (**Figure 3-2**). The emotion type most accurately and consistently recognised by participants was Cozmo's anger (mean recognition rate = 78.40%), followed by Cozmo's sadness (recognition rate = 69.18%), happiness (recognition rate = 62.38%), and surprise (recognition rate = 63.35%). For neutral animations, participants' recognition was less in consensus. On average, only 19.42% of participants perceived the neutral videos as neutral. 18.2% of them reported "I don't know" and 17.48% of participants classified them as "happy".

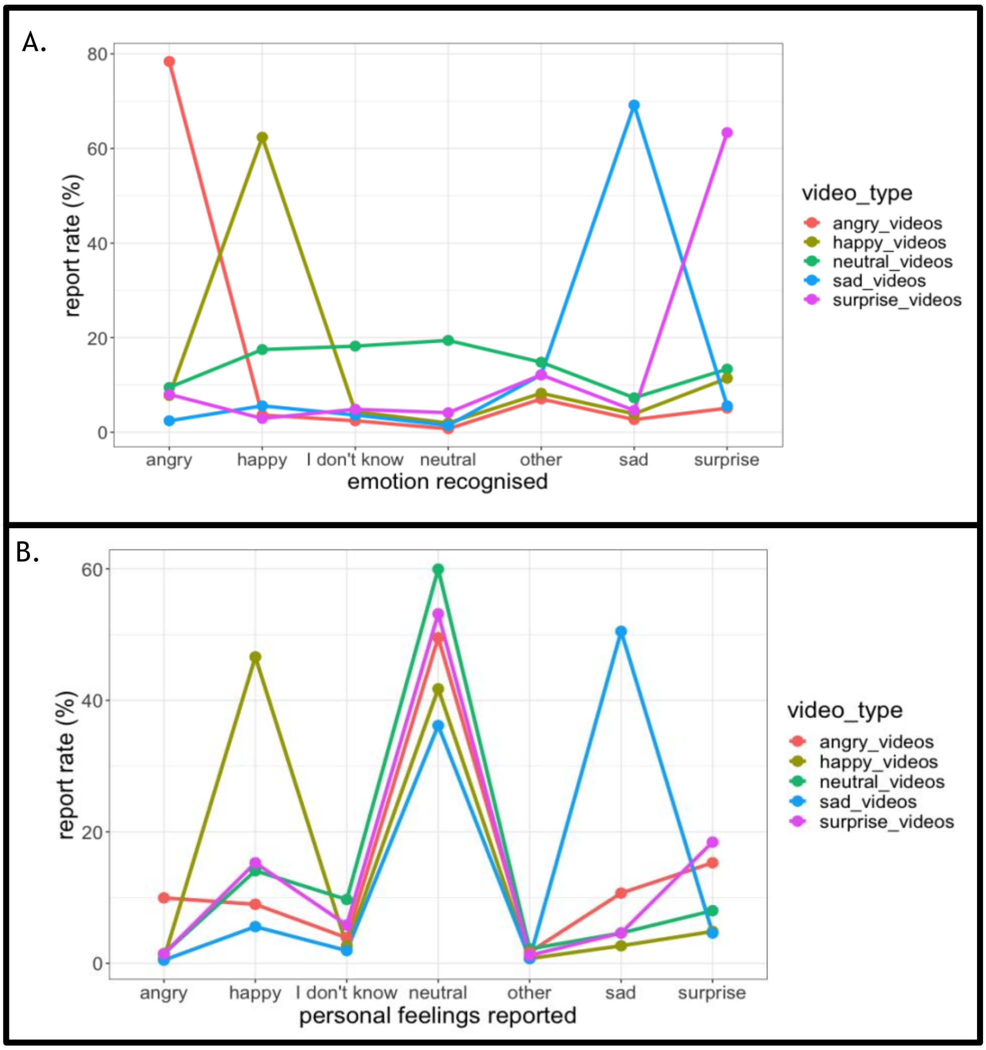


**Figure 3-2.** (A) Recognition rates of Cozmo's emotional displays. (B) Participants' report rates of their subjective feelings after watching the robot's animation videos.

As for participants' subjective feelings after watching Cozmo's different emotional displays, happy and sad animations were the emotion categories that showed stronger effects of emotional contagion. 46.60% of participants felt happy after the robot's happy displays and 50.49% of them felt sad after the robot's sad displays. For angry, surprising, and neutral videos, participants mostly felt neutral after watching them: 49.52% of them felt neutral after the robot's angry displays (compared to only 9.95% of them feeling angry); 53.16% of them felt neutral after surprising displays (compared to 18.45% of them feeling surprised); 59.95% of them felt neutral after watching neutral displays.

### *3.4.2. Dispositional Empathy and Emotion Recognition of Cozmo*

We calculated the mean scores of participants' IRI reports ($M$ = 2.55, $SD$ = 0.43; on a 5-point Likert scale from 0 to 4) and the means of their IRI subscale scores (perspective taking: $M$ = 2.77, $SD$ = 0.7; fantasy: $M$ = 2.66, $SD$ = 0.75; empathic concern: $M$ = 2.98, $SD$ = 0.66; personal distress: $M$ = 1.77, $SD$ = 0.88). Reliability analysis revealed that Cronbach's alpha for IRI is .76. We then analysed the correlations between IRI scores and emotion recognition rates (**Figure 3-3**). None of the Pearson's correlation coefficients between variables was significant. Overall, the relationship between emotion recognition of all emotions and IRI scores was $r$ = -0.14, $p$ = .150.

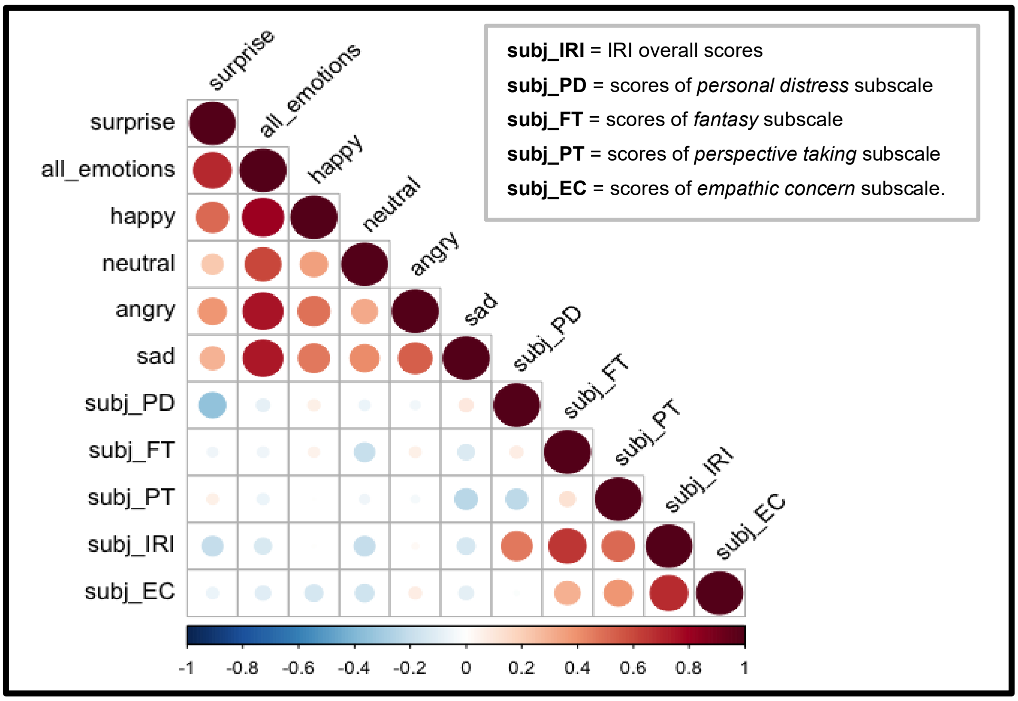


**Figure 3-3.** Correlations between emotion recognition and IRI scores. Redder and bigger dots represent stronger positive correlations, and the bluer and bigger dots show stronger negative correlations.

## *3.4.3. The Influence of Dispositional Empathy on Emotion Recognition and Emotion Contagion*

### The Influence of Dispositional Empathy on Emotion Recognition

We ran a generalised linear mixed effects model with the lme4 package (v1.1.23) (Bates et al., 2015) to examine the impact of empathic traits on participants' trial-by-trial emotion recognition (correctly recognising an emotional display was coded as 1; incorrectly recognising a display was 0). In the model, we had IRI overall scores as the fixed factor, emotion recognition accuracy as the binary dependent variable, and controlled subject-level and trial-level random effects. In the result, the effect of empathic traits was non-significant on trial-by-trial recognition, *$\beta$* = -0.46, *95% CI* [-1.02, 0.11], *p* = .116. Considering previous evidence showing that empathy subtypes could differentially impact recognition of human facial expressions (Besel & Yuille, 2010), we conducted another generalised linear mixed effects model with the four IRI subscales (*PT, PD, FT, EC*) as fixed factors while the rest of model design remained the same. The results showed that none of the subscales significantly impacted emotion recognition: *perspective taking (PT) — $\beta$* = -0.10, *95% CI* [-0.47, 0.27], *p* = .604; *personal distress (PD) — $\beta$* = -0.14, *95% CI* [-0.42, 0.14], *p* = .315; *fantasy (FT) — $\beta$* = -0.04, *95% CI* [-0.38, 0.29], *p* = .795; *empathic concern (EC) — $\beta$* = -0.17, *95% CI* [-0.58, 0.23], *p* = .402.

### The Influence of Dispositional Empathy on Emotion Contagion.

To investigate the influence of empathic traits on emotion contagion of Cozmo's expressions, we conducted a generalised linear mixed effects model, with IRI scores as the fixed factor, emotion contagion as the binary dependent variable (if what they felt was the same as what they recognised from the videos, it was coded as 1; otherwise it was 0). We controlled subject-level and trial-level random intercepts. We did not find a significant effect from subjects' IRI overall scores, *$\beta$* = -0.10, *95% CI* [-0.65, 0.44], *p* = .711. Again, we explored whether the four IRI subscales had differential influences on emotion contagion, and ran another model with the four subscales as the fixed factors while keeping the rest of the model design the same. We found a significant effect of *empathic concern*

*(EC)* subscale ($\beta$ = -0.40, *95% CI* [-0.78, -0.01], *p* = .042), but the other three subscales were non-significant (*PT*: $\beta$ = 0.009, *95% CI* [-0.34, 0.36], *p* = .957; *PD*: $\beta$ = 0.02, *95% CI* [-0.25, 0.28], *p* = .909; *FT*: $\beta$ = 0.26, *95% CI* [-0.06, 0.58], *p* = .113). The effects of the four IRI subscales were visualised with the R package "effects" (v4.1.4) (Fox, 2003) in **Figure 3-4**.

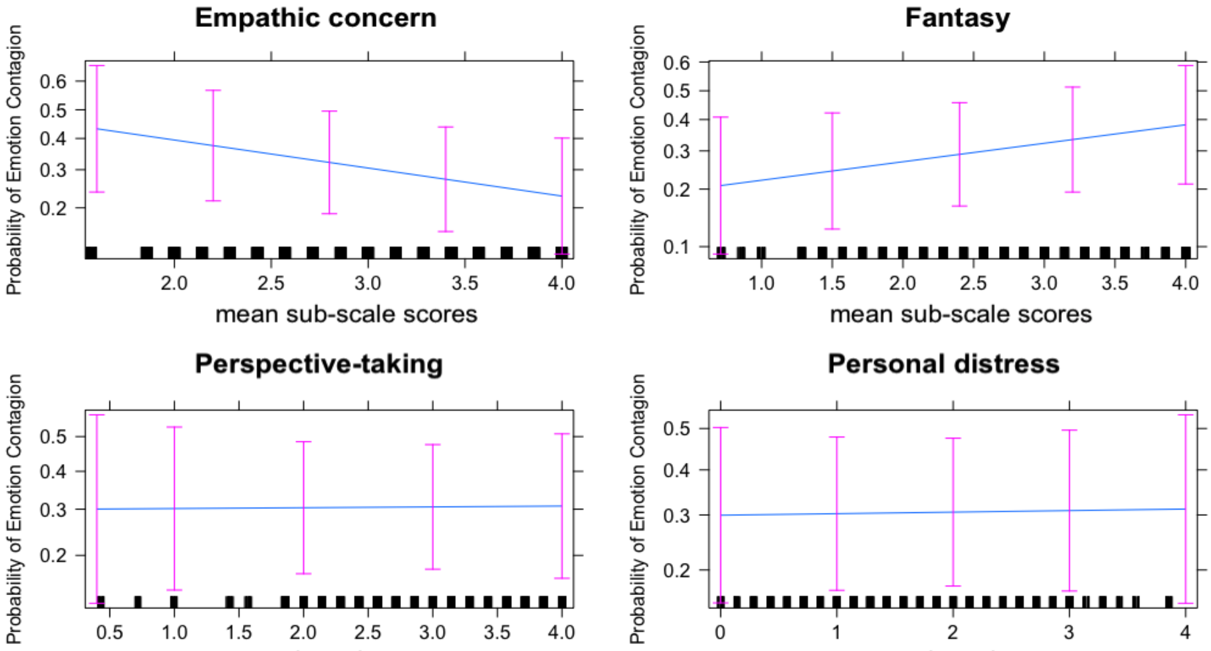


**Figure. 3-4.** The effects of the four empathy subtypes (IRI subscales) on emotion contagion participants experienced after watching the Cozmo's emotional expressions in videos. Only the "empathic concern" subscale was found to significantly predict emotion contagion. The items of these subscales were all rated on a five-point Likert scale from 0 (does not describe me well) to 4 (describes me very well). The effect plot was generated with the R package "effects" (v4.1.4)

## 3.5. Discussion

Here we designed an online experiment to investigate people's emotion recognition of a Cozmo robot's emotional expressions and whether such emotion recognition is shaped by individuals' dispositional empathic traits (measured by the IRI (Davis, 1983a). We also explored the extent to which participants' affective states might synchronize with the robot after watching the robot's emotional displays, which is also known as emotion contagion — an important aspect of empathy. We expected empathic participants to be more accurate in recognising the robot's emotional displays and also to report the displays more emotionally contagious. Below we discuss each part of our findings in detail.

First, the emotions participants recognised from Cozmo's videos were generally in line with the experimenters' predictions, except for the neutral videos. Contrary to human emotion recognition evidence suggesting that happiness is the most easily recognised emotion (Montagne et al., 2007) and usually shows high agreement rates among testing samples (Barrett et al., 2019), our results show that participants most consistently recognised Cozmo's anger. Moreover, as we compared the current emotion recognition rates with the mean recognition rates of 43 previous HRI studies reviewed in Stock-Homburg's paper (Stock-Homburg, 2021), we found that Cozmo's anger (recognition rate = 78.40%), sadness (recognition rate = 69.18%), and happiness (recognition rate = 62.38%) performed better than the literature's average recognition rates of robotic emotions displayed by both facial and bodily expressions (anger: 56.77%; sadness: 55.95%; happiness: 62.09%; Stock-Homburg, 2021). However, Cozmo's surprise (recognition rate = 63.35%) performed worse than the average of the literature (76.08%). The findings validate that, even in the context of online experiment, Cozmo is capable of displaying perceivable and recognisable emotion animations. It is also worth noting that participants recognised various different emotions – such as happiness, surprise, curiosity, fear – from the videos we regarded as neutral. The diverse responses we received for the neutral stimuli point to Kuleshov effect, which proposes that people evaluate the emotion of a neutral face by contextual cues (such as the emotional stimuli preceding the face) (Barratt et al., 2016; Mobbs et al., 2006). Consequently, researchers who wish to manipulate a robot to be neutral in expression (e.g., in a control condition) should be aware of the potential Kuleshov effect, especially in online experiments where we have less control over participants' environments.

Second, we explored the influence of empathic traits on the emotion recognition and emotion contagion effects of Cozmo by mixed effects models. None of the empathy variables – neither the overall IRI scores nor the scores of IRI subscales – significantly impacted recognition of Cozmo's emotional displays. However, when we looked into the relationship between empathic traits and emotion contagion, we found a significant effect from the *empathic concern (EC)* subscale of IRI. Surprisingly, people who scored higher in the EC subscale were *less* likely to report the same feeling as what they had just recognised from

Cozmo's display. Although the result confirms that empathy subtypes could have unique links with emotional mental processes like emotion recognition (Besel & Yuille, 2010) and facial mimicry (Perugia et al., 2020), we urge replication of this finding before attempting to explain why the relationship was opposite to our prediction. It is worth reiterating that IRI is a scale to measure individuals' empathy toward other people (not robots), and therefore it might not be a suitable or precise measure for this research question. Further research is needed to clarify this and to gain insights into the mechanism(s) underpinning emotion contagion effects of robots and influence of personal empathy traits. Future research could also deploy an embodied Cozmo robot to investigate emotion recognition embodied emotion displays, since physical embodiment crucially shapes real-life HRI (Grossman et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018). Also, researchers could include additional emotion categories for Cozmo to display, to test whether it is able to display an even more diverse range of emotional cues.

**Acknowledgment:** We thank Dominic Munro and Nadine Mekari for stimuli selection.

# Chapter 4 People’s dispositional cooperative tendencies towards robots are unaffected by robots’ negative emotional displays in prisoner’s dilemma games

The chapter is an exact copy of the registered report published in *Cognition and Emotion*:

Te-Yi Hsieh & Emily S. Cross (2022): People’s dispositional cooperative tendencies towards robots are unaffected by robots’ negative emotional displays in prisoner’s dilemma games, *Cognition and Emotion*, DOI: 10.1080/02699931.2022.2054781

## 4.1. Abstract

The study explores the impact of robots’ emotional displays on people’s tendency to cooperate with a robot opponent in prisoner’s dilemma games. Participants played iterated prisoner’s dilemma games with a non-expressive robot (as a measure of cooperative baseline), followed by an angry, and a sad robot, in turn. Based on the Emotion as Social Information model, we expected participants with higher cooperative predispositions to cooperate less when a robot displayed anger, and cooperate more when the robot displayed sadness. Contrarily, according to this model, participants with lower cooperative predispositions should cooperate more with an angry robot and less with a sad robot. The results of 60 participants failed to support the predictions. Only the participants’ cooperative predispositions significantly predicted their cooperative tendencies during gameplay. Participants who cooperated more in the baseline measure also cooperated more with the robots displaying sadness and anger. In exploratory analyses, we found that participants who accurately recognised the robots’ sad and angry displays tended to cooperate less with them overall. The study highlights the impact of personal factors in human-robot cooperation, and how these factors might surpass the influence of bottom-up emotional displays by the robots in the present experimental scenario.

**KEYWORDS**

EASI model, social robotics, Human-robot interaction, prisoner's dilemma games, social decision making

## 4.2. Introduction

Social robots are becoming increasingly valuable tools for assisting people in industrial, educational, and health care settings (Broadbent, 2017b; Dautenhahn, 2007). The COVID-19 pandemic has further highlighted the potential utility for robots in replacing human labour to reduce the risk of infection, but also for their social abilities, such as helping to alleviate loneliness during lockdown (Kim et al., 2021; Odekerken-Schröder et al., 2020; Yang et al., 2020). As the world is likely to embrace a “new normal” after COVID-19, including remote education, increased working from home culture, and more autonomous industry (Cahapay, 2020; Jamaludin et al., 2020), the necessity of welcoming social robots into our lives is becoming even clearer. It is consequently imperative to gain deeper understanding of the factors shaping people’s willingness to work with robots in their households and workplaces, and how best to promote the social and cooperative behaviours during human-robot interaction (HRI).

Previous research has used economic games as an analogy of real-life social decision-making settings to investigate human cooperative behaviours (Bland et al., 2017; Chaudhuri et al., 2002; Rand & Nowak, 2013; Rapoport & Chammah, 1967). By manipulating the payoffs rewarded to participants after making a decision (for example, to cooperate or not), researchers can test the boundaries of people’s willingness to cooperate across various settings, and more importantly, examine the factors that induce cooperative behaviours (Bland et al., 2017; Pothos et al., 2011a; Rapoport, 1967). One pivotal factor that affects our decision-making process is the extent to which, and how, others display emotion  (George & Dane, 2016; Lerner et al., 2015; Rick & Loewenstein, 2008; Van Kleef, 2009). As social animals, we use other people’s emotions to make sense of current situations; thus, our decision-making is susceptible to influence by others’ emotional expressions (Darwin & Prodger, 1998; Kjell & Thompson, 2013; Moors et al., 2013). Our sensitivity to emotion displays is so pronounced that even if an agent that displays emotions is artificial by nature (e.g., an animated avatar or a manufactured robot), research evidence is accumulating to suggest such emotional displays are similarly influential in shaping people’s social decisions (de Melo et al., 2010; de Melo, Gratch, et al., 2014a; Kayukawa et al., 2017; Terada & Takeuchi, 2017).

However, most of this evidence informing our knowledge of cooperative behaviours in HRI comes from online studies (for example, de Melo et al., 2010, 2014, 2019; Hoegen et al., 2018). While online research provides a useful point of departure for understanding people's cooperative tendencies, physically embodied interaction is a key feature of real-life HRI and it can be regarded as a distinct scenario from screen-mediated interaction (Grossman et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018; S. A. Lee & Liang, 2016; Wykowska et al., 2016). For example, people gave more positive evaluation and showed more empathy towards an embodied robot than a disembodied one (Kwak et al., 2013; K. M. Lee et al., 2006) In order to clarify the psychological mechanisms supporting human–robot cooperation, the present study focused on the impact of robots' emotional expressions and displays on people's tendency to cooperate with a robot opponent in prisoner's dilemma games. A clearer understanding of the role a robot's emotion display plays on human cooperative behaviour can bring crucial insight into the design of social robots that can be effectively deployed as assistants in our society across several settings (e.g., education, healthcare and workplace support).

### *4.2.1. The social functions of emotions in human psychology literature*

Emotional expressions are prominent social cues that influence decision making during interpersonal interactions (George & Dane, 2016; Lerner et al., 2015; Rick & Loewenstein, 2008; Van Kleef, 2009). Others' emotions offer useful information for us to infer their feelings, intentions, and desire, and help us reason about the current situation (Frijda, 1986; Moors et al., 2013; Roseman & Smith, 2001). Furthermore, others' emotions often have context-dependent meaning, and impact on our own behaviours, as claimed in the Emotion as Social Information (EASI) model (Van Kleef et al., 2010). In competitive situations, people have been shown to make strategic and epistemic judgements in response to opponents' emotions. For instance, people are more likely to concede to angry emotion displays (to avoid destructive dispute), while they might either become irresponsive to or seize the chance to exploit sad opponents. Conversely, in cooperative settings, the EASI model proposes that humans prioritise social harmony over strategy, and thus seeing others' angry displays, which erodes the cooperative atmosphere, makes us less willing to cooperate with those who act or express angrily. However, observing another express sadness evokes empathy and promotes cooperative and supportive behaviours within a group (Van Kleef et al., 2010).

In the present study, we focused on the impact of robots' displays of anger and sadness. In contrast to positive emotions, which imply fulfilment and satisfaction, negative emotions often connote a goal unfulfilled or dissatisfaction with an outcome (Frijda, 1986; Moors et al., 2013; Roseman & Smith, 2001; Van Kleef et al., 2010). This is precisely the crucial situation where social cues promoting cooperation are likely to be needed in real-life settings. In human psychology, researchers have attempted to validate the interpersonal impact of angry and sad displays by either online or in-person experiments. For example, using computer-mediated interactions, Van Kleef et al. (2004) found that people made more concessions to the negotiator who sent an angry message about the offer (e.g., "This offer makes me really angry,"), in comparison to the negotiator who sent a happy message about the offer (e.g., "I am happy with this offer"). In another more interactive scenario, Kopelman et al. (2006) examined the impact of positive, negative, and neutral emotions in negotiation situations with two different approaches of emotional manipulation: first, coaching participants to express specific emotions in their negotiation dyads, and second, playing pre-recorded videotapes of a professional actor displaying the three types of emotions while giving a business offer. The researchers found that participants were more likely to make a business deal with negotiators with the positive manner than with the negative or neutral one. However, Kopelman et al. (2006) also acknowledged the limitations of such emotional manipulation that might be constrained by individuals' emotional expressivity (people feign negative emotions worse than positive emotions) and by the unnatural and artificial aspect of interacting with a videotaped person.

Given the difficulty in manipulating human emotions to examine the interpersonal impact of emotion displays on social decisions, evidence supporting the appraisal theory or EASI model was mainly derived from studies examining computer-mediated interactions (Van Dijk et al., 2008, 2018; Van Kleef et al., 2004, 2006) or interactions without rigorous control of the emotional stimuli (Kopelman et al., 2006). Fortunately, these limitations are greatly diminished in the context of HRI where robots can be programmed to perform identical behaviours, and can thus convey embodied emotional stimuli precisely for every participant and every trial.

### *4.2.2. Artificial agents' emotion displays in human–robot cooperation*

Considering the vital role of emotional expressions in our social life, an increasing number of artificial agents (robots and virtual agents) are being built to display human-readable emotions by facial or bodily expressions (Hortensius et al., 2018). Some researchers report that people behave similarly with artificial agents and with human agents in economic games (de Melo et al., 2010; Krach et al., 2008; Wu et al., 2016), and provided empirical findings on the utility of artificial agents' emotion displays to promote cooperative behaviours (de Melo et al., 2011; de Melo, Gratch, et al., 2014a; Terada & Takeuchi, 2017). For instance, in online gaming settings, manipulation of virtual agents' facial expressions (showing joy after mutual cooperation and guilt after making a selfish decision) according to the appraisal theory of emotion have been proved effective in eliciting people's cooperative behaviours in economic games with artificial agents (de Melo et al., 2010, 2011; de Melo, Gratch, et al., 2014a). The social functions of agents' facial expressions were not only found by highly human-like virtual agents. Terada and Takeuchi (2017) have demonstrated that emotions displayed by an embodied robot's simple line drawing face (showing on its monitor head) could induce people's altruistic behaviours in ultimatum games. However, when emotions were displayed merely by modalities like bodily movements and verbal expressions (rather than by facial expressions) the emotional impact on cooperative behaviours was less clear. Kayukawa and colleagues (2017) applied de Melo et al.'s (2010) emotional manipulation to an embodied Nao robot (manufactured by SoftBank Robotics) but found that the Nao being programmed to induce cooperation via different emotional responses (i.e., displaying joy after mutual cooperation, anger after being betrayed, shame after betraying, and sadness in a lose-lose situation) did not bring about more cooperative behaviours among participants in prisoner's dilemma games (which the authors suspect could also be due to the limited sample size of 14 subjects). Nevertheless, the participants did regard the emotional Nao robot as more friendly and cheerful than the non-expressive Nao (Kayukawa et al., 2017).

In addition to manipulating artificial agents' emotion displays based on emotion theories, Hoegen et al. (2018) programmed virtual human characters to mimic participants' facial expressions during prisoner's dilemma games and found a correlation between perceived rapport and cooperation rates only when interacting with the agent mimicking. All in all, according to the literature reviewed above, legitimate emotion displays (either based on psychological emotion theories or in congruence with people's own emotional states) by virtual humans appears to be at least somewhat effective in shaping people's cooperative decisions (de Melo et al., 2010, 2011; de Melo, Gratch, et al., 2014a; Hoegen et al., 2018). However, evidence from HRI is still not sufficient for us to decisively and reliably understand the relationship between embodied robots' emotion displays and people's cooperative behaviours. Furthermore, this topic warrants empirical examination now if we are to develop real-life robot assistants to appropriately serve people's social needs with apt and effective emotion displays. Our study therefore aimed to address this question through a study performed with the highly expressive Cozmo robots (detailed in Method) and to examine the impact of the robots' emotion displays on cooperative behaviours in the context of human-robot prisoner's dilemma games.

### *4.2.3. Prisoner's dilemma games*

To study human cooperative behaviours, the prisoner's dilemma (PD) game is one of the most widely used paradigms in research spanning the social sciences (Pothos et al., 2011a; Rapoport, 1967; Rapoport & Chammah, 1967). A classic PD game involves two people making simultaneous decisions to cooperate or to defect. Each player's payoff depends on both players' decisions, as illustrated in **Figure 4-1**. In the situation of mutual cooperation, both players are rewarded with a moderate amount of endowment (***R*** in Figure 1; £7 each, for example). Meanwhile, players might be tempted by the highest profit (***T***; e.g., £10) for being the only one who defects, and render the other who cooperates in the worse situation (***S***; e.g., £0). However, choosing to defect also comes with a risk. If both players opt to defect, they both receive punishment of little gain (***P***; e.g., £1).

|  | **Player 1 cooperates** | **Player 1 defects** |
|---|---|---|
| **Player 2 cooperates** | ***R*** (£7)<br>***R*** (£7) | ***T*** (£10)<br>***S*** (£0) |
| **Player 2 defects** | ***S*** (£0)<br>***T*** (£10) | ***P*** (£1)<br>***P*** (£1) |

**Figure 4-1.** An exemplified payoff matrix in prisoner's dilemma games. R = rewards; T = temptation; S = sucker's payoff; P = punishment. The dilemma is defined by two rules: T > R > P > S, and 2R > T + S. Adapted from Hsieh, TY., Chaudhury, B. & Cross, E. S. (2020). Human-robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots. PsyArXiv. https://psyarxiv.com/q6pv7/

In this scenario, a social dilemma happens when collective group profit is at odds with individual profit, and as a cooperative decision involves the risk of being exploited, and players have the freedom to choose between the two opposite actions to take. An extensive body of literature on interpersonal PD games has used both experiments and data simulation to model and theorise on the emergence and evolution of human cooperative behaviours (Axelrod & Hamilton, 1981; Embrey et al., 2018; Rapoport & Chammah, 1967). With mathematical modelling, more recent research has provided considerable insights into the mechanisms and factors supporting or hampering cooperation across various social dilemma situations (e.g., in dyads and in groups) (Bravo et al., 2012; Ito & Tanimoto, 2018; Kopp et al., 2018; Perc et al., 2017).  Also, from empirical evidence of interpersonal PD games, multiple factors are at play during people's decision-making process in the scenario, such as the trust in the other player (Chaudhuri et al., 2002; Janssen, 2008; Wu et al., 2016), their social value orientation (Pletzer et al., 2018), and perceived environmental cooperativeness/competitiveness (Elliot et al., 2018; Moisan et al., 2018). However, when it comes to PD games played with robots (let alone the Cozmo robotic platform specifically), our current understanding of people's decision-making process remains limited. Recent research on human-robot PD games has provided preliminarily insights into the impacts of reciprocity (Sandoval et al., 2016), trust (Paeng et al., 2016), dialogic verbal reactions (Maggioni & Rossignoli, 2021), and a Nao robot's emotion displays (Kayukawa et al., 2017) on HRI. Yet, the preliminary evidence raises more questions than answers at this stage, especially with respect to the effects of robots' emotion displays in PD games.

Meanwhile, researchers in HRI are becoming increasingly alert to generalisability concerns that empirical findings from research performed with a specific robotic platform might not necessarily apply to a different robot (Henschel et al., 2020; Hortensius et al., 2018; Hortensius & Cross, 2018). Therefore, in order to eliminate any confounding impact from robot-specific or context-specific factors (like people's trust and perceived agency towards Cozmo), we employed a baseline measure of people's cooperative tendencies (where the emotional manipulation was not yet administered), to be compared with the cooperative behaviours under the impact of the robots' emotion displays. This comparable baseline measure was more appropriate than a human condition (where, for example, a human confederate was trained to perform sad and angry expressions) for distilling the difference made by robots' emotions, since our aim was to examine the utility and social impact of robots' emotion displays, instead of comparing and contrasting the emotional effects of robots than that of humans.

Another advantage of having a baseline measure of cooperative tendencies was that we were able to further investigate whether the impact of the robots' emotions differ by people's baseline cooperative tendencies. According to the EASI model, the meaning and impact of emotional cues can depend on the nature of context (Van Kleef, 2009; Van Kleef et al., 2010). In the scenario of PD games, the perceived nature of such context might be individual-dependent. Some people might opt for mutual profit and strive to build cooperative relationship, but others might act strategically and resort to the highest self-gain (Balliet et al., 2009). It is hence plausible that the factor of robots' emotion displays would very to some degree across individuals given the personal differences in social-decision and emotion processing (Franken & Muris, 2005; Hamann & Canli, 2004). Specifically, we were intrigued to examine whether the emotional effects depend on individuals' baseline cooperative tendencies, in an attempt to identify the precise and effective emotions for robots to display to bolster people's cooperative behaviours in HRI.

### *4.2.4. The current study*

In the present study, we wished to examine whether the context-dependent impact of emotions proposed in the EASI model (Van Kleef et al., 2010) still holds true when (1) the discrimination of competitive and cooperative context is defined subjectively by people's cooperative baseline, as opposed to by experimental manipulation of a task (e.g., Adam & Brett, 2015; M. Lee et al., 2018; Novak et al., 2014); and (2) the emotions are displayed by a robot opponent. Based on the EASI model (Van Kleef et al., 2010), we hypothesised that the social meaning and consequent effects of sad and angry emotions diverge between people with high and low cooperative predispositions. Here we used the term 'predisposition' to refer to the default cooperative tendency people have when facing prisoner's dilemmas, independent of any external factor related to an opponent. More specifically, we predicted that a robot that exhibits sad emotional displays leads participants with more cooperative predispositions to behave more cooperatively (here sadness should be seen as a cue of needing support), while the same sad emotional displays should lead participants with more competitive predispositions to play even more competitively (in this case, sadness should be seen as a sign of weakness in an opponent that can be exploited). On the other hand, an angry robot should induce more cooperative actions among participants with a competitive predisposition (where anger is seen as a warning of a bigger dispute on the horizon), but reduce cooperative intentions among participants with more cooperative predispositions (where anger is perceived to signal an inadequate collaborator) (Van Kleef et al., 2010).

People's willingness to cooperate in PD games denotes the intention of building cooperative relationship with the other while forgoing the possibility of the highest self-gain (Rapoport & Chammah, 1967), which, in the context of HRI, could be seen as a social milestone for people to accept robots as their social partners and commit to a collective task. Past research has also substantiated that people's decisions made in PD games reflect their temperamental cooperative willingness and real-life social-decision making process and behaviours (Balliet et al., 2009; Mokros et al., 2008; Pothos et al., 2011a; Viola et al., 2019). Our research here could provide insight into the possible factors promoting human-robot cooperation and highlight the possibility that bottom-up emotional cues might interact with top-down personal factors, thus making one-size-fits-all robotic programming problematic, and establishing further empirical foundations for adaptive and bespoke programming for social robots. Moreover, investigation into the topic could have several practical consequences as well. First, social dilemmas emerging between humans and robots have the potential to someday, possibly soon, feature in daily life, where robots need to decide between benefits of individual people and the collective interests of human society. These types of discussion are already well underway in the autonomous vehicle development community, where debate and discussion continues over the situations in which people might accept their self-driving cars to sacrifice their own lives to save the lives of (multiple) pedestrians (Bonnefon et al., 2016; Perc et al., 2019). Second, some research evidence has verified that experimental procedures to promote people's cooperative tendencies and altruism (for example, by moral nudging) could have cross-situational effects on their real-life charitable behaviours (Capraro et al., 2019; Capraro & Perc, 2021). Our research here could therefore have implications for real-life HRI, especially to the utility of social robots' emotion displays to enhance the social quality in human-robot cooperation.

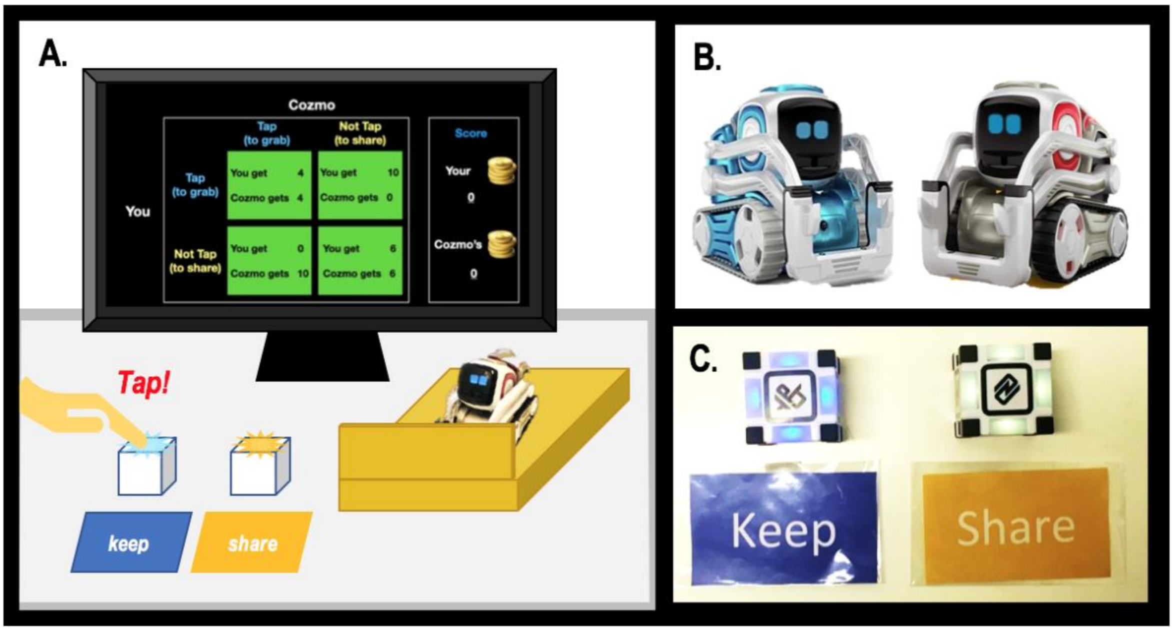
## 4.3. Methods

### *4.3.1. Open science statement*

Prior to data collection, we reported our pilot data, stimuli, and power analysis codes on our Open Science Framework (OSF) page: https://osf.io/tjs8m/?view_only=d52ffba154ed4236b07c663291a5b053. Additionally, we had anonymous data, analysis codes, and materials associated with the study freely available on this OSF page after the study was finished, in keeping with the best research practices proposed by the open science initiatives (Galak et al., 2012; Munafò, 2016).

### 4.3.2. Setup and apparatus

We used the commercially-available Cozmo edutainment robots (manufactured by Anki Inc., **Figure 4-2A&B**) in the experiment as participants' opponents in PD games. The Cozmo robot has been chosen for its capability of expressing diverse facial expressions with its LED face screen (128 x 64 pixel resolution). Additionally, Cozmo is portable (5 x 7.2 x 10 inches in size), affordable, and is flexibly programmed and manipulated via its software development kit (SDK), which make it especially suitable for HRI experimental research (Chaudhury et al., 2020; Cross, Riddoch, et al., 2019). We deployed two separate Cozmo robots for the actual PD games, a blue Cozmo model (named Botz) and a red Cozmo model (named Roxon). One of the robots would consistently display anger, and the other would consistently display sadness (colour and emotion pairing were counterbalanced across participants). By having different coloured Cozmos associated with the two different emotions, this should help prevent the undesirable situation that people would think the same robot was displaying sadness and anger.



| | Tap (to grab) | Not Tap (to share) |
|---|---|---|
| Tap (to grab) | You get 4<br>Cozmo gets 4 | You get 10<br>Cozmo gets 0 |
| Not Tap (to share) | You get 0<br>Cozmo gets 10 | You get 6<br>Cozmo gets 6 |

**Figure 4-2.** Setup and apparatus. (A) Illustration of the experimental setup. During the experiment, participants played games with the robot situated in front of them on a desk, and made game responses by tapping the cubes on the desk. The payoff matrix and real-time game outcomes were shown by a monitor before them. (B) The blue Cozmo (Botz) and the red Cozmo (Roxon) used in the experiment. (C) The interactive cubes that players tapped to make game decisions.

Cooperative and non-cooperative decisions in the current PD game were framed as sharing coins with the other or keeping all coins for oneself, respectively. In each game round, a certain amount of coin endowment was provided to both players, and each was required to make an individual and simultaneous decision as to whether they wanted to share or keep the coins. The exact amount given to each player depended on both of their choices (detailed in the "Game design" section, **Figure 4-5**). During the PD games, a monitor showing the payoff matrix and real-time game outcomes was placed in front of participants (**Figure 4-2A**). Every participant was provided two interactive cubes (**Figure 4-2C**), which illuminated with different colours representing different decisions (blue meant to keep coins for oneself, and yellow meant to share coins with Roxon or Botz). Participants tapped one of these cubes in a round to make a game decision, and the robots used only one interactive cube in games to prevent participants from trying to anticipate the robots' choice by observing the direction it drove to. Also to avoid people peeking over the robots' decision during the responding time, the robots' cube was hidden from participants' sight using a partition between participants and the robot. However, this partition sat above a 4.3 cm thick cardboard box, to ensure the body and expressions of the robots can be fully seen by participants (**Figure 4-2A**). In reality, the robots' game decisions were pre-programmed and they tapped the cube only to make participants believe that the robots were making decisions in real time. All the cubes and the robots were connected via WiFi to the Cozmo application installed on a tablet, and the tablet was paired with a laptop which ran the Python programme to operate the game and the robot, and to record players' game responses by Python log files. The experimental setup followed that developed by previous work by Hsieh et al. (2020).

### *4.3.3. Manipulation and stimuli*

We manipulated the robots' game strategy to always start with a fixed sequence in the first five rounds (share, share, keep, keep, share), followed by a tit-for-tat strategy (i.e., repeating a human player's previous decision) (**Figure 4-3**). This strategy manipulation was adopted by previous studies (de Melo et al., 2010; Kayukawa et al., 2017) to diminish the predictability of agents' actions, and to increase the possibility of experiencing all the four outcomes in the payoff matrix and therefore a higher chance of being exposed to the robots' emotions in the initial five rounds.

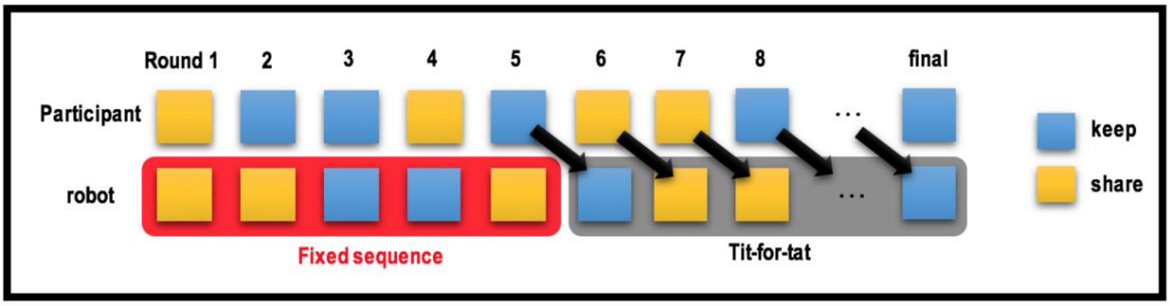


**Figure 4-3.** The strategy manipulation of the robots. In this exemplified game block, the robot started with a fixed sequence of five decisions and followed tit-for-tat strategy till the end. Details of the block design are in the "Game design" section.

The robot expressed emotions not only by its face, but also via vocal interjections (like sighs, laughter, and grunts) and by body movements from its forklift-like arm, head motion, and track directions. In order to select the most appropriate and representative emotional expressions for the robots to display in the main experiment, we required four categories of emotional stimuli (happy, angry, sad, and neutral expressions), with happiness shown after mutual cooperation, anger or sadness displayed after the robots being betrayed by a human, and neutral expression in the rest of situations. We carried out an online pilot experiment via formR platform (Arslan et al., 2020), where participants (*n* = 64, $M_{age}$ = 27.6, 43 females) watched video clips (around 10 seconds each) of a Cozmo robot performing one of the four kinds of emotional animations (happy, angry, sad, or neutral), and answered following each short video clip whether they perceived the expression to be “happy”, “angry”, “sad”, “neutral”, “other” (needed to specify in text), or “I don’t know”. When the answers were happy, angry, or sad, participants were also asked to rate the intensity of the emotion, with slider ratings from “very slight” (1) to “extreme” (100).

The stimulus set for the pilot involved 13 videos clips selected by the experimenters after reviewing all Cozmo’s repertoire of default animations (a total of 348 animations are available on the Github repository – https://github.com/cozmo4hri/animations – created by Chaudhury et al., 2020). Three animations were chosen for each of the three categories – happy (animation numbers: 103, 338, 348), angry (55, 84, 130), and sad (59, 63, 134) – and four (69, 91, 158, 169) for neutral since it is more ambiguous to determine what made neutral expressions. We analysed the mean accuracy rates (the number of answers matching the experimenters pre-defined emotion label / the total number of participants) for each emotional animation, as well as the mean emotional intensity rated by the subjects. The animations with the highest accuracy rate in each category were chosen, which included animation number 348 for happy (*accuracy* = 81.2%, $M_{intensity}$ = 76.1), number 84 for angry (*accuracy* = 98.4%, $M_{intensity}$ = 85.4), number 63 for sad (*accuracy* = 90.6%, $M_{intensity}$ = 57.0), and number 68 (*accuracy* = 39.1%) for neutral. The low accuracy rate for the neutral animations corresponds to the Kuleshov effect, which suggests that people tend to interpret a neutral face or expression by its context or what immediately preceded it, and may perceive a constant face to express different emotions given different contexts (Barratt et al., 2016; Mobbs et al., 2006). Participants in our pilot also reported diverse emotions perceived from the animation number 68, such as doubtful, confused, and surprise. To prevent the possibility that people in the main experiment will also overly interpret the animation which is supposed to be depict neutral emotion, we removed the neutral expression from our manipulation and let the robots directly move on to the next round without displaying any animation. Stimuli and analysis codes for the pilot experiment are available on the OSF page: https://osf.io/tjs8m/?view_only=d52ffba154ed4236b07c663291a5b053

**Figure 4-4** shows the demos of the final set of emotion animations to be used to programme the robots in the main PD game experiment. For the anger animation, the robot’s fork arm hits the table violently, frowns, utters sharp and rapid sounds, and drives left and right repeatedly with apparent agitation (**Figure 4-4A**). For the sad animation, the robot shows a downcast face, sighs, and slowly drops its head down (**Figure 4-4B**). Finally, the happy robot animation features laughing sounds, smiling eyes, arm waving, and driving in circles with excitement (**Figure 4-4C**).

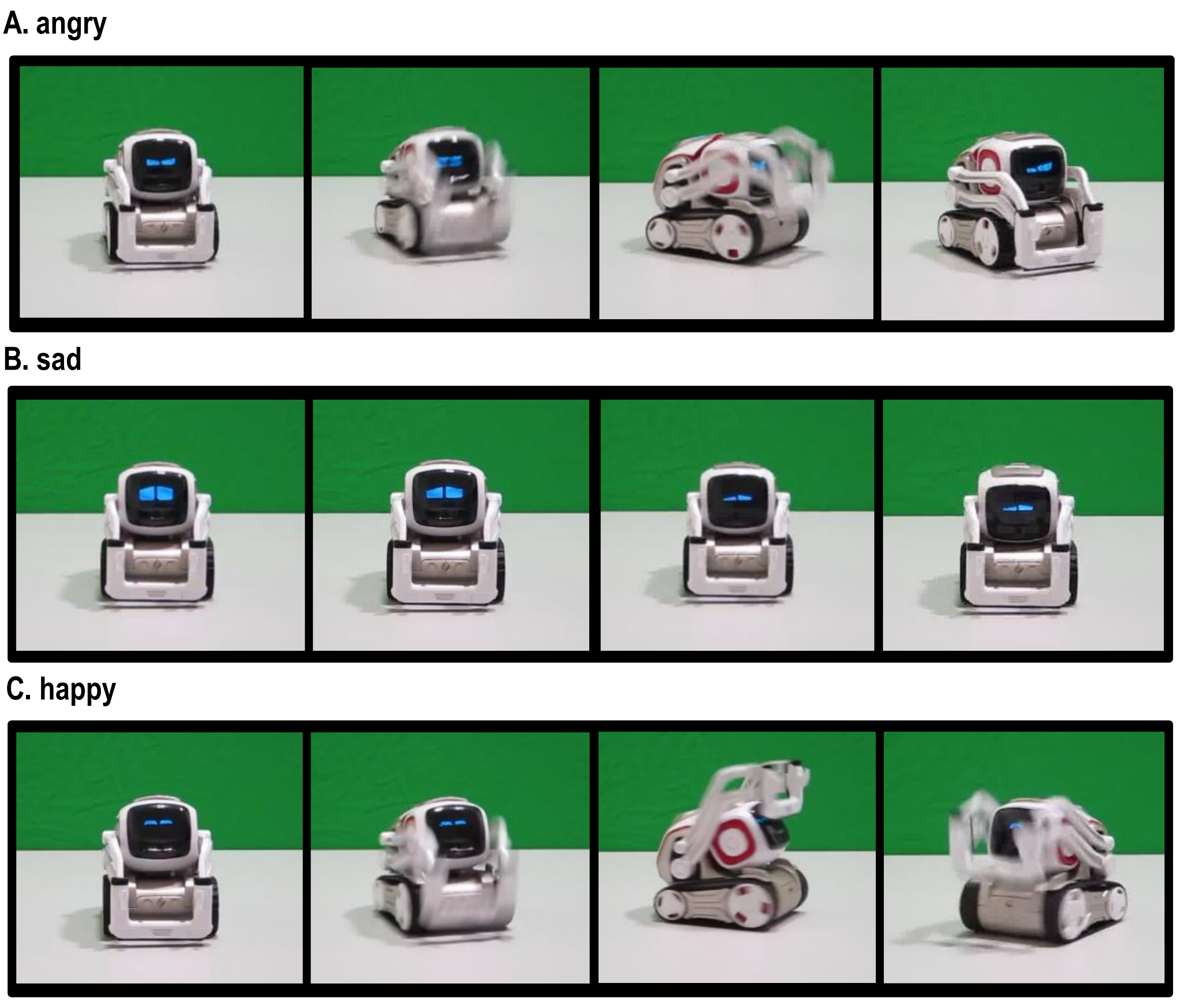


**Figure 4-4.** Demos of the robots’ emotional expressions. (A) Angry expression. (B) Sad expression. (C) Happy expression. Video records of theses demos are available on the OSF page: https://osf.io/tjs8m/?view_only=d52ffba154ed4236b07c663291a5b053

### *4.3.4. Game design*

The experiment was introduced to participants as a robot competition where the experiments wished to know which robot (Roxon or Botz) was the most competent at playing economic games with human interaction partners. The winner robot would be used for future studies, whereas the loser robot would be erased its memory and left on the shelf. The script of memory erasure was adapted from Seo et al.'s (2015) study and has been proved effective to convince participants of the real consequences of the games to robot players (Hsieh et al., 2020). Participants, on the other hand, were monetarily incentivised. The average performance of the last two games blocks would determine their chances of winning a £20 shopping voucher as an extra prize in addition to the standard remuneration for their time.

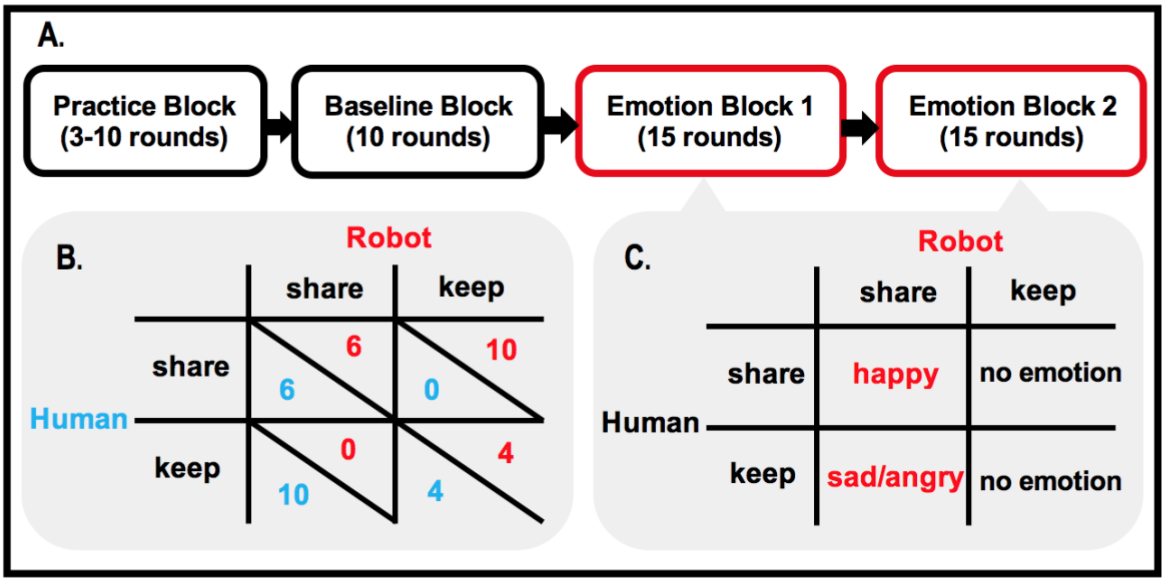

A.
Practice Block (3-10 rounds)
Baseline Block (10 rounds)
Emotion Block 1 (15 rounds)
Emotion Block 2 (15 rounds)
B.
Robot
share
keep
Human
share
keep
6
6
10
0
0
10
4
4
C.
Robot
share
keep
Human
share
keep
happy
no emotion
sad/angry
no emotion

**Figure 4-5.** Experimental design. (A) The order and game rounds planned for the four blocks. Participants firstly familiarised themselves with the game rules in the practice block, and played with a non-expressive Cozmo in the baseline block (as a measure of their cooperative disposition). Finally, they played with Roxon and Botz (one programmed to be sad and the other to be angry) in turn in emotion block 1 and 2. (B) Payoff matrix design. (C) Emotion manipulation of the robots in emotion block 1 and 2. The main manipulation of the robot's sad and angry emotional displays happened after a human player chose to keep coins, but the robot decided to share. The robots' emotion manipulation for the rest of three game outcomes remained the same across emotion block 1 and 2.

The experiment involved one practice block and three blocks of iterated PD games (**Figure 4-5A**). In each round PD game, players would decide to share coins with the other player or to keep all the coins by themselves. Different amounts of coins would be given to players depending on both of their decisions (**Figure 4-5B**). Since prior evidence shows that different designs of payoff matrices in PD games lead to different cooperation rates among human players (Moisan et al., 2018; Rapoport, 1967), we deliberately selected the present payoff from Hsieh et al.'s (2020) study, where two different designs of payoff matrices, one with higher incentives for cooperation and the other with lower incentives, were compared in human-robot PD games. The results revealed that the impact of incentive structures was only significant in the first game round, and over the 20 iterated PD game rounds, participants' cooperative behaviours toward a Cozmo robot were similar in general (mean cooperation rate: 0.40 for the high-incentive game and 0.34 for the low-incentive game). In the high-incentive game condition, participants made significantly more cooperative decisions in the initial game round, which was followed by a quick reduction of cooperation. However, people's decisions in the low-incentive game remained at a constant level throughout the whole game (Hsieh et al., 2020). Consequently, here we adopted the game design with relatively lower cooperatives (**Figure 4-5B**) to forestall the possible initial spikes in cooperative decisions induced by the structure of payoff matrix, and meanwhile ensure that the game context would not bring about ceiling or floor effects on people's cooperative decisions. Designs and content of the four blocks are:

First, in the practice block, participants would familiarise themselves with the skills and the timing of tapping the cubes. The game screen placed in front of participants showed a goal sentence in each round (e.g., "try to earn 10 coins in this round."). Participants only needed to take a corresponding action to make the goal possible (i.e., choosing to keep, in the example). The Cozmo robot used in the practice and the following baseline blocks was an extra robot in addition to Roxon and Botz, and it would always make correct responses to reach the same goal during the practice. By doing so, participants can become more familiar with the payoff matrix and the ways of tapping cubes, without starting to develop their strategies and confounding the following PD game. The length of the practice depended on participants' performance. They can pass the practice by making three consecutive correct and successfully registered responses, otherwise, the practice game ended after 10 rounds. The experimenter supervised participants during the practice to ensure they fully understand how to play the game before moving on.

Second, the baseline block involved ten rounds of PD games played with a non-expressive Cozmo which did not have any emotional animation programmed after either game outcome. The block served as a baseline measure and an indicator of participants' default behavioural tendency in the PD game context before having more extensive interaction with Cozmo robots. We used participants' cooperation rates in the baseline block to predict how they would be influenced by Roxon's and Botz's emotional expression in the analyses.

Third, participants took turns playing PD games with Roxon and Botz, with one displaying sadness and the other showing anger (order and colours counterbalanced). Each emotion block involved 15 rounds of iterated PD games. The robot's negative emotion (sadness/anger) was manipulated after a human player chose to keep but the robot shared. We focused on the particular situation because, firstly, it was a reasonable timing for the robot to show negative expressions as it was betrayed by a human; secondly, it may involve important practical implication to examine whether robot's negative emotions (either sadness or anger) can increase people's cooperative willingness after they already demonstrated non-cooperative behaviours. Throughout emotion block 1 and 2, the robots showed the happy expression after mutual cooperation, as a general signal of cooperative intention. All in all, both robots in the PD games were programmed to send cooperative signals through emotional expressions but in two different ways —— one through showing anger after being betrayed, and the other through displaying sadness after defection. We anticipated the two negative emotions would differentially influence people with different cooperative inclination and baselines in PD games. Participants were not aware the emotion manipulation before actual interaction with the two robots, but only knew that the two robots had different 'personality' and might act diversely.

### 4.3.5. Measures and manipulation check

The main measure of the study was people's decisions made in the three game blocks. Their binomial decisions (to keep or to share) were saved directly with Python log files in the controlling laptop, and were used to compute the cooperation rates (the times sharing/ the total round) in each block.

After participants completed the four blocks of games. We asked them to describe Roxon and Botz respectively, in terms of their emotionality and strategy, and also to report their own strategies adopted when playing with the robots in games. These open-ended questions helped us evaluate the validity of the manipulation on the robots' emotions and strategy, and acquire the qualitative data of how people responded to the two different robots. The manipulation check questionnaire was administered via formR platform (Arslan et al., 2020) on a lab PC.

### *4.3.6. Procedure*

The experiment was planned to be conducted in quiet research laboratory booths located within the institute of Neuroscience and Psychology at the University of Glasgow and within the Department of Cognitive Science at Macquarie University, once behavioural testing was considered safe according to the UK government's, the Australian government's and both University's guidelines concerning COVID-19. Considering the pandemic situation in both sites when the research plan was written, data collection could commence at Macquarie University as soon as a decision was reached on our registered report submission. If lab-based experiments at Glasgow became feasible while data collection was still proceeding, we planned to collect data across both sites to increase participant numbers and diversity. Whenever data collection was carried out in two lab spaces, we would run additional analyses (detailed in "Sampling and analysis plan") to confirm that no systematic difference occurred due to the data collection site. Participants and the experimenter would wear face masks at all times during the study, and we had spare masks prepared if participants required a new or additional mask. In order to reduce unnecessary face-to-face contact, introduction and instruction of the experiment were given to participants by playing a short video on the desktop PC in the lab. After participants provided their written informed consent and showed sufficient performance in the practice block, they were left alone playing games with the robots. The experimenter was seated outside the lab and because the games and robots were operated by a tablet and a laptop connected through the robots' wifi, the experimenter can still monitor the game progress without being present. Finally, participants completed a series of open-ended questions on a PC for manipulation check, as well as their demographics. The whole experiment took approximately one to one and a half hour(s). Participants were debriefed, paid (£6 per hour), and thanked in the end.

### 4.3.7. Participants

We planned to recruit participants aged 18 to 59, with normal or corrected to normal eyesight, and without neurological or psychiatric history. We also aimed to recruit participants who were naïve to robots and to our study. Consequently, people who owned a Cozmo robot, worked with robots on a daily bases, or had participated in our previous experiment (Hsieh et al., 2020) were eligible to the current experiment. Based on a simulation-based power analysis, a sample size of 180 was needed to have 0.9 power finding a significant interaction between the robots' emotions and people's cooperative predisposition on cooperation rates in PD games. The power analysis was carried out with the simglm (v0.8.0.) (LeBeau, 2019) and simr (1.0.5) (Green & Macleod, 2016) R packages, by the following steps.

Firstly, to simulate data for the planned model – **cooperative rate ~ cooperative predisposition*emotion + (1|subject)** – we used relevant meta-analysis results (Balliet et al., 2009; Lench et al., 2011; Pletzer et al., 2018) for our beta weight estimation. For the emotional effects on human judgment, Lench et al. (2011) reported the effect size of *Hedges' g* = 0.18 (from 25 previous studies) when comparing the impact of sad and anger emotions in particular. As to the effect of cooperative predisposition on decisions in economic games, there was no comparable experimental design we can find in the literature and the closest concept is social value orientation (SVO), which refers to people's temperamental motivation to care for others (Murphy & Ackermann, 2014). Over two meta-analysis studies, SVO showed a consistent small to medium effect size on cooperative behaviours in economic games (*r* = 0.30 in Balliet et al.'s, 2009; *r* = 0.32 in Pletzer et al.'s, 2018). However, what we aimed to measure was not people's general traits but their default behavioural tendency in social dilemmas, albeit the two concepts might be closely related. We therefore adopted the 'consevative smallest effect size of interest' (SESOI) strategy (Anvari & Lakens, 2019) and used *r* = 0.20 (or the equivolent *Hedges' g* = 0.40) for our parameter estimation. The interaction of the fixed effects would be generated automatically during the process of data simulation with simglm package (LeBeau, 2019), so we did not need to manually specify the beta weight of interaction.

Second, we simulated data based on aforementioned evidence and calculate statistical power (with the simr package, Green & Macleod, 2016) by the function of sample sizes (**Figure 4-6**). Our main research focus was the interaction between the robots' emotion and people's cooperative predisposition (measured in the baseline block), and the result showed that we needed 180 participants to have 0.9 power finding a significant interaction.

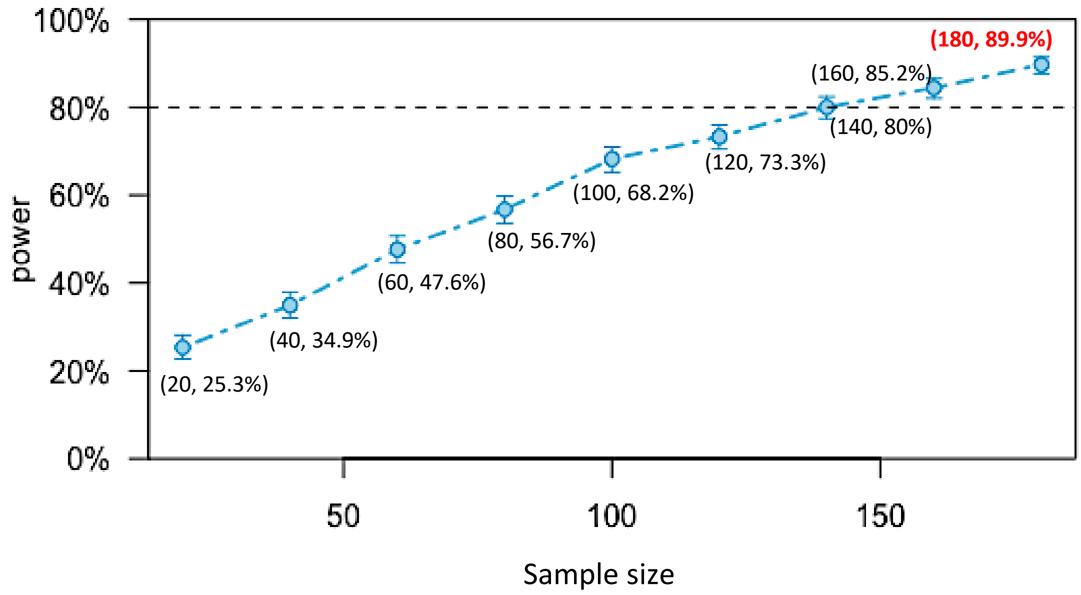


**Figure 4-6.** Power curve for finding an interaction between the robots' emotion and people's cooperative predisposition. Each data point is noted by (sample size, power). The result of simulation suggests that 90% power can be achieved if the sample size reaches 180 (participants).

### *4.3.8. Sampling plan*

Given the large sample size we might need to achieve high power for the effect of interest, we administered sequential analyses to collect data more efficiently (Lakens, 2014b). We planned to perform two interim analyses after 60 participants and 100 participants were recruited, with alpha levels adjusted by Pocock boundary (*p* = 0.0221 for three planned analyses, Pocock, 1977). Following each interim analysis, we would stop data collection early if one of the two conditions was fulfilled: first, if the hypothesis was supported and we found a significant interaction between the robots' emotion and people's cooperative predisposition by the criterion of $p$ = 0.0221; second, if the effect size of interaction was significantly smaller than SESOI ($f^2$ < 0.02, Cohen, 1988).

### 4.3.9. Analysis plan

#### Main analysis

All data analyses would be carried out in R v4.0.1 (R Core Team, 2020). Our hypothesis was that people with higher cooperative predisposition (i.e., high cooperation rates in the baseline block) in PD games cooperate even more when the robot responded with sadness, and would cooperate less when the robot displayed anger, and conversely, people with more competitive predisposition (i.e., low cooperation rates in the baseline block) would cooperate more after the robot displayed anger but became more competitive following the robot's display of sadness. Cooperative and competitive decisions were framed as sharing (coded as 1) and keeping coins (coded as 0) in the current game context. Cooperative rates in the baseline block and in the two emotion blocks would be log-transformed before being feed into our model, where their normally distributed nature would enable values to range from positive to negative values (Benoit, 2011).

The main research question would be examined by a linear mixed effects regression model with the lme4 package (Bates et al., 2015). We would have participants' log-transformed cooperative rates in emotion block 1 and 2 as the dependent variable, and the robots' emotions (anger and sadness) and participants' cooperative predisposition as the fixed factors. For random effects, we would start from the model design specified as follows:

***cooperation ~ emotion*coop_predisposition + (1 | subj_id)***

If the results showed failure in model convergence or a singular fit, we would remove the random intercept term and ran the model as a multiple regression. We expected to find a significant interplay of the robots' emotions and people's cooperative predisposition in participants' cooperative decisions in prisoner's dilemma games (**Figure 4-7**). Post hoc analyses following a significant interaction would be conducted by the effects (v4.1.4) (Fox, 2003) and the emmeans package (v1.4.7) (Lenth, 2020). We planned to examine the impact of cooperative predisposition for sad and angry emotion separately, and anticipated the effects of cooperative predisposition would be opposite in sad and angry conditions -- high cooperative predisposition predicted more cooperative behaviours in sad condition but fewer cooperative behaviours in angry condition (**Figure 4-7**).

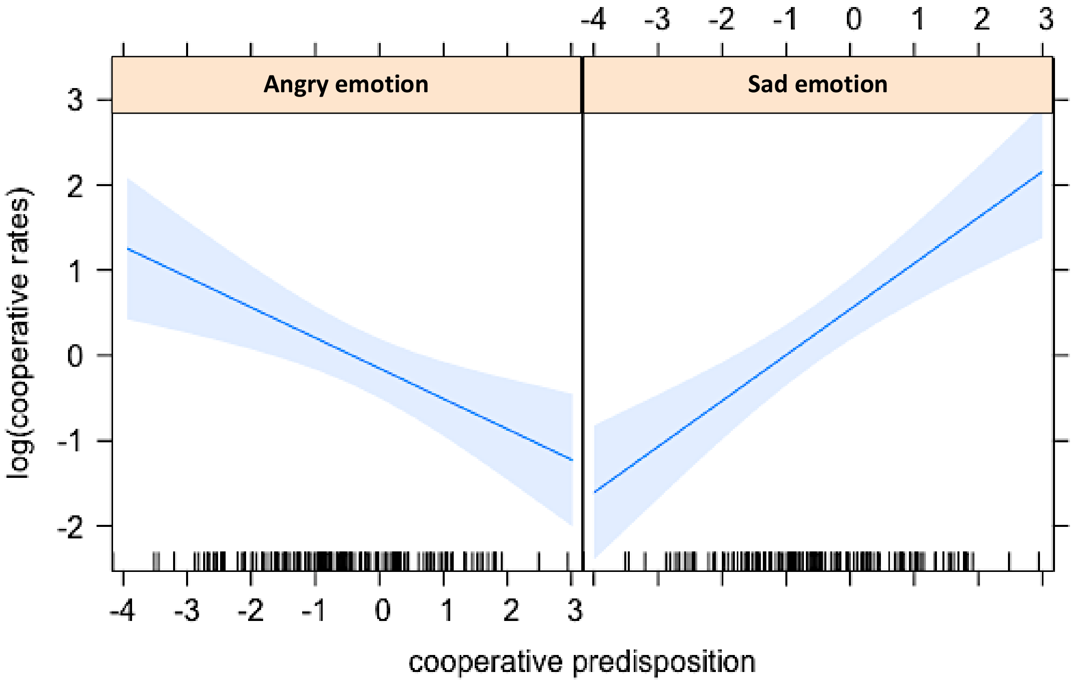


**Figure 4-7.** Hypothetical plot of the expected interaction between the robots' emotions (sad and angry) and people's cooperative predisposition (log-transformed cooperation rates in the baseline block). Participants with higher cooperative predisposition were predicted to become less cooperative by the robot's angry emotion but more cooperative by sad emotion. On the contrary, participants with lower cooperative predisposition were hypothesised to become cooperative by the robot's anger but even less cooperative by its sadness.

### Exploratory analysis

Even though our pilot experiment validated the emotion animations selected for the robots' emotional manipulation for this proposed study, we appreciated that individual variation in human emotion perception, as shown in previous finding on human faces (Barrett et al., 2019), could still emerge among our participant sample. Also, due to the online nature of the pilot experiment, it was plausible to question whether people engaged in playing an embodied human-robot PD game would perceive the robots' emotion displays in the same way as participants did in the online pilot experiment. Therefore, we planned to run an exploratory model with an additional factor – whether participants accurately perceived the robots' emotion displays (*subj_perception*) – to examine whether the subjective perception of the robots' emotion displays was an influential factor shaping the emotional effects:

***cooperation ~ emotion*coop_predisposition*subj_perception + (1 | subj_id)***

This '*subj_percpetion*' factor was derived from participants' subjective reports on "*Did you see the robot displaying any emotion during the game? If you did, what emotion(s) did it display?*" in the post-game questionnaires. When participants' reports of perceived emotions were consistent with the actual emotion manipulation, their answer would be coded as "yes" (i.e., accurately perceived), otherwise their reports would be coded as "no" (i.e., did not accurately perceived). The coding process would be carried out by at least two researchers who were fluent in English. The inter-rater reliability would be analysed with kappa statistics (McHugh, 2012), and we aimed for a minimum of 90% agreement among raters.

Additionally, if the data collection was conducted in both University of Glasgow and Macquarie University, we would run a second exploratory model to control for the possible random variation caused by collecting data across two sites:

***cooperation ~ emotion*coop_predisposition + (1 | collection_site / subj_id)***

The term '*(1 | collection_site / subj_id)*' was to express the nested random effects of subjects within collection sites. Similarly, we would also run the model with the factor '*subject_perception*' added to examine the possible impact from participants' subjective perception of the robots' emotion displays:

***cooperation ~ emotion*coop_predisposition*subj_perception + (1 | collection_site / subj_id)***

The above exploratory models would be compared with the main model by the anova() function in R, to examine the possible improvement in model fit by adding an additional factor or random structure. The model with the best model fit would be reported as the main result of the study, while all the other model output and the process of model selection would also be presented explicitly in our result section.

## 4.4. Results

We carried out the preregistered analyses when 60 participants (mean age = 24.8; 39 females, 17 males, and 4 non-binary) were recruited as per our preregistered sequential analysis plan. Among this sample, 51.67% of participants were White; 38.33% were Asian or Asian British; 1.67% were Black, African, Black British or Caribbean; 1.67% belonged to mixed or multiple ethnic groups; 5% were from other ethnic groups; and 1.67% preferred not to report. Considering the COVID-related restrictions on in-person testing at University of Glasgow and Macquarie University between September and December 2021, all data were collected at the University of Glasgow. Therefore, the exploratory model to control for the potential random effects induced by collecting data at two sites were not performed. We measured participants' daily exposure to robots (L. D. Riek et al., 2011) to ensure that they were generally naïve to robots. In the question of how many robot-related films participants had seen before (from a list of 14 films including Westworld, Real Humans, etc), the median number of robot films seen was 3, with an interquartile range (IQR) of 3. When asking participants how often they engaged with robots in their daily life on a scale from 1 (Never) to 7 (Daily), the median response was 2 (IQR = 2). The results confirmed that participants did not have extensive experience with robots before taking part in this study, and therefore their a priori understanding of robots was unlikely to impact the current HRI.

First, we visualised the distribution of participants' binomial game decisions (to share coins with the robots or not) in the three blocks in Figure 8. From **Figure 4-8**, we could see that the cooperative trends of the three game blocks were similar. Participants started from a higher cooperative tendency in the beginning of each block, and this tendency decreased until the end of the game. The only visible difference between the baseline block and the two emotion blocks was that participants were making slightly more cooperative choices near the end of the block. However, since we did not inform participants of the total number of rounds for each block, it was unlikely that the increasing cooperative decisions were planned deliberately by participants.

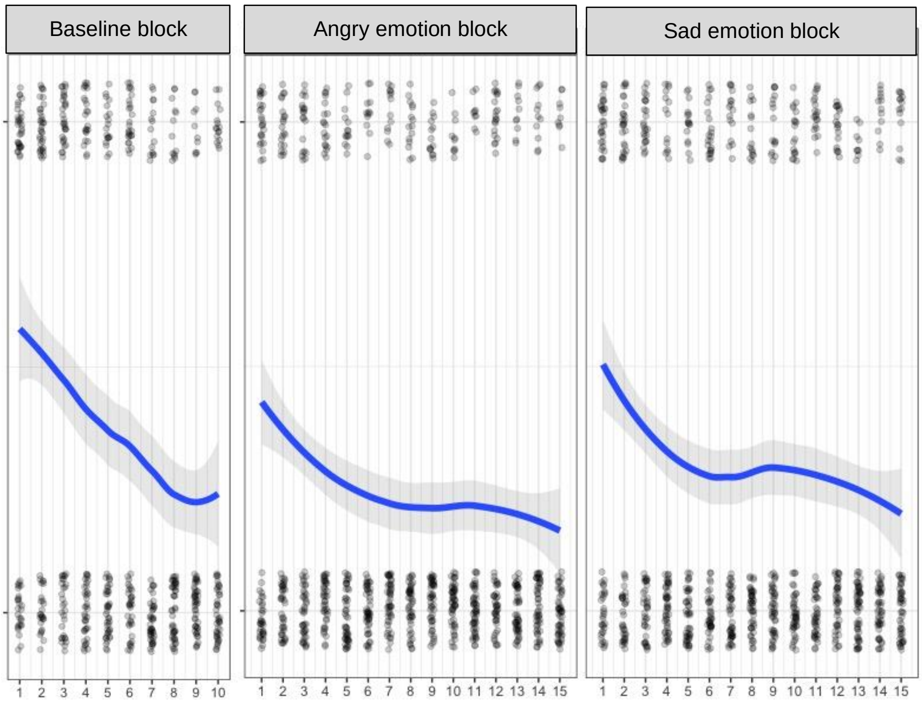


**Figure 4-8.** Binomial game decision distribution across the three game blocks (sharing coded as 1; keeping coded as 0). Nonparametric smoothed curves were added to show the cooperative trends.

Second, we calculated the mean cooperation rates for each block by dividing the numbers of participants' cooperative decisions by the total numbers of game rounds (10 rounds in the baseline block and 15 rounds for each emotion block). In the baseline block, the mean cooperation rate was 37.13%; in the angry block it was 24.83%; in the sad block it was 30.34%. Following the registered analysis plan, we reported the main result of a linear mixed effects model to examine whether there was an interplay between cooperative predisposition and the robots' emotions. For exploratory analyses, we presented the results of the registered model which included an additional factor of participants' emotion perception accuracy. Additionally, we conducted and reported the results of unregistered exploratory analyses, which were the logistic version of the registered models. The logistic models used participants' binomial decisions as the dependent variable, instead of the log-transformed cooperation rates. We carried out this additional modelling because we realised the process of log-transformation (in order to feed the data of cooperation rates to linear models) led to information loss, while using mixed effects logistic regression models on the raw dataset might bring about higher power to detect the effects of interest. Below we present each part of these analyses in detail.

### 4.4.1. Main model results

The model successfully converged with the pre-registered model design. We included the fixed factors of the robots' emotions (anger and sadness) and participants' cooperative predisposition (i.e., log-transformed cooperation rates in the baseline block), the dependent variable of the log-transformed cooperation rates in the two emotion blocks, and the random effects of subject-level random intercepts. As mentioned above, we adopted sequential analyses (with two interim analyses) and therefore we used $p$ = .0221 as the adjusted alpha level (Pocock, 1977). We found a significant factor of participants' cooperative predisposition in this model ($\beta$ = 0.54, *95% CI* [0.17, 0.92], $p$ = .004, $\eta_p^2$ = .23). However, neither the fixed effect of the robots' emotions ($\beta$ = 0.34, *95% CI* [-0.01, 0.69], $p$ = .058, $\eta_p^2$ = .07) nor the interaction between the two factors ($\beta$ = 0.06, *95% CI* [-0.41, 0.53], $p$ = .795, $\eta_p^2$ = .001) was significant. Based on our registered sampling plan of sequential analyses, the data collection was stopped given that the effect size (*Cohen's* $f^2$ = 0.0004) of the interaction (the main effect of interest) is smaller than the SESOI ($f^2$ = 0.02). Namely, the true effect size of the interaction might be smaller than what was considered to be practically meaningful. Therefore, we decided not to pursue such a minor effect with a bigger sample size. Overall, the $R^2$ of the model was .330, with the fixed effects $R^2$ = .178 and the random effects $R^2$ = .153.

### 4.4.2. Registered exploratory model results

In the registered exploratory model, we included an additional fixed factor — the binomial records of whether participants had accurately perceived the robots' emotion as we expected — into the design of the main model. The answers we coded as "successfully perceived the robot's anger" included participants' reports of "angry", "anger", "furious" that were used to describe the robot programmed to display anger; the answers we coded as "successfully perceived the robot's sadness" were the reports that explicitly used the words of "sad" or "sadness" to describe the robot programmed to display sadness. Since the manipulation check was measure by open-ended questions and we did not provide any word bank for participants to choose from, a few participants would use the words that were more ambiguous, like "disappointed", "frustrated", "discontent", "displeasure", to describe the robots' emotional displays. We did not include those answers as evidence of successfully perceiving the emotional manipulation. Also, three participants reported perceiving both negative emotions in a single emotion block: two said they perceived both sadness and anger from the robot programmed to display sad expressions, and one perceived both anger and sadness from the robot programmed to display angry expressions. We also excluded these reports from correct emotional recognition. All in all, the successful perception rate for the robot's angry display was 66.7%, and the rate for the sad display was 51.7%.

We then added this binomial variable of whether participants perceived the robots' emotional manipulation into the model, to examine the extent to which individual differences in emotion perception might influence the results. The model output was presented in **Table 4-1**. We found that none of the fixed factors, nor their interactions, significantly impacted people's cooperative tendencies.

**Table 4-1**

Results of the linear mixed effects model that examined the effects of the robots' emotions, participants' cooperative predisposition, and their emotion perception accuracy on subjects' log-transformed cooperation rates

| | Registered exploratory model | | | | | |
|---|---|---|---|---|---|---|
| | *cooperation ~ emotion*coop_predisposition*subj_perception + (1 \| subj_id)* | | | | | |
| | *Estimate* | *SE* | *Low CI* | *High CI* | *z* | *p-value* |
| intercept | -0.88 | 0.28 | -1.43 | -0.33 | | **.002*** |
| emotion [sad-angry] | 0.38 | 0.33 | -0.27 | 1.03 | .06 | .250 |
| coop_predisposition | -0.02 | 0.53 | -1.04 | 1.01 | .08 | .977 |
| subj_perception [correct-incorrect] | 0.04 | 0.32 | -0.60 | 0.67 | .00005 | .914 |
| emotion* coop_predisposition | 0.63 | 0.57 | -0.48 | 1.74 | .01 | .265 |
| emotion * subj_perception | -0.04 | 0.43 | -0.89 | 0.81 | .00009 | .926 |
| coop_predisposition * subj_perception | 0.63 | 0.56 | -0.47 | 1.73 | .008 | .264 |
| emotion* coop_predisposition * subj_perception | -0.65 | 0.66 | -1.95 | 0.65 | .01 | .326 |
| Subject-level random intercepts | 0.39 | | | | | |
| Residuals | 0.75 | | | | | |

*CI* = 95% confidence interval. **p* < *.0221*

Abbreviations: *SE* = standard error; *CI* = confidence interval.

Overall, the $R^2$ of the registered exploratory model was .347, with the fixed effects $R^2$ = .187 and the random effects $R^2$ = .160. We conducted a model comparison test by the R function anova() to examine whether inclusion of the additional factor ("*subj_perception*") improved the model fit. The result suggested that the difference between the main model (without the "*subj_perception*" factor) and the registered exploratory model (with the "*subj_perception*" factor) was not significant, $\chi^2$(4, 106) = 1.72, *p* = .79.

### 4.4.3. Unregistered exploratory model results

Although the usage of log-transformed cooperation rates allowed the dependent variable values to range from negative to positive, instead of 0 to 1 (Benoit, 2011), we lost some data points because if a participant made no cooperative decision in a game block, the 0 cooperation rate would lead to negative infinity after being log-transformed. This resulted in us having to exclude 14 data points (which resulted in this negative infinity value after log transformation) in order to run linear mixed effects models. Excluding these data points caused crucial information loss since those data represented performances by the most competitive individuals. Therefore, we conducted additional mixed effects logistic regression models to examine if the effects of interest would be better to detect by performing analyses on the raw and complete dataset (binomial game decisions: cooperative decisions coded as 1 and noncooperative decisions coded as 0).

First, in the logistic version of the main model (Model 1 in **Table 4-2&3**), we used participants' binomial game decisions as the dependent variable and added the random intercepts of game rounds into the random effect structure. The rest of the model design remained the same as the main linear model. Similar to the results of the main model, we found a significant effect from participants' cooperative predisposition ($\beta$ = 3.71, *95% CI* [2.16, 5.26], *p* < .001) whereas the main effect of the robots' emotions ($\beta$ = 0.25, *95% CI* [-0.41, 0.92], *p* = .452) and the interaction between the two factors ($\beta$ = 0.15, *95% CI* [-1.39, 1.69], *p* = .851) were nonsignificant.

Second, we ran a logistic version of the registered exploratory model which included the factor of individuals' emotion perception. Again, we controlled for the round-level random effects in the logistic models. We started with the most complex random structure (Barr et al., 2013) for round-level random effects – *(1 + emotion*subj_perception | round)* – but the model failed to converge and we therefore run the model with only the random intercepts of subjects (Model 2 in **Table 4-2**). Results yielded a significant effect from subjects' emotion perception accuracy ($\beta$ = -1.74, *95% CI* [-3.15, -0.33], *p* = .015) whereas all other fixed effects and their interaction were nonsignificant (Model 2 in **Table 4-3**). In general, people who correctly perceived the robots' angry and sad emotions were less likely to cooperate with the robots in emotion blocks. Given the complexity of the three factors involved in the model, we visualised the overall results of the Model 2 in **Figure 4-9** by the R package "effects" (v4.1.4) (Fox & Weisberg, 2018). From **Figure 4-9**, it is possible to see a positive correlation between people's cooperative tendencies in the baseline block and their cooperative probability in emotion blocks, and the correlation might be shaped by people's emotion perception accuracy (albeit the interaction was not significant *p* = .059, by the alpha level of *p* = .0221).

Finally, for exploratory purposes, we ran the Model 3 without the factor of the robots' emotions since its effect did not seem significant in either Model 1or Model 2. In the result of Model 3, the effect of people's cooperative predisposition became significant ($\beta$ = 3.05, *95% CI* [1.23, 4.87], *p* = .001), and the effect of subjects' emotion perception accuracy was not significant ($\beta$ = -0.70, *95% CI* [-1.58, 0.17], *p* = .115) given the pre-defined alpha level of .0221. The output summary of three models and the result of model comparison are reported in **Table 4-3**. Among the three logistic models, none of these three models showed significant improvement in model fit compared to the other two models.

**Table 4-2**

The designs of the three unregistered exploratory models to examine the effects of the robots' emotions, participants' cooperative predisposition, and their emotion perception accuracy on subjects' binomial game decisions

| | Model design | | |
|---|---|---|---|
| | *Fixed factor(s)* | *Random effects* | *Dependent variable* |
| Model 1 | *emotion*coop_predisposition* | *(1 \| subj_id) + (1 \| round)* | *game decisions* |
| Model 2 | *emotion*coop_predisposition* subj_perception* | *(1 \| subj_id)* | *game decisions* |
| Model 3 | *coop_predisposition* subj_perception* | *(1 \| subj_id) + (1 \| round)* | *game decisions* |

**Table 4-3**

The result summary of the three unregistered exploratory models and the outcome of the model comparison.

| | Unregistered exploratory models | | | | | |
|---|---|---|---|---|---|---|
| | Model 1 | | Model2 | | Model 3 | |
| | *Estimate (SE)* | *p* | *Estimate (SE)* | *p* | *Estimate (SE)* | *p* |
| intercept | -2.68<br>(0.35) | **<.001*** | -1.20<br>(0.64) | .061 | -2.03<br>(0.42) | **<.001*** |
| emotion [sad-angry] | 0.25<br>(0.34) | .452 | -1.07<br>(0.71) | .129 | | |
| coop_predisposition | 3.71<br>(0.79) | **<.001*** | 1.09<br>(1.47) | .460 | 3.05<br>(0.93) | **.001*** |
| subj_perception [correct-incorrect] | | | -1.74<br>(0.72) | **.015*** | -0.70<br>(0.45) | .115 |
| emotion* coop_predisposition | 0.15<br>(0.79) | .851 | 2.52<br>(1.59) | .113 | | |
| emotion * subj_perception | | | 1.55<br>(0.88) | .077 | | |
| coop_predisposition* subj_perception | | | 3.09<br>(1.64) | .059 | 0.95<br>(1.00) | .342 |
| emotion* coop_predisposition* subj_perception | | | -2.71<br>(2.04) | .183 | | |
| df | | | 3 | | 0 | |
| AIC | 1973 | | 1998 | | 1976 | |
| BIC | 2006 | | 2048 | | 2009 | |
| Log-likelihood | -980 | | -990 | | -982 | |
| $\chi^2$ | | | 0.00 | | 0.00 | |
| *p* | | | **1** | | **1** | |

*CI* = 95% confidence interval. **p* < *.0221*

Abbreviations: *SE* = standard error.

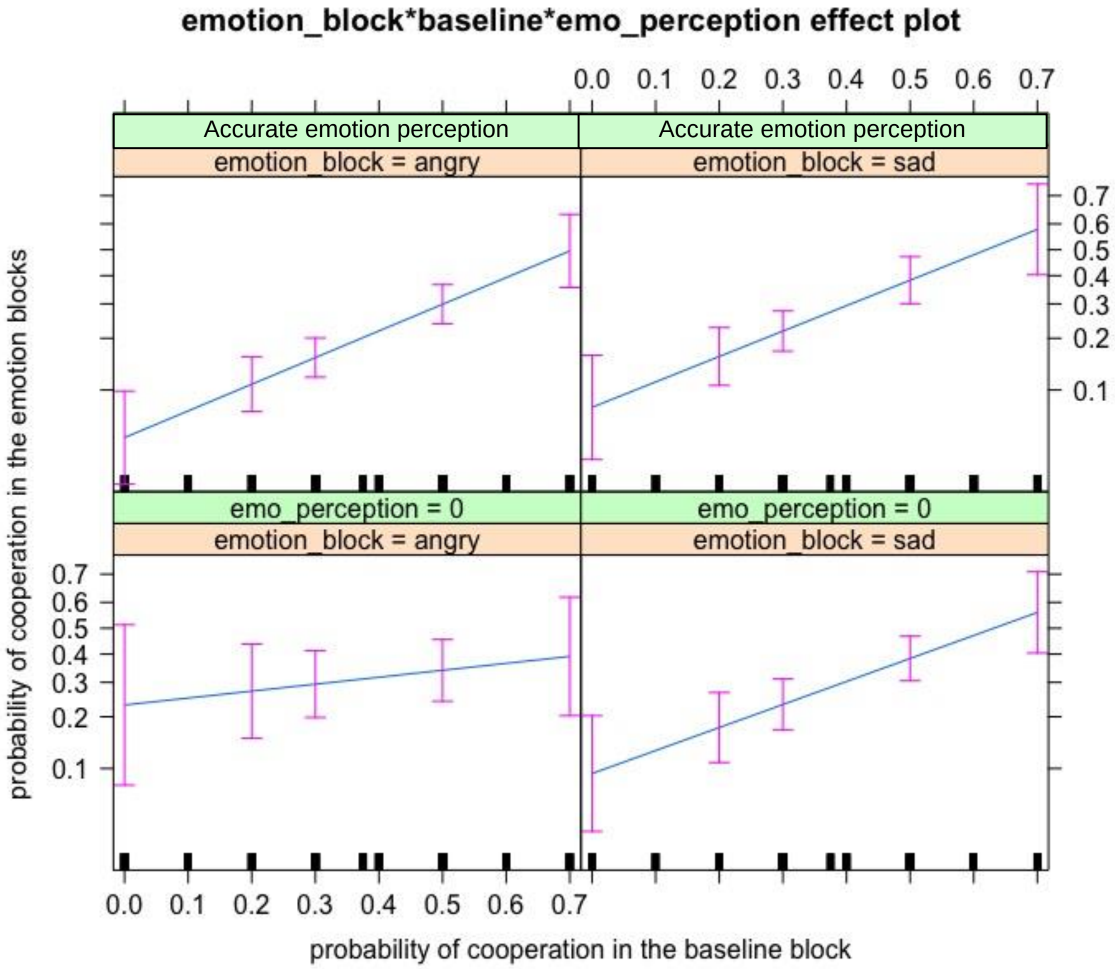


**Figure 4-9.** Effect plot of the unregistered Model 2. The model examined the effects of the robots' emotions, participants' cooperative predisposition and emotion perception accuracy on participants' binomial decisions in the PD games.

## 4.5. Discussion

In this study, we sought to examine the extent to which people's cooperative tendencies in prisoner's dilemma (PD) games are influenced by robots' negative emotion displays and whether the influence of robotic emotion displays is shaped by individual participants' cooperative predispositions (measured in a baseline game block where the robot did not display any emotion). Based on Van Kleef et al.'s (2010) Emotion as Social Information (EASI) model, we predicted that participants who were more cooperative in the baseline block would become even more cooperative when the robot displayed sadness (to show compassion), but less cooperative when the robot displayed anger (to punish who eroded cooperative atmosphere), whereas participants who were competitive in the baseline block would be made to cooperate by the robot's anger (to avoid lose-lose dispute) and would be even more competitive by the robot's sadness (to take advantage of the signs of weakness). The first interim analysis carried out when 60 participants were recruited failed to support these predictions. What has emerged is a significant effect of people's cooperative predispositions on their cooperative tendencies towards both emotional robots. Based on our preregistered sequential analysis plan, we did not continue further data collection given that the effect size of the main effect of interest (the interaction between the robots' emotions and people's cooperative predisposition) was smaller than the pre-defined SESOI. Below we discuss our findings in detail.

We performed a linear mixed effects model to examine the main research question and expected to find a significant interaction between the robots' emotions and participants cooperative predispositions on their log-transformed cooperation rates in emotion blocks. However, only the main effect of people's cooperative predisposition was found to be significant with a large effect size. Participants who showed stronger cooperative tendencies in the baseline block were more likely to cooperate with the robots in emotion blocks. The effect of cooperative predisposition was confirmed by the logistic version of the main model, which used people's binomial decisions as the dependent variable instead of the log-transformed data. This high behavioural consistency within individuals might imply that participants' game decisions in the baseline block reflected their innate cooperative attitudes in this context. Previous research has pointed out the concept of Social Value Orientation (SVO), which refers to people’s dispositional prosocial tendencies during interpersonal interactions (Murphy & Ackermann, 2014). The impact of SVO on cooperative decisions when faced with a social dilemma has been confirmed by at least two meta-analysis studies (Balliet et al., 2009; Pletzer et al., 2018). This work has shown a consistent medium effect of SVO on social decisions. In the present study, we did not include a self-report SVO measure (e.g., the scale by Murphy et al., 2011), because our previous work addressing related questions (Hsieh et al., 2020), did not provide any evidence for a significant relationship between participants’ SVO scores and their cooperative decisions during PD games played with a Cozmo robot, and almost all participants were categorised to the “prosocial” SVO type. It is consequently worth questioning to what extent people’s self-reports of SVO are influenced by social desirability and whether there is a link between the SVO measure (which was not specifically designed to measure the attitudes towards robots) and people’s actual cooperative behaviours in HRI. Although we cannot say for sure if participants’ consistent cooperative tendencies throughout the three game blocks were associated with SVO, our current finding highlights the strong effects of personal factors in cooperation with robots. Furthermore, it seems such top-down personal effects might surpass the bottom-up emotional displays presented by the robots in our experiment. Also, this finding confirms the utility of the baseline measure. Even though our baseline block only involved 10 game rounds (compare to 15 rounds in each emotion block), participants’ cooperation rates were still predictive as to what they would do in similar scenarios.

However, we were surprised to find that participants’ cooperative decisions in the final two rounds of the baseline block seemed to increase a little, reversing the decline in cooperation rates that was observed in previous rounds of the baseline block and in both emotion blocks (**Figure 4-8**). One possible reason behind this could be the robot’s reciprocal (tit-for-tat) game strategy adopted in the second half of the baseline block. A previous study has shown that a robot’s tit-for-tat strategy, compared to a random strategy, in PD games led to higher cooperation rates among participants (Sandoval et al., 2016). We programmed our three robot players to always start with a fixed sequence of decisions, followed by a tit-for-tat strategy, across all the game blocks, in order to make their game strategies less predictable and to increase the chances of exposing participants to the robot emotion manipulation. Still, it was possible that near the end of the baseline block, participants realised the robot's tit-for-tat strategy, especially when the robot did not display any emotional reaction to distract them, and therefore became more willing to cooperate. However, this interpretation remains speculative at this stage, and we futher research will be required to substantiate this explanation. Currently, we cannot exclude the possibility that this finding was simply due to random variance within our sample.

In light of the well-documented individual differences in emotion perception of human facial expressions (Barrett et al., 2019) and of robots' emotion displays (Stock-Homburg, 2021), we planned to explore if the variation in emotion perception would influence participants' cooperative tendencies in PD games and the effects of the robots' emotions. In participants' self-report data concerning observed emotions from the two emotional robots, we did find considerable individual differences in perceiving and reporting the robots' emotional displays. Although more than half of the participants correctly recognised that one of the robots showed sad expressions and the other was angry, some participants described them only in comparative terms (e.g., saying one robot was less angry than the other) or were not aware of any emotional displays by the robots. Quite a few participants seemed to perceive and describe only the negativity of the emotions displayed by the robots and reported the expressions as "displeasure", "frustration", or "disappointment", without explicitly identifying them as sadness or anger. The result of the accuracy rates in perceiving the robots' sadness and anger suggested that the robot's angry expression was easier for participants to recognise, which verifies the conclusion of Stock-Homburg's (2021) review paper suggesting that robots' higher arousal emotions (e.g., anger and happiness) are more consistently and accurately perceived by people (Stock-Homburg, 2021). In the review paper, Stock-Homburg (2021) extensively reviewed 43 studies that examined the emotional expressions displayed by (1) the robots that only have robotic faces (e.g., Barthoc robot, EMYS robot); (2) the robots with anthropomorphic full bodies (e.g., NAO, Pepper robot); and (3) zoomorphic robots (e.g., Keepon robot, KAROTZ robot). Our findings of Cozmo robots therefore added another example of non-humanlike robots whose high-arousal emotional displays are better recognised by people.

To statistically examine the impact of individual differences in perceiving robots' emotional displays, we ran both linear and logistic mixed effects models. We found a significant effect of emotion perception only in the logistic model with all the factors – including the robots' emotions, people's cooperative predisposition and individual emotion perception – involved (Model 2 in Table 3). Participants who correctly perceived the robots' negative emotions displayed after being betrayed by a human player in PD games were less likely to cooperate with the robots in PD games. However, the effect was not significantly shaped by the robot's emotion types (sadness or anger), nor by people's cooperative predisposition, against our predictions. In the current study, the effect of the robots' emotional displays might be constrained by the low recognition rates for robotic emotions in the embodied human–robot PD games (66.7% accuracy for anger; 51.7% for sadness), which were much lower than the recognition rates we measured in our online pilot (98.4% accuracy for anger; 90.6% for sadness). When engaging in economic games played with embodied robots, people might attend mostly to strategic decision-making in order to win, and have limited attention paid to the robot opponents' emotional expressions during games. Although we manipulated the robots so that their emotional displays occurred after each round, when participants were not required to make any other game response, it is still possible that participants were more focused on their next step in the game, and therefore were not fully aware (or focussed on) what the robots were doing.

Contrarily, when examining the influence of individual emotion perception via a linear mixed effects model on the log-transformed dependent variable, we did not find any significant effect from the fixed factors and their interactions. We think these results can be explained by the fact that, when running the linear model, we excluded 14 data points to fix the issue of zero cooperation rates leading to values of negative infinity. This data exclusion also meant we lost performance data from the most competitive participants. Therefore, the usage of mixed effects logistic regression models gave us more power to the detect the effects of interest, and brought about more complete results since the analyses were performed on the entire dataset. The reason why we did not plan on logistic models in the first place was due to the difficulty in performing beta weight estimation for power analyses given the limited number of studies adopting logistic mixed effects model approach in the literature. One study by Moisan et al. (2018) that used this statistical approach focused on the effects of incentive structures on cooperation in interpersonal PD games, rather than robots' emotional displays in human-robot PD games. Consequently, we suggest that more research could consider using mixed effects logistic regression models for analysing such binomial decision data. The strengths of mixed effects models to control for subject-level and stimulus-level random variation also make them outperform ANOVAs or t-tests in many cases (Debruine & Barr, 2019; Field & Wright, 2011).

Among the three exploratory logistic models we conducted, only the personal factors (including cooperative predisposition and individual emotion perception) were found to be relevant to people's cooperative tendencies towards the robots in PD games. Individual differences in emotion perception and cooperative predisposition, compared to the robots' emotion displays, seemed to play a more important role in explaining people's cooperative decisions in the current human-robot PD games. Similar to our finding in the main model, the personal factors drove participants' game decisions more than the robots' emotion types did. Kjell and Thompson's (2013) study also demonstrated the power of personal factors in social decision-making process and found that individuals' SVO outweighed the influence of the essay emotion manipulation tasks on the subjects' cooperative decisions in a computer-mediated PD game. However, since the emotion recognition rates for Cozmo's sad and angry displays were lower than our expectations in this current study, we are unable to state decisively whether personal factors are more relevant than robots' emotional displays to people's cooperative willingness during HRIs in general. Follow-up studies are warranted for a more robust understanding of the effects of robots' emotional dis- plays on people's cooperative decisions in embodied HRIs, and for clarifying how the effects of robotic emotions relate to personal factors, such as coopera- tive predispositions and emotion perception. Future research could consider adopting less cognitive demanding game scenarios to examine the effects of robotic emotional displays on people’s cooperative tendencies, in order to ensure participants have the cognitive resources available to process robots’ emotional displays (and other responses) while enga- ging in social decision-making tasks.

So far, we cannot reject the null hypothesis and cannot claim that people’s cooperative decisions in the human–robot PD games are influenced by the interaction between the robots’ emotions displays (anger and sadness) and people’s cooperative predisposition in the way as the EASI model proposed (Van Kleef et al., 2010). However, it is important to emphasise that the EASI model was derived from human psychological literature and was originally intended to explain and predict interpersonal effects of emotional cues during interpersonal interactions between two people. Therefore, the EASI model might not be the most suitable model to predict the impact of embodied robots’ emotional displays on people’s cooperative decisions. This also demonstrates the limitations of understanding HRIs merely through the lens of human social cognition, while disregarding the fact that social robots may be seen or categorised variably across a continuum that ranges from simple inanimate objects through to humans, given the vast variety in robots’ physical features and social character- istics (Cross & Ramsey, 2021). As such, a robot-specific theoretical framework would be helpful if we are to better explain and predict the social effects of artificial agents’ emotional displays on people’s behaviours.

Moreover, other factors are also likely to influence people’s cooperative tendencies towards robots that were not adequately captured in this study, such as individuals’ intergroup perceptions towards robots (De Jong et al., 2021; Fraune et al., 2017), anthropo- morphism (Torta et al., 2013), trust towards robots (Paeng et al., 2016; Tulk & Wiese, 2018; Wu et al., 2016) and the type of game strategy adopted by robot opponents (de Melo & Terada, 2020). In this study, we focused exclusively on the effects of Cozmo robots’ sad and angry displays, while attempting to control for other individual random variation via mixed effects modelling. Future studies have the opportunity to expand the present investigation by examining the social effects of other robotic emotional displays, since current evidence has shown that virtual agents’ joy and regret expressions might be particularly impactful on people’s cooperative tendencies, compared to displays of sadness and anger (de Melo, Carnevale, et al., 2014; de Melo & Terada, 2019, 2020). Also, follow-up studies could further investigate additional personal, robotic, and contextual factors in PD games for an in-depth and comprehensive understanding of the decision- making process in human-robot cooperation.

Nevertheless, the present findings underscore the utility and importance of performing a manipulation check for emotion manipulation on robots and deploying a baseline measure for people's dispositional cooperative tendencies. Especially for between-subject design or small sample size studies, it is essential to ensure that people's cooperative decisions are driven by the experimental manipulation, rather than by their innate cooperative tendencies or by individual differences in perception. Also, when investigating the social effects of embodied robotic emotions, it is worth conducting pilot studies in more realistic scenarios where people are observing real-life, embodied HRIs, rather than simply checking stimulus validity via online experiments (c.f., Cross & Ramsey, 2021; Henschel, Hortensius & Cross, 2020). It could be the case that the actual effectiveness of emotional manipulation on robots is overestimated in complex and dynamic embodied HRIs. By taking these considerations into account, researchers could truly reveal the potential effects of robots' emotional displays on shaping people's cooperative decisions.

**Acknowledgements:** We thank Kohinoor Darda and Nathan Caruana for the feedback on the manuscript submitted for stage 1 registered report review and Bishakha Chaudhury and Amol Deshmukh for the technical support. The work was supported by the European Research Council under the European Union's 2020 research and innovation program (Grant agreement 677270 to ESC) and Leverhulme Trust (PLP-2018-152 to ESC).

# Chapter 5 General discussion

Through this thesis, I developed a research approach that integrates psychology, Open Science initiatives, and game theory paradigms to rigorously and structurally examine social behavioural aspects of human—robot cooperation. Through three empirical studies (two lab-based and one online), I have provided preliminary evidence revealing relevant factors that shape our cooperative tendencies towards the small, playful Cozmo robots. In this final chapter, I summarise the current findings and discuss the implications and limitations of my work. Furthermore, I reflect on the contribution and challenges of this research approach, and also provide some future directions for researchers in the relevant fields.

## 5.1. Factors shaping cooperative tendencies towards robots in prisoner's dilemma games

Social interaction is, by its nature, dynamic and shaped by multi-level determinants (Fehr & Fischbacher, 2004; C. D. Frith & Singer, 2008; Van Lange et al., 2013). Similarly, in the context of human interactions with artificial agents, researchers have adopted multi-dimensional perspectives to understand human behaviour and perception towards physically embodied robots and virtual agents (Epley et al., 2007; Fiebich, 2018). For example, in Epley et al.'s (2007) 'three-factor theory to anthropomorphism', people's tendencies to humanise non-human agents depend on personal knowledge of the non-human agents, cognitive motivations for understanding current situations, and social motivations for relational connection. Furthermore, these three factors are at play at the levels of an individual's disposition, the situation they find themselves in, their stage of development, and the cultural context (Epley et al., 2007). If we take sociality motivation as an example, people's need to connect with others could be shaped by their current affective states, how lonely they are in a situation, their attachment styles formed in developmental processes, as well as their cultural background (e.g., whether they come from a more individualistic or collectivist culture; Epley et al., 2007). On the other hand, Fiebich (2018) claimed that ideal cooperation entails three key dimensions: (1) shared intentions between two (or more) agents; (2) a certain level of behavioural coordination; and (3) interrelated affective states between collaborators. Therefore, from a robotic design perspective, three domains of skills — including behavioural, cognitive, and affective domains — should be attended to in order to model cooperative robots (Fiebich, 2018).

This thesis took the perspective of human-users in HRIs, and delved into human—robot cooperation via psychological approaches. In line with the theoretical frameworks proposed by Epley et al. (2007) and Fiebich (2018), the current research also incorporates multi-aspect investigations — including contextual, personal, and robotic factors — in order to examine their relevance in human—robot cooperation. Throughout the three empirical studies (Chapters 2 to 4), some factors that shape people's cooperative tendencies in prisoner's dilemma games were revealed that warrant further discussion and synthesis, which the following sections detail.

### *5.1.1. Contextual factors*

In Chapter 2, we investigated people's situational cooperative tendencies towards the Cozmo robot via manipulating the incentive structures of prisoner's dilemma games. Though we did not find evidence to support that people's cooperative decisions throughout the whole game (20 rounds) were significantly influenced by these incentive structures, exploratory analyses revealed that participants' first game decisions were significantly different between high-incentive and low-incentive game conditions. Higher contextual incentives for cooperation led people to make more cooperative decisions towards the robot in the first game rounds, even though defection was always a more profitable decision for individual payoffs in either condition. However, the higher cooperative tendencies in the high-incentive condition dropped off quickly after the initial game rounds, and the rest of people's decisions were mainly driven by the robot opponent's decisions. In both incentive conditions, people demonstrated a reciprocal (tit-for-tat) behavioural pattern in response to the robot. These findings further informed us of the experimental design of the study in Chapter 4, where we only adopted the low-incentive game structure to better distil the possible effects from the robots' emotional displays.

Over the two human—robot prisoner's dilemma game studies, we found a consistent behavioural pattern that people started from higher cooperative tendencies and these tendencies gradually decreased until the end of the game, regardless of the games' incentive structure (**Figure 5-1**). A similar behavioural pattern has also been demonstrated in interpersonal prisoner's dilemma games (Gunnthorsdottir et al., 2007; Houser & Kurzban, 2002; Rand et al., 2011).

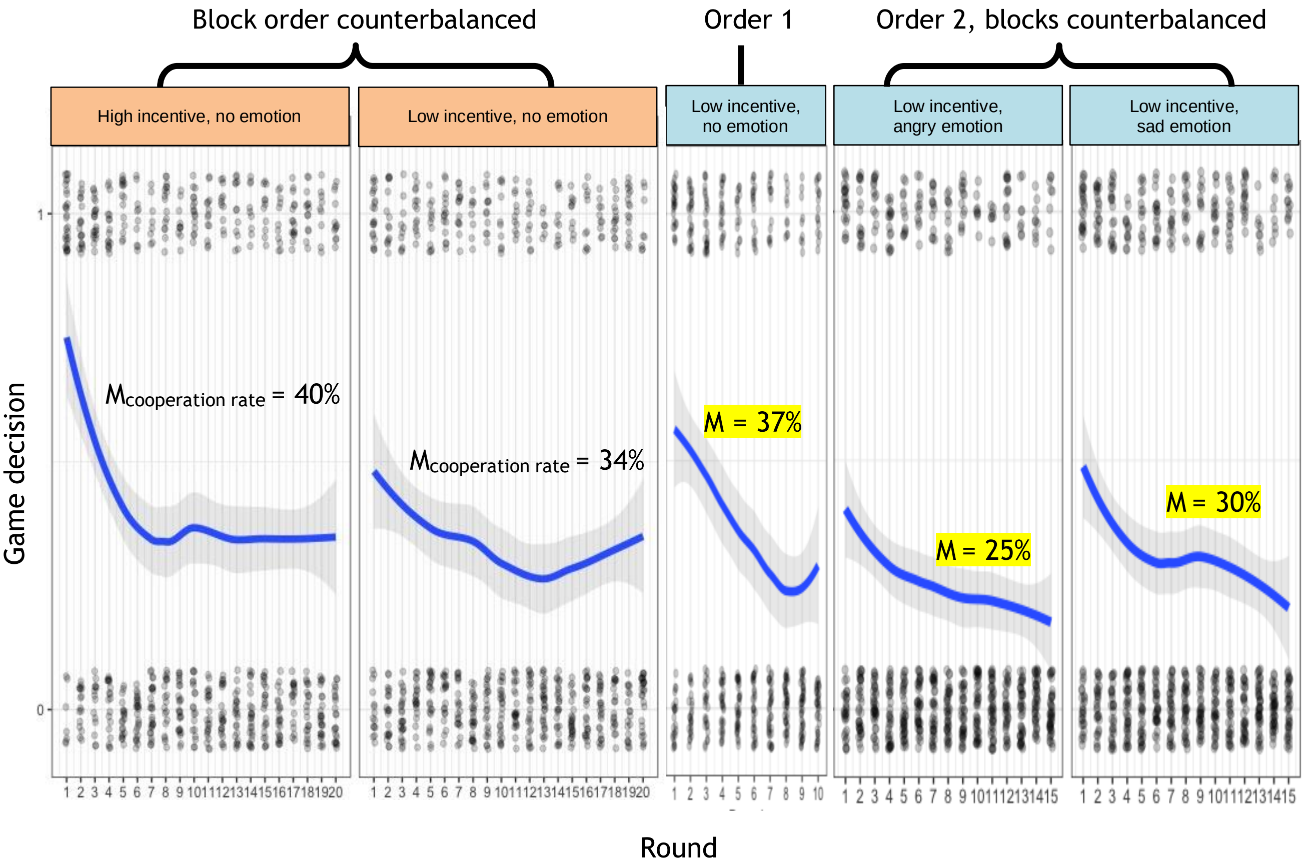


**Figure 5-1.** Binomial cooperative decision distribution across the two lab-based studies on prisoner's dilemma games played with Cozmo robots (sharing/cooperating coded as 1; keeping/defecting coded as 0). 'High incentive' is defined by Rapoport's K-index = 0.6; 'Low K-index' is defined by K-index = 0.2. Nonparametric smoothed curves are added to visualise the cooperative trends. Mean cooperation rate for each game block is calculated by dividing the numbers of participants' cooperative decisions by the total numbers of game rounds. The left two panel plots (orange tags) are from the results of the Chapter 2 study where the robot adopted a random strategy, and the right three panel plots (blue tags) are from the Chapter 4 study where the robots started from a fixed sequence of five decisions and played a tit-for-tat (reciprocal) strategy.

In order to explore the impact of incentive structures on people's cooperative decisions across the two main laboratory studies (Chapter 2 and 4), I carried out exploratory analyses to examine the impact of incentive structures on the combined datasets of the two studies (Chapter 2 and 4; **Figure 5-1**). I only analysed the first game decisions because the first decisions represent people's initial cooperative intentions without being influenced by other factors like the robots' game strategies or the robots' emotional displays. The mixed effects logistic regression model used for this new analysis is specified as follows:

***Cooperative decision ~ incentive_structure + (1|study)***

In this model, participants' game decisions (1 as cooperative decisions; 0 as non-cooperative decisions) are treated as a binomial dependent variable, and the fixed factor is the incentive structure of a game, which is either high-incentive (K-index = 0.6) or low-incentive (K-index = 0.2). The random variation between the two studies (random intercepts) are controlled for in this model. The results do not reveal a significant impact by incentive structures on the first game decisions, $\beta$ = -0.78, *95% CI* [-1.90, 0.34], *p* = .171. As one of the advantages of mixed effects model is to deal with unbalanced designs and missing data (DeBruine & Barr, 2021), the model result should not be biased too much by the unequal sample sizes between the high-incentive (n = 70) and low-incentive (n = 130) conditions. This finding suggests that, at least when playing prisoner's dilemma games with an embodied Cozmo robot, people's decisions are not as impacted by incentive structures as evidence from online versions of these kinds of games played with human opponents suggests (Moisan et al., 2018). However, as already mentioned in Chapter 2, it is important to acknowledge that we only compare the difference between two inventive structures (K-index = 0.6 and 0.2). Further investigations which, for example, include more K-index levels and use different robotic platforms will be required to more robustly and reliably understand the potential impact of incentive structures on people's cooperative decisions in prisoner's dilemma games played with robots.

In addition to incentive structures of prisoner's dilemma games, current literature on interpersonal games has revealed other relevant contextual factors that promote cooperative relationships, including increased time pressure (Rand et al., 2014) and dynamic social networks (i.e., allowing participants to play with new players in games; Rand et al., 2011). Finally, apart from the attempts to uncover contextual factors that enhance people's cooperative willingness towards robots, it might be worth also considering the situations where we might be 'too cooperative' to robots. For example, in Salem et al.'s (2015) study, which was set up in an actual house than in a lab, the majority of participants followed an embodied robot's instructions to help with uncommon and unethical tasks, such as logging into the experimenter's laptop with the password told by the robot (100% of the 40 participants), and pouring orange juice into a plant (67.5% of the 40 participants; Salem et al., 2015). Though these behaviours by the participants (who might have been aware that the robot was intentionally programmed by the experimenter for the experiment) might not necessarily equate to what people would actually do in real-life HRIs, their qualitative data of participants reporting being in 'autopilot' when following the robot's suggestions provide valuable insights for designing (and studying) future scenarios of real-life human—robot cooperation (Salem et al., 2015).

Taken together, throughout the two empirical studies on human—robot prisoner's dilemma games in this thesis, we do not provide compelling evidence suggesting that people's cooperative decisions toward robots are influenced by the incentive structures of the games. Instead, we found consistent declines in cooperative willingness toward a robot opponent as more game rounds were played. Future studies could probe other contextual factors, such as time pressure and dynamic social networks (Rand et al., 2011, 2014), in the context of HRIs, as well as further explore people's cooperative relationships with social robots in other experimental set-ups in addition to economic games.

### *5.1.2. Personal factors*

Several personal factors were explored in this thesis given the well-documented individual differences in social decision-making (Andrighetto et al., 2020; Murphy & Ackermann, 2014; Pletzer et al., 2018) and emotion perception (Barrett et al., 2019). **Figure 5-2** summarises the significant personal factors revealed by each chapter. In this section, I focus my discussion on the factors of people's dispositional anthropomorphism, cooperative predisposition and individual differences in perceiving embodied robots' emotional displays. The empathic trait factor we studied in Chapter 3 is discussed in the next section ("5.1.3

Robotic factors”), since this specific investigation was more related to the emotional contagion effects of the Cozmo robot’s emotion displays.

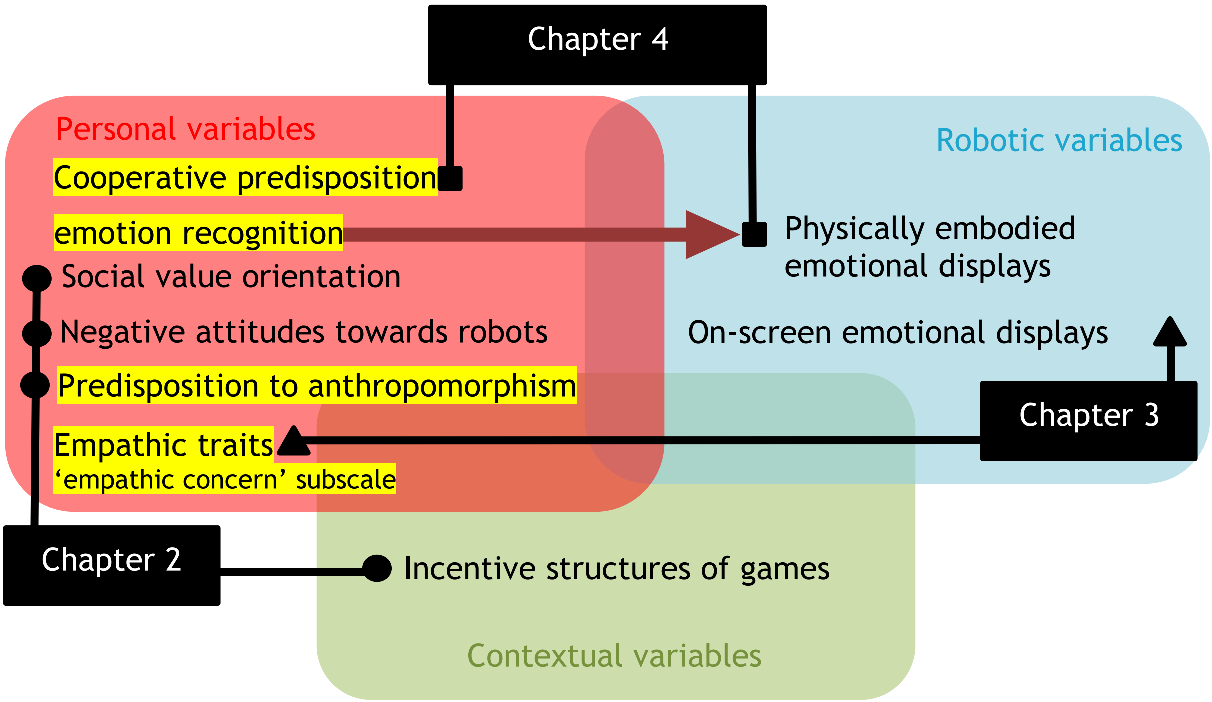

Chapter 4
Personal variables
Cooperative predisposition
emotion recognition
Social value orientation
Negative attitudes towards robots
Predisposition to anthropomorphism
Empathic traits
‘empathic concern’ subscale
Robotic variables
Physically embodied emotional displays
On-screen emotional displays
Chapter 3
Chapter 2
Incentive structures of games
Contextual variables

**Figure 5-2.** An overview of the variables examined across the three empirical studies (Chapter 2 and 4: human—robot cooperation in prisoner’s dilemma games played with a physically embodied robot; Chapter 3: emotion recognition and emotion contagion of robotic emotional displays viewed as videos online). The particularly relevant and influential personal factors found by this thesis are highlighted in yellow.

### (1) Dispositional anthropomorphism

In Chapter 2, we explored three personal factors that might influence people’s cooperative decisions in prisoner’s dilemma games played with a Cozmo robot, including negative attitudes towards robots (Syrdal et al., 2009), social value orientation (Murphy et al., 2011), and predisposition to anthropomorphism (Ruijten et al., 2019). We found a marginally significant effect from people’s predisposition to anthropomorphism on their overall cooperative rates in prisoner’s dilemma games played with an embodied Cozmo robot, suggesting that dispositional anthropomorphism traits predict higher cooperative rates with this particular robot in this particular task. Anthropomorphism involves both the cognitive aspect of attributing human-like characteristics to non-human agents and also the behavioural aspect of treating non-human agents in a similar way to how we respond to other people (Fischer, 2021; Ruijten et al., 2019). In the current literature on HRIs, relatively less attention has been paid on dispositional anthropomorphism, compared to situational anthropomorphism. Most HRI studies have investigated situational anthropomorphism in human—robot economic games via manipulating the agents' human-like features to examine if this induces stronger anthropomorphising responses towards these artificial agents (Fraune, 2020; Nishio et al., 2018; Torta et al., 2013). For example, in Torta et al.'s (2013) online ultimatum game experiments, participants accepted the unfair offers made by a human opponent and a humanoid opponent more frequently than they did those made by a computer opponent. As shown by the scores of anthropomorphism scales towards the three agents (anthropomorphism: human > humanoid > computer), Torta et al. (2013) suggested that participants' differential responses to the three players were due to the level of anthropomorphism (Torta et al., 2013). Another study on human—robot ultimatum games by Nishio et al. (2018) found that having short verbal dialogs with an android opponent (e.g., the android greeted participants by saying "Hello. How are things going?") made people's game responses to it more similar to the responses to a human player. As a consequence, participants were thus becoming less likely to reject unfair offers made by the android. However, the effect of verbal dialog (which the authors considered to engage mentalisation processes) was not found on the less human-like agents included in this study, including a computer and a humanoid robot (Nishio et al., 2018). The authors explained this as an interaction between agents' appearances and verbal dialog on (situational) anthropomorphism (Nishio et al., 2018). Similar to this, Pipitone et al.'s (2021) preliminary results suggested that a robot's inner speech (i.e., talking to itself while collaboratively setting a table for a meal according to etiquette rules with participants) made people anthropomorphise and like it more. However, Pipitone et al. (2021) did not measure the levels of people's cooperative tendencies, nor did they look at participants' social decisions in these recent experiments. These studies focus on how to enhance anthropomorphism towards robots via more human-like robotic forms or behaviours (mostly verbal behaviours). So far, much less is known about the role played by dispositional anthropomorphism in HRIs or in prisoner's dilemma games played with robots. It is important to investigate these questions further, because this could help researchers better interpret and clarify the extent to which people's anthropomorphising responses are driven by robots' characteristics or by individual differences in anthropomorphism tendencies.

In human psychology literature, the importance of measuring individual differences in anthropomorphism has been revealed, and measures of dispositional anthropomorphism have been found to predict people's differential moral decisions and behavioural responses towards non-human agents (Epley et al., 2007; Ruijten et al., 2019). Furthermore, as introduced before, Epley et al. (2007) proposed that anthropomorphism should be studied from dispositional, situational, developmental, and cultural perspectives. Neurocognitive evidence has also demonstrated that dispositional and situational anthropomorphism are correlated, but differentially associated, with the Theory-of-Mind brain network (Hortensius et al., 2021). Finally, this thesis provides evidence supporting the relevance of dispositional anthropomorphism in shaping people's cooperative decisions in human—robot prisoner's dilemma games. Given the previous and current work, it is crucial to consider both dispositional and situational anthropomorphism when researching cooperative tendencies during HRIs.

### (2) Cooperative predisposition and individual differences in perceiving embodied robots' emotional displays

In the second lab-based study on people's cooperative tendencies in prisoner's dilemma games played with Cozmo robots (Chapter 4), we used a within-subject design and found that participants' cooperative decisions (throughout the three game blocks) demonstrated high consistency within individuals, regardless of the emotion manipulations of the robots. In other words, cooperative/competitive participants tended to remain similarly cooperative/competitive even when Cozmo showed sad and angry expressions after being betrayed by human players. The dominant impact of people's dispositional cooperative tendencies has also been studied and affirmed in interpersonal prisoner's dilemma games where people's cooperative decisions are mainly driven by their social value orientation, instead of the emotional manipulations on participants' subjective affective states (Kjell & Thompson, 2013).

Further exploratory analyses (Chapter 4, **Table 4-3**) revealed that participants' cooperative decisions in games could be best explained by their cooperative predisposition (i.e., cooperative rates in the baseline block) and their emotion recognition of Cozmos' emotions (i.e., whether they had accurately recognised the robots' sad and angry displays). Our data demonstrate profound personal differences in perceiving artificial emotions displayed by embodied robots (Cozmo robots, specifically). In general, people who correctly recognised Cozmo's emotions were less likely to cooperate with them, regardless of the emotions displayed by the robots (sadness or anger).

Currently, we failed to provide evidence to support that Van Kleef et al.'s (2010) Emotion as Social Information (EASI) model still applies when (1) emotional stimuli are displayed by embodied robots; or (2) differentiation of cooperative and competitive contexts is defined by baseline measures of people's cooperative tendencies in prisoner's dilemma games. However, these findings highlight the necessity to deploy manipulation checks and baseline measures when evaluating people's cooperative tendencies towards social robots, as well as the potential biases when comparing small-sample and between-subject-design conditions.

### *5.1.3. Robotic factors*

In this thesis, the focus on robotic factors is on robots' emotional displays. We explored people's emotion recognition of both disembodied Cozmo robots presented on a screen (Chapter 3 and the online pilot in Chapter 4) and physically embodied Cozmo robots (Chapter 4). Additionally, we looked into the emotion contagion effects of the robots' emotional displays (Chapter 3), and the effects of embodied artificial emotions on people's cooperative tendencies in prisoner's dilemma games (Chapter 4). In the following sections, I discuss each of these points in detail.

### (1) Emotion recognition of physically embodied and on-screen Cozmo robots

The results of the mean recognition rates measured in the three experiments are summarised in **Table 5-1**.

**Table 5-1.** Recognition rates (%) of Cozmo's emotional displays in the three experiments.

| | | | Recognition rates (%) | | | | |
|---|---|---|---|---|---|---|---|
| | *n* | *embodiment* | ***Angry*** | ***Sad*** | ***Happy*** | ***Surprise*** | ***Neutral*** |
| Chapter 3 | 103 | On-screen | 78.4 | 69.2 | 62.4 | 63.4 | 19.4 |
| Chapter 4 pilot | 64 | On-screen | 90.6 | 85.4 | 75.0 | | 26.6 |
| Chapter 4 | 60 | Physically embodied | 66.7 | 51.7 | | | |

Recognition rate is defined by the percentage of people who recognise the same emotion category as what the experimenters intend the robot to display.

Across the three experiments that measured people's emotion recognition of Cozmo robots (**Table 5-1**), we found that the robots' angry expressions, compared to sad expressions, were more consistently and accurately recognised by people. Additionally, recognition rates for on-screen emotional displays (which were measured in online experiments) were generally higher than embodied emotions displayed by physical Cozmos in the lab. However, it is worth noting that, the recognition rates of embodied emotions were measured during the prisoner's dilemma games played with the Cozmo robots (Chapter 4), which served as a manipulation check. Therefore, the nature of the tasks to measure emotion recognition were considerably different between the Chapter 4 experiment and the other two experiments. Nevertheless, the findings here emphasise the importance of administering manipulation checks when examining the effects of embodied robots' emotional displays. Although we had administered a stimulus validation pilot to confirm the validity of robotic emotional stimuli, the actual emotion recognition rates measured in in-person testing were generally lower than the results of the online pilot. This suggests that it might be more difficult for people to realise and recognise the emotions that embodied robots are displaying when engaging in a task with them. In hindsight, it would have been valuable to run an emotion recognition task/pilot with the embodied robots in the laboratory, to determine whether the low recognition rates during game play are due to the fact that the robot is physically embodied vs. screen based (less likely), or the fact that participants' focus and attention is divided due to the prisoner's dilemma task requirements (more likely).

### (2) Emotion contagion of Cozmo's emotional displays and its relationships with empathic traits

Chapter 3 explored the extent to which people's emotion recognition of an online Cozmo robot's emotional displays are affected by individual empathic traits (measured by Interpersonal Reactivity Index, IRI; Davis, 1983), and the extent to which people's subjective feelings might synchronise with the robot's emotional expressions (which is known as emotion contagion). Only the 'empathic concern' subscale was found to be negatively associated with the emotion contagion effects of the robot's emotional displays. Specifically, participants who reported higher tendencies in 'empathic concern' (i.e., more likely to have vicarious feelings for others' situations) tended to find the robot's emotional expressions *less* contagious.

According to a psychological theoretical framework of empathy, empathy is a multi-dimensional construct, involving the cognitive component of understanding others' perspectives and the affective component of feeling for others' situations (Davis, 1983a; Zaki, 2014). The current finding provides evidence supporting the existence of empathic subtypes and these subtypes might be associate with different mental processes and affective responses (Besel & Yuille, 2010; Perugia et al., 2020). However, one of the limitations of this study is that in the videos participants watched and rated for the experiment described in Chapter 3, the Cozmo robot was displaying emotions devoid of any specific context. Therefore, the relevance of the robotic emotion rating tasks to real-world empathy might be limited. This online study was derived from the stimulus validation pilot in Chapter 4, and the main purpose was to examine whether Cozmo is a suitable robotic platform for displaying recognisable artificial emotion stimuli. For future research that specifically focuses on the role of empathy traits in affective interactions with social robots, other experimental tasks might be more suitable, for example, the tasks to measure people’s hesitance to harm a robot (Darling et al., 2015; Riddoch & Cross, 2021) and people’s reactions when witnessing a robot in pain (Cross, Riddoch, et al., 2019; Rosenthal-von der Pütten et al., 2014; Seo et al., 2015). Nonetheless, the work described in Chapter 3 provides preliminary evidence suggesting that people’s dispositional empathic traits could shape their affective reactions towards a Cozmo robot’s emotional displays and counterintuitively, people who reported higher tendencies in feeling for others’ situations (affective empathy) showed less emotion contagion effects of the robot’s emotional displays.

### (3) The impact of Cozmo’s sad and angry displays on people’s cooperative tendencies

When examining the effects of the Cozmo robots’ negative emotional displays (sadness and anger) on people’s cooperative tendencies (Chapter 4), we did not find evidence for the differential effects from the robots’ sad and angry displays, in contrast to what the EASI model proposes (Van Kleef et al., 2010). Instead, our current results revealed that people who accurately recognised Cozmo’s sad and angry expressions tended to cooperate less with the robots, regardless of the emotion types. These findings contradict evidence reported from online prisoner’s dilemma games played with virtual agents, which suggests that virtual agents’ differential emotional displays do indeed shape people’s cooperative decisions in a similar way to how we are influenced by other people’s emotional expressions in the real world (de Melo et al., 2010, 2011; de Melo, Gratch, et al., 2014a; Hoegen et al., 2018). However, our results are somewhat consistent with Kayukawa et al.'s (2017) study which failed to replicate de Melo et al.'s (2010) findings when applying the same emotional manipulation to an embodied NAO robot. In order words, Kayukawa et al (2017) found that people preferred the emotionally expressive NAO, compared to the non-expressive NAO, but emotional manipulations of the NAO robot did not lead to more cooperative decisions, as demonstrated in de Melo et al.'s (2010) online study. Our study, along with Kayukawa et al.'s (2017), might denote that it is more challenging to produce effective and strong emotional manipulations on embodied robots that can surpass people’s pre-existing cooperative tendencies. However, the discrepant findings reported across screen-based and physically embodied artificial agents in this domain will require further examination in order to understand whether differences are due to the specific type of artificial agent being studied (more likely) or the agent's physical presence/embodiment (less likely).

The finding of robots' negative emotional displays impeding human—robot cooperation is consistent with Kopelman et al.'s (2006) study on interpersonal negotiation. Kopelman et al. (2006) report that participants were more like to make a business deal and cooperate in the future with negotiators who showed kindness and positive attitudes, but not with those who strategically expressed hostility in negotiations (Kopelman et al., 2006). Also, people tended to make more demanding requests to hostile and tough negotiators compared to the positive negotiators (Kopelman et al., 2006). However, this result stands in contrast with Van Kleef et al.'s (2004) findings that people made concessions more easily to angry negotiators but not happy ones. These inconsistent findings between the two studies might be at least partly explained by their different ways of emotional manipulations. In Kopelman et al.'s (2006) study, the authors trained participants to display either positive or negative emotions during negotiations; in Van Kleef et al.'s (2004) study, the authors adopted computer-mediated communicational approach for negotiators to interact. Though psychologists have developed a great amount of theoretical work on the social meanings of interpersonal emotion displays (Manstead & Fischer, 2001; Moors et al., 2013; Van Kleef et al., 2010), to empirically examine the effects of interpersonal emotions in a way that is both well-control and socially natural presents significant challenges. On the other hand, although robots can be used as flexible and reliable research tools for displaying well-controlled emotional stimuli, an outstanding challenge for HRI researchers might be to equip embodied robots with the abilities to express recognisable and effective emotional expressions in dynamic social interactions, given profound interpersonal differences in emotion perception (Barrett et al., 2019; Stock-Homburg, 2021).

There has been growing interest among the affective computing community to understand and characterise the social impact of artificial emotion displays during HRIs (Stock-Homburg, 2021). However, so far, the effects of embodied robots' emotional expressions on people's cooperative tendencies are still unclear, and the artificial emotions displayed by the currently available robotic platforms — including the Cozmo robots used here, and the popular and widely available NAO robot (used by Kayukawa et al., 2017) — seem ineffective in shaping people's cooperative decisions in prisoner's dilemma games. This thesis suggests that there are more factors to take into consideration when understanding the effects of social robots' emotional displays during human—robot cooperation, including people's cooperative predispositions and individual differences in recognising artificial emotions.

## 5.2. Contribution, limitations and future directions

### *5.2.1. Contribution*

This thesis proposes an integrative approach to investigate questions related to HRI that draws on theory and methods from a number of complimentary disciplinary perspectives. Below I outline the three key messages that the current work contributes to the field. These contributions mainly concern the methodology of HRI research, reproducibility, and human—robot cooperation.

**(1) Structural experimental designs for more rigorous HRI investigations**

This approach emphasises the value of adopting perspectives drawn from established psychological theories and empirical findings to design rigorous experimental manipulations and define predictions of people's social decisions during HRIs. This work demonstrates that conducting research in structural HRI (e.g., human—robot economic games) could have several advantages especially when examining causal relationships between factors is the main research interest. First, structural experimental contexts allow researchers to explore people's situational social responses and the impact of environmental factors. In Chapter 2, I revealed that people's initial cooperative decisions towards Cozmo robots would be shaped by the slightly different game structures. This finding highlighted the importance of taking experimental contexts into account when examining people’s social decisions towards robots, and further informed the design of the follow-up study in Chapter 4.

Second, another advantage of structural and evidence-based experimental designs is that studies can be linked with and be comparable to the previous literature using similar experimental contexts. For example, the use of prisoner’s dilemma games in this thesis allowed me to standardise and manipulate the incentive structures of the game contexts according to the well-developed literature (Moisan et al., 2018; Rapoport, 1967; Rapoport & Chammah, 1967), and to compare the present findings with the evidence from interpersonal prisoner’s dilemma games (Moisan et al., 2018) or games played with virtual agents (de Melo et al., 2010; de Melo, Gratch, et al., 2014a).

In short, the empirical investigations in the thesis demonstrate a feasible way to structurally and rigorously investigate people’s cooperative tendencies towards small-size robots by adopting the prisoner’s dilemma game paradigm. Future research could build evidence on the topic by adopting the similar experimental context to examine, for example, whether people’s cooperative decisions in this type of games differ by different robot opponents or by different characteristics of robot opponents.

### (2) Incorporate open science practices in HRI studies to ensure reproducibility of the field

Given the current Open Science movement, which emphasises high quality, more rigorous research practices (including preregistration, accessible research materials and data, and performing high-powered experiments), I suggest that these research practices should be more common in the HRI fields for more reproducible science. The Open Science practices mentioned in the method sections of the empirical chapters in the thesis provide examples of some steps HRI researchers may wish to consider taking in an attempt to conduct more reproducible science. In the three empirical chapters of the thesis, I hope to have highlighted the importance and value of power analyses and pre-defined data collection plans. Especially with the adoption of sequential analyses in Chapter 4, data collection could be more efficient and come to an earlier stop given the adjusted and stricter alpha level. Moreover, all the research materials in relation to the thesis are freely available on the Open Science Framework platform. This not only shows the transparency of the research processes but also makes it easier for future replication studies or relevant work.

Having a clear research proposal planned prior to data collection could force researchers think rigorously about their study designs and sampling plans. This could help tackle the research validity threats which have currently been identified in the field (Innes & Morrison, 2020). This present work therefore contributes to the field by pointing out the importance of Open Science practices and provides actual examples of how these practices can be incorporated into the research processes of HRI.

### (3) Take personal factors into account when investigating human—robot cooperation

From the empirical studies in Chapter 2 and 4, I provide evidence suggesting that people's cooperative tendencies towards robots in human—robot prisoner's dilemma games are mainly driven by individual factors (i.e., a person's cooperative predisposition), and less by external factors including the incentive structures of the games or by the robot's emotional displays. These findings urge that social interactions between people and robots should be investigated thorough a more comprehensive perspective. Namely, the impact of personal factors, as well as contextual factors, in HRIs should be considered when examining people's social responses towards robots. At least in the study in Chapter 4, participants' individual differences in perceiving Cozmo's emotional displays were profound. Therefore, measurements of individual differences or manipulation checks should be administered for a valid and rigorous conclusion of the social effects of robots' emotional cues.

Furthermore, in this thesis, I reveal the utility and strengths of mixed effects models to control individual-level variation when personal factors are not the main research focus. To ensure research validity and reproducibility of research findings in the field, more powerful and sophisticated analysis approaches should be adopted (Belpaeme, 2020). The current work points to the necessity of taking personal factors into consideration when looking into the social dynamics of HRI, and provides an approach (mixed effects models) to control individual variation for valid interpretations of research findings.

In the following sections, I consider several limitations related to these findings and the current research approach. Finally, several ideas for future directions to address these limitations are provided in each of the below sections.

### *5.2.2. Limitations and future directions*

#### (1) The use of Cozmo robots in HRI research

While I contend that this research aids our current understanding in human—robot cooperation, some limitations exist in the thesis that also require careful consideration. In the present research, we only used Cozmo robots for understanding people's cooperative willingness. Our choice of a single robotic platform necessarily constrains the generalisability of these findings to other (let alone all) social robotic platforms (Henschel et al., 2020; Hortensius et al., 2018; Hortensius & Cross, 2018).

Cozmo robots (manufactured by Anki inc.) are highly flexible and customisable, and can be relatively easily programmed to carry out autonomous behaviours for specific research purposes. Furthermore, their affordability and portability make them highly suitable for HRI investigations (Chaudhury et al., 2020). A growing number of researchers have also used these robots for exploring the social and cognitive mechanisms underpinning HRIs (Abubshait et al., 2020; Ciardo et al., 2020; Cross, Riddoch, et al., 2019; Currie & Wiese, 2019; De Jong et al., 2021; Lefkeli et al., 2021; Tan et al., 2018). Additionally, in Chapter 3, we have demonstrated that the emotional expressions (anger, sadness, and happiness) displayed by a Cozmo robot on screen are highly recognisable, and the emotion recognition rates we recorded for Cozmo's emotional displays are considerably better than the average recognition rates from the previous 43 HRI studies (using various other robots, such as NAO, Pepper, Barthoc, Keepon, etc.) recently reviewed by Stock-Homburg (2021).

Although the utility and advantages of Cozmo robots for HRI investigations have been demonstrated by a number of studies, it remains questionable whether the findings based on this specific robotic platform can apply to other different robots (related to the section in Chapter 1 “1.2.3 Generalisability of empirical HRI studies”). More careful interpretations of psychological and HRI findings should be stressed, in order to ensure this field progresses in a reliable, replicable and robust manner (Ramsey, 2021). For the findings presented this thesis, replications on other robotic platforms would be valuable to better understand how Cozmo’s behaviours generalise to different kinds of social robots. The purpose of this thesis is to provide a point of departure for building a better understanding of the mechanisms and consequences of human—robot cooperation, and the evidence provided here should be used and interpreted keeping these limitations in mind.

### (2**) Ecological validity of prisoner’s dilemma games**

There has been a long history of researchers adopting the prisoner’s dilemma paradigm for understanding human cooperative behaviours (Axelrod, 1984; Rapoport & Chammah, 1967; Van Lange et al., 2013). The game represents an analogy of real-life decision-making processes, and evidence suggests clear links between social decisions in economic games and real-world moral judgements and charitable behaviours (Capraro et al., 2019; Capraro & Perc, 2021). In classic prisoner’s dilemma games, a cooperative decision means a participant forgoes short-term individual profit in favour of potentially bigger and longer-term collective interest. This decision-making process seems like it holds some real-life relevance to decisions that might be undertaken during real-world HRIs. For example, business owners might need to decide whether or not to make monetary investments (as well as physical and mental efforts; i.e., short-term individual interests) to use robots for potentially boosting productivity (i.e., bigger long-term collective profit). Though the purchase and maintenance of robots could be expensive and new skills and knowledge will be required for businesses to incorporate robot assistants, in the long run, adoption of these robots might cut down the expenses of labour cost and might generate higher productivity as robots do not require rest or sleep the same way human workers do. This possible future scenario provides a window into the real-world implications that investigations of human—robot prisoner’s dilemma games could apply to. However, as psychologists have been well aware for some time, real-life social interactions and decision-making processes are usually far more complex and dynamic than what can be captured in laboratory experiments (Sanfey, 2007; Van Lange et al., 2013). We will need to collect further explicit and concrete evidence to substantiate the relationships between cooperative decisions measured in lab-based prisoner’s dilemma games and actual cooperative willingness in real-world and long-term HRIs.

### (3**) Beyond behavioural measures**

Another limitation of the thesis relates to the fact that I only investigated people’s cooperative tendencies at a behavioural level, leaving a number of interesting and important questions related to the underlying cognitive or neuropsychological mechanisms supporting cooperative decisions towards robots unexplored. A growing number of researchers are emphasising the importance of incorporating the knowledge and methodology from social cognition and neuroscience to gain a fuller and more in-depth understanding of HRIs (Chaminade et al., 2012; Cross, Hortensius, et al., 2019; Cross, Riddoch, et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018; Wykowska et al., 2016). Also, current empirical studies have used the neurocognitive and physiological techniques like functional magnetic resonance imaging (fMRI) (Chaminade et al., 2012; Cross, Riddoch, et al., 2019; Rosenthal-von der Pütten et al., 2014), eye-tracking (Berg et al., 2019; Peshkovskaya et al., 2017), and electroencephalogram (EEG) (Kompatsiari et al., 2018; Wykowska et al., 2016) to acquire a fuller mind, brain, and behavioural understanding of HRIs.

In addition to quantitative measures, which mainly address questions of *how* we interact with social robots and *if* specific manipulations will affect the ways we interact with them, qualitative approaches shed light on *why* individuals respond to robots in certain ways (Lyons et al., 2019; Riddoch & Cross, 2021). In this thesis, my focus was on the extent to which people are willing to cooperate with Cozmo robots and whether this willingness is shaped by contextual, personal, and robotic factors. Many questions have arisen from this work that warrant answers, which might be well-served by qualitative approaches. These include why people decide to cooperate (or not) with a robot and the thought process people undergo when interpreting the meaning(s) of emotions displayed by a robot during prisoner's dilemma games. Future research on the topic could adopt a multi-method approach (e.g., Hortensius et al., 2021) to expand and deepen the current understanding of human—robot cooperation via neurocognitive, physiological, *and* qualitative measures.

## 5.3. Conclusion

The thesis presents an integrative research approach that draws together psychological perspectives, Open Science practices, and game theory paradigms in an attempt to generate robust and reproducible evidence to advance our understanding of how humans cooperate with robots. Through two lab-based studies on people's cooperative tendencies towards embodied Cozmo robots in prisoner's dilemma games, we did not find evidence supporting that people's cooperative decisions are affected by incentive structures of the games or by the robots' negative emotional displays (sad and angry emotions). Instead, people's behaviours in games demonstrate a strong reciprocal tendency towards the robot opponent, and their cooperative tendencies show high consistency within individuals across a series of game blocks, regardless of the emotion manipulations on the robots. Additionally, via an online investigation of people's emotion recognition of Cozmo's emotional displays on screen, Cozmo's emotional expressions, especially anger, sadness, and happiness, are highly recognisable to participants, which further validates the utility of this particular robotic platform as a suitable tool for examining emotion-related questions in online experiments. However, the current findings also underscore that significant challenges will still need to be overcome to manipulate effective emotional displays on embodied robots. Although our robotic emotional stimuli have been validated via online experiments, the actual recognition rates of embodied Cozmo robots' emotional displays were generally lower than the results of online measures. Future research on this topic could examine the generalisability of current findings by using other robotic platforms and adopt neuropsychological or qualitative approaches to gain better understanding of the cognitive mechanisms underpinning people's cooperative willingness towards social robots. Overall, this thesis contributes to the HRI fields by presenting a research approach that examines the cooperative and social aspects of HRIs in a structural, multidisciplinary and rigorous manner. As social robotics is still in a nascent state, it is important to develop in-depth and comprehensive perspectives of how people interact and collaborate with embodied robots, in order to set solid foundations for future robotic designs that better meet our expectations and that enable us to form cooperative relationships with robots at a psychological level.

# Appendix A:

# Supplementary materials for Chapter 2

**Table S1**

Participants' nationality distribution

| nationality | n | percentage |
|---|---|---|
| Australia | 1 | 1.43 |
| Brazil | 1 | 1.43 |
| Bulgaria | 2 | 2.86 |
| China | 8 | 11.43 |
| Colombia | 1 | 1.43 |
| Dual nationality | 2 | 2.86 |
| Greece | 1 | 1.43 |
| Honduras | 1 | 1.43 |
| Hong Kong | 1 | 1.43 |
| India | 4 | 5.71 |
| Ireland | 1 | 1.43 |
| Israel | 1 | 1.43 |
| Mexico | 1 | 1.43 |
| Nepal | 1 | 1.43 |
| Nigeria | 1 | 1.43 |
| Pakistan | 1 | 1.43 |
| Philippines | 1 | 1.43 |
| Polland | 2 | 2.86 |
| Portugal | 1 | 1.43 |
| Singapore | 3 | 4.29 |
| Spain | 3 | 4.29 |
| Thailand | 1 | 1.43 |
| UK | 25 | 35.71 |
| US | 6 | 8.57 |

Percentage of each nationality category was calculated by n/70 (sample size)

**Table S2**

Participants' cooperation rates (%) across 20 game rounds.

| | **High K-index condition (condition 1)** | **Low K-index condition (condition 2)** |
|---|---|---|
| **Round1** | 80.00 | 57.14 |
| **Round2** | 54.29 | 40.00 |
| **Round3** | 60.00 | 28.57 |
| **Round4** | 45.71 | 40.00 |
| **Round5** | 42.86 | 34.29 |
| **Round6** | 28.57 | 42.86 |
| **Round7** | 22.86 | 40.00 |
| **Round8** | 45.71 | 22.86 |
| **Round9** | 28.57 | 34.29 |
| **Round10** | 45.71 | 28.57 |
| **Round11** | 37.14 | 31.43 |
| **Round12** | 31.43 | 31.43 |
| **Round13** | 34.29 | 17.14 |
| **Round14** | 25.71 | 28.57 |
| **Round15** | 31.43 | 28.57 |
| **Round16** | 31.43 | 28.57 |
| **Round17** | 45.71 | 28.57 |
| **Round18** | 37.14 | 40.00 |
| **Round19** | 25.71 | 45.71 |
| **Round20** | 34.29 | 22.86 |

Cooperation rate (per round) = n of subjects who shared / 70 (sample size)

# Appendix B:

# Rebuttal for "Human–robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots."

NB: Shared with the permission from the chief editor of *Royal Society Open Science*.

## Reviewer 1

Comments to the Author(s):

**This paper reports an experiment testing the effect of incentives on iterated cooperation in human-robot interactions.**

**The paper is well written and interesting. I am positive towards it. However, I have to report a major shortcoming: the sample size is very small, 70 subjects. I am not sure what is the policy of the journal regarding single-study papers based on only 70 subjects. I am aware of (and sensible to) the Replicability Crisis and I tend to not recommend publication of small studies. Maybe the authors can consider replicating their study? Below are the detailed comments that I have taken while reading the paper.**

Authors' response:

We thank Reviewer 1 for their positive comment, and for raising this question regarding the sample size. As mentioned in our manuscript, the sample size was determined by a simulation-based power analysis, with the simr R package (v1.0.5) (Green & Macleod, 2016) and with the parameters estimated by Moisan et al.'s (2018) study. The sample size and study procedures were preregistered prior to data collection, adhering to the open science initiatives (Galak et al., 2012; Munafò, 2016).

Although we agree that a replication study will be useful to validate current findings, we still regard the current manuscript as worth publishing at this stage, since it points to the necessity of taking payoff structure into account when assessing human cooperative tendency in human-robot prisoner's dilemma games. A recent paper by Innes and Morrison (2020) has proposed that studies in human–robot interaction (HRI) are generally lacking a rigorous experimental approach. Our study aims to take positive steps toward rectifying this issue, and presents the utility of game theory and structural economic games to examine human social behaviours in HRI. As research on human–robot cooperation has begun to proliferate in recent years, we argue that it will  be imperative for researchers in this field to at least be aware of the possible impact of incentive structure and of the relevant factors in the context as early as possible, to prevent false interpretation and to ensure research validity. Below we responded to each of Reviewer 1's additional comments in detail.

**R1-1:**

**Page 3, line 4. Another work showing that the level of cooperation in PD is related to the benefit of cooperation is Capraro, Jordan and Rand (2014). I think the authors should also mention that this is inconsistent with neoclassical economic theory, which predicts defection regardless of the payoff structure. However, this is consistent with other models of behaviour, such that preferences for social efficiency (Charness and Rabin, 2002) and the cooperative equilibrium model (Capraro, 2013).**

Authors' response:

We thank Reviewer 1 for providing suggestions for us to include additional relevant work in this area. We now include these papers into our introduction (p.2) to link the current topic with previous broader literature better.

P. 2: "*…The propositions of Rapoport's K-index are in line with several social behaviour models, such as preferences for social efficiency [19] and the cooperative equilibrium model [20]. These models, coupled with empirical evidence from interpersonal PD games [17,21], suggest that people's cooperative tendency is shaped by payoff structures in PD games. This stands in* *contrast to the neoclassical economic theory's prediction [22] that people should act rationally to maximise self-gain and therefore defect all along."*

**R1-2:**

**Figure 1. I don't think it is needed a figure to describe the PD in an academic paper.**

Authors' response:

Considering the broad readership of RSOS, we included Figure 1 to illustrate the basic logic underpinning a prisoner's dilemma scenario, and to allow readers to easily make sense of the K-index equation, as well as the two rules defining a prisoner's dilemma ($T > R > P > S$; $2R > T + S$).

Several previous papers also used both text and a figure (or a table) of a typical payoff matrix to introduce prisoner's dilemma (Bell et al., 2017; Moisan et al., 2018; Sandoval et al., 2016). We therefore believe that there is added value in Figure 1, which should help non-specialist readers to understand the fundamental background of the research topic better.

**R1-3:**

**Page 4, line 28. Which two groups? The high K-index and the low K-index? At this stage of the paper the design is not yet clear, so it is not clear which groups you are referring to.**

Authors' response:

We thank Reviewer 1 for bringing this point to our attention, and have amended the relevant text for clarification.

P. 3: "*…Two Wilcoxon rank sum tests were performed to test whether the participants* ***in high K-index and low K-index conditions*** *differed in their daily engagement with robots or in the number of films featuring robots seen, in order to control for possible confounds in prior experience."*

**R1-4:**

**Page 4, line 31. A p-value of 0.083 is rather small, especially for such a small sample.**

<u>Authors' response</u>:

We used the benchmark of p = .05 to decide whether there was a significant difference between the two groups of subjects in previous experience with robots, and therefore concluded that a result of p = .083 did not reach our significance threshold, and thus we cannot conclude that a significant difference emerged between the two groups. However, we understand that the two groups of participants would not be exactly the same and the between-subject design could have its potential weakness and limitations. Yet, with the adoption of mixed effects models, we consider the impact of possible subject-level random differences should be controlled and minimised.

We answer the question regarding our sample size in detail in comment (14).

**<u>R1-5:</u>**

**Page 6, line 49. The fact that Cozmo's choices were manipulated in this way should be clarified earlier in the paper, in my opinion.**

<u>Authors' response</u>:

We added a sentence about Cozmo's strategy in introduction (p. 2) to make the context of current investigation more specific, and kept the detailed illustration in the Materials and Methods section.

P. 2: "*…we predict that participants who play a high K-index PD game against a robot will make more cooperative decisions than those who play a low K-index game, **regardless of a robot opponent's random ordered game decisions**.*"

**<u>R1-6:</u>**

**Page 6, line 60. A useful meta-analysis on the relationship between svo and cooperation has been recently published by Andrighetto et al (2020). It could be useful.**

<u>Authors' response</u>:

We thank Reviewer 1 for bringing this meta-analysis to our attention. We have now added this into the “Measures” section and our reference list.

P. 5: “*The SVO scale has a significant relationship with cooperative decisions in interpersonal social dilemmas [36,**37**].*”

P. 14: “*37. Andrighetto G, Capraro V, Guido A, Szekely A. 2020 Cooperation, Response Time, and Social Value Orientation: A Meta-Analysis. (doi:10.31234/osf.io/cbakz)*”

**R1-7:**

**Figure 5 is not very informative. To be honest, it’s the first time that I see someone to report the variability of intercepts and fixed effects. The fact that, in the first rounds, the K-index does have an effect is already very clear from Figure 6. It is sufficient to report the stats in the main text.**

Authors’ response:

We appreciate the reviewer’s comment and removed the original Figure 5.

**R1-8:**

**Also, Figure 7 does not add much above Figure 6, so it can be deleted. Instead of describing figures with words, it would much better to report the statistical analysis. Perhaps the authors might consider adding a table with average cooperation by round and by treatment and, for each round, report the statistical difference. (In the current version of the paper, the authors report the stat only of round 1)**

Authors’ response:

We thank Reviewer 1 for sharing this feedback. We could see that the original Figures 6 and 7 (now Figures 5 and 6 in the revised manuscript) were conveying similar information; however, we believe that Figure 7 (now Figure 6 in the revised manuscript), which visualises cooperation rates across 20 rounds, still adds value, as such cooperation trends and fluctuations are difficult to discern either in Figure 6 or a table. Therefore, we prefer to keep Figure 7, but as per Reviewer 1’s suggestion, we have also added a table (Table S1) that lists the exact values of average cooperation rates per round and per condition, in supplementary materials.

Our original intention of running an analysis on the subset of first game decisions was because this demonstrated participants' initial cooperative intention in the games before being impacted by Cozmo's decisions from round 2 onwards. For the remaining game rounds (rounds 2–20), we did not intend to run an individual analysis per round due to concerns over multiple comparisons (Shaffer, 1995), and also due to our motivation to examine the general impact of incentive structure (as demonstrated by the results of mixed effects model, Table 1) rather than its impact per game round.

However, we appreciate that readers might be interested in different perspectives of our data beyond the pre-registered analyses that we focus on in the current report. For that reason, all anonymous data and analysis codes are available on the study's OSF page (https://osf.io/res67/) to enable others to make use of these research materials for their individual interests and purposes.

**R1-9:**

**Page 9, line 50: "inventive structure" -> incentive**

Authors' response:

Thanks for spotting the typo. We have now corrected it.

P. 8: "…*We then trimmed the complexity to arrive at a model that converged by removing random slopes for* ***incentive*** *structure…*"

**R1-10:**

**Pag 10: "we observed an interesting phenomenon…" To be honest, I don't find this to be very surprising: when Cozmo's score is high and subject's score is low, it means that it is more likely that, in the previous round(s), Cozmo defected and the subject cooperated, therefore in the next round the subject is more likely to defect. On the other hand, when  Cozmo's score is high and subject score is high too, it means that is more likely that, in the previous round(s), Cozmo cooperated and the subject cooperated, therefore**

**in the next round, the subject is more likely to cooperate. In sum, I think that you are just re-seeing tit-for-tat from another angle.**

Authors' response:

We thank the reviewer for providing an explanation of the result. We removed the possibly controversial word ("interesting") and adjusted the phrasing into: "*From this analysis, observed that…*" (p. 8).

Also, we incorporated this possible explanation raised by Reviewer 1 into the discussion.

P. 10-11: "*…Nevertheless, an alternative explanation could be that the interaction between Cozmo's and participants' scores on cooperation tendency was an outcome of participants' reciprocal behaviours in games. Specifically, we observed that the participants, when earning little scores, were less likely to cooperate with Cozmo especially when its score was much higher. This was likely because Cozmo had taken advantage of them (i.e., participants cooperated while Cozmo defected) previously for multiple times. Imaginably, people would be unwilling to cooperate with Cozmo after the robot beat them in scores by being uncooperative with them. On the other hand, we found that the participants, when earning high scores already, were more likely to cooperate with Cozmo especially when its scores were also high. This could be explained by previous mutual cooperation and therefore mutual beneficial outcomes. After such win-win cooperative experiences, participants would presumably keep cooperating and reciprocate Cozmo's prior cooperation. Granted, in this study we are not able to provide a decisive answer as to what participants' underlying social and psychological motives were for their game play decision. Nevertheless, the current study provides evidence of dynamic cooperative willingness which changes with the status of human-robot PD games. An important challenge for future research to address will be the factors underpinning people's decision-making process in these scenarios.* "

**R1-11:**

**Human factors. The coefficients of SVO and NARS are similar to that of the anthropomorphism scale, the p-values are also relatively small. It's possible**

**that you do have an effect, but you were unable to detect it because of insufficient power (the sample size is very small).**

Authors' response:

Here again, we used the benchmark of p = .05 to decide whether a predicter significantly impacted the dependent variable (i.e., cooperation rates), and therefore regarded the p-value of .137 (SVO) and the p-value of .145 (NARS) as non-significant predictors, given our pre-registered sample size.

However, we acknowledge the possibility that the current sample size might be insufficient to detect all human factors as it was not specifically calculated to find human factors. Instead, the sample size was estimated by a simulation-based power analysis, with the simr R package (v1.0.5) (Green & Macleod, 2016) and parameters estimated by Moisan et al.'s (2018) study, to detect a significant fixed factor of incentive structure in a mixed effects model. This is also the reason why we included this in exploratory analyses, mainly to provide relevant human factors in the context for future researchers to be aware of.

**R1-12:**

**Did you incentivise cooperation guesses?**

Authors' response:

No. The cooperation rate guesses were involved in the post-game questionnaires and served as a manipulation check of the robot's game strategy.

**R1-13:**

**Ah, another paper is actually Gunnthorsdottir et al. (2007)**

We thank Reviewer 1 for pointing out this relevant paper, which we have now incorporated into our revised discussion.

P. 10: "*…However, the quick decay of cooperation rates and people's reciprocal tendencies were consistent with prior evidence from interpersonal economic games showing that people are less likely to cooperate or make public contributions after experiencing others' uncooperativeness [**43**,44].*"

P. 15: "*43. Gunnthorsdottir A, Houser D, McCabe K. 2007 Disposition, history and contributions in public goods experiments. J. Econ. Behav. Organ. 62, 304–315. (doi:10.1016/j.jebo.2005.03.008)*"

R1-14:

**A major limitation of this work is the small sample size. Perhaps it makes sense to replicate the study.**

Authors' response:

We appreciate Reviewer 1's advice, and do not disagree that direct replications are a service to science. We are also well aware of the importance of achieving sufficient statistical power with a large enough sample size to detect effects of interest, and the proper scientific practices of conducting a power analysis for sample size estimation prior to data collection to ensure any given study is appropriately powered to detect the effect size(s) of interest (Cohen, 1988; Maxwell et al., 2008). Therefore, as mentioned in our manuscript, our sample size (N = 70) was determined by a simulation-based power analysis, with the simr R package (v1.0.5) (Green & Macleod, 2016) and parameters estimated by Moisan et al.'s (2018) study, rather than by any other rule of thumb or subjective experiences.

We acknowledge that the current research topic is rather novel and any chosen power analysis can provide only a statistical estimation of a required sample size rather than the definite answer. Further, we agree that research findings in general should be examined with extensive replication studies. However, although our results did not perfectly fit our initial hypotheses, it was not necessarily caused by an insufficient sample size, but could possibly be people's unique behavioural responses to robot opponents in prisoner's dilemma games, which has not been well studied before. We thus would argue that the current study, which strictly followed open science practices, is worth publishing at this stage, in order for researchers in the nascent field of HRI to gain insights from our findings and experimental approach. Further, we would argue that, by such collective efforts, researchers from HRI and any other field that studies human behaviour will be moving in the right direction for building knowledge and understanding on a topic in a cumulative and rigorous way that enables the community to improve the quality of experimental designs and methods based on others' past experiences (Munafò et al., 2017; Open Science Collaboration, 2017).

**R1-15:**

**Moreover, the fact that you did not find incentive effect with two incentives does not mean that there is no incentive effect in general.**

Authors' response:

This is indeed a very good point that Reviewer 1 raises, which we now include as one of the study's limitations in the discussion.

p. 11: "*…Finally, in the current study we only examined the difference between K-indices of 0.6 and of 0.2. Future research could include more levels of K-indices to acquire a fuller understanding of how our willingness to cooperate with a robot changes according to different incentive structures of human-robot PD games.*"

# Reviewer 2

**Comments to the Author(s):**

**In this manuscript, the authors develop a computer-mediated human-robot PD game and examine the frequencies of participants sharing coins (cooperating) with a Cozmo robot in high and low K-index conditions. They mainly study the impact of incentive structures on cooperative decisions and explore the impact of interactive behavior on participants and the impact of real-time game performance on participants. The results show that the incentive structures of a human-robot PD game have an effect on human cooperation only at the beginning of the game. Throughout the whole game, participants' cooperative/noncooperative decisions are driven more by the robot's decision (following a tit-for-tat strategy) and the presentation of game scores in each round. I think that this work is interesting. Here I have some following comments or questions about this work.**

R2-1:

**In Figure 7, the cooperation rate of some low K-index is obviously higher than that of high K-index. I suggest the author provide some detailed explanations for this.**

Authors' response:

The fact that some cooperation rates in the low K-index conditions were higher than that in the high K-index condition could be explained by the first part of our exploratory analyses, where we found that participants' decisions after the first game round were mainly driven by Cozmo's previous decisions rather than the incentive structures. This appears to have caused random fluctuations in cooperation rates across both conditions, due to the way we manipulated Cozmo's game strategies (random but resulting a 50% cooperation rate).
We thank Reviewer 2 for raising this issue and we have now added a sentence before the exploratory section to make it clearer.

p. 8: *"…Below we present exploratory analyses which we conducted in order to identify possible factors driving the fluctuations in participants' cooperation tendencies."*

R2-2:

**The result shows that empathetic response could confound results to some unknown degree. I wonder whether changing the degree of the punishment can reduce its impact on the results.**

Authors' response:

In the discussion, we pointed out the possible confounding effect caused by the cover story of the experiment (i.e., erasing Cozmo's memory if it lost) as a caveat for future researchers if they wish to adopt the same (or a similar) script. However, in the current study, this kind of individual-level random effect should be already controlled by the random effects modelling of our mixed effects models, and should not confound our results. We appreciated Reviewer 2's question nonetheless, and have rephrased the description to prevent confusion.

P. 11: "*…We acknowledge the possible confounding impact caused by individuals' empathetic responses and therefore adopted mixed effects models to better control for possible subject-level random effects.…*"
As to the question regarding changing the degree of punishment, we think any script that is effective in convincing participants of the real consequences of the games to a robot can be a valid manipulation in this context. We used this specific script since it has previously been proven effective by Seo et al.'s (2015) study, but also agree that future research could further explore the best way to make the game scenario as meaningful as possible to both human and to robot players alike.

R2-3:

**If the punishment for Cozmo is increased, I wonder whether the final results of the experiment will be affected.**

Authors' response:
We would not expect any significant change in the final results if Cozmo's punishment were increased, because our mixed effects models already control such subject-level random effects (if any). However, we agree with Reviewer 2 that it would be insightful for future studies to manipulate the degrees and the content of the rewards and punishment to a robot, to find out if there is a more suitable way to frame the game scenario to be meaningful to both humans and robots. Thus, we now have added this point into our discussion as one of the directions for future researchers to look further into:

P. 11: "*Future studies could use more structured quantitative measures to assess how meaningful each participant thinks an economic game is to a robot or any other non-human agent, to ensure the validity of this kind of paradigm.* ***For example, researchers could manipulate (e.g., increase or decrease) the extent of punishment and rewards a robot receives in human-robot PD games, and measure whether and how participants' perception and cooperative willingness are changed.***"

# Appendix C:

# Rebuttal for "Examining how the display of emotions influences human–robot cooperation in a prisoner's dilemma game."

NB: Shared with the permission from the anonymous reviewers and editor at *Cognition and Emotion*

## Reviewer 1

**How and why cooperation emerges in social dilemmas is an intensely investigated subject with obvious practical ramifications. Methods of mathematical modeling and network science have been applied successfully and with much effect in recent years to shed light on the problem from many different perspectives, and also to outline many different ways on how solutions could be obtained. The authors are certainly right in pointing out that the interactions between humans and robots are bound to increase in the near future, and that thus the subject is very much worth investigating. I would be happy to review a revised submission that takes into account the following comments.**

Authors' response:
We thank Reviewer 1 for their positive comment and agreement on the importance of the topic. Below we respond to each of Reviewer 1's additional comments in detail.

**R1-1:**
**There are a couple of useful reviews that would fit very well to the introduction and to the subject in general, and also promote reading across fields, namely Social and juristic challenges of artificial intelligence, Palgrave Commun. 5, 61 (2019) and Mathematical foundations of moral preferences, J. R. Soc. Interface 18, 20200880 (2021), which both concern social**

**dilemmas and how this could play out in settings involving human-machine interactions.**

Authors' response:

We thank Reviewer 1 for bringing these relevant review papers into our attention. We have now incorporated them into our introduction, to link the research topic better with real-life human-machine interactions.

- Page 12, line 254-266:

"Moreover, investigation into the topic could have several practical consequences as well. First, social dilemmas emerging between humans and robots have the potential to someday, possibly soon, feature in daily life, where robots need to decide between benefits of individual people and the collective interests of human society. These types of discussion are already well underway in the autonomous vehicle development community, where debate and discussion continues over the situations in which people might accept their self-driving cars to sacrifice their own lives to save the lives of (multiple) pedestrians (Bonnefon et al., 2016; Perc et al., 2019). Second, some research evidence has verified that experimental procedures to promote people's cooperative tendencies and altruism (for example, by moral nudging) could have cross-situational effects on their real-life charitable behaviours (Capraro et al., 2019; Capraro & Perc, 2021). Our research here could therefore have implications for real-life HRI, especially to the utility of social robots' emotion displays to enhance the social quality in human-robot cooperation."

R1-2:

**In general, I would expect the introduction to cover a bit better research dedicated to cooperation in the prisoner's dilemma game, especially also related to modeling, where a lot of results has accumulated in recent years concerning various settings and interactions. Here misinformation and trust strike me as particularly relevant in human-robot cooperation, and the introduction and discussion could in this regard be much improved and brought more up-to-date.**

Authors' response:

We thank Reviewer 1 for the suggestion to enrich our introduction with coverage of more recent studies on prisoner's dilemma games. We now include more recent research modelling human-decision making process in social dilemmas, and current findings on human-robot prisoner's dilemma games. We hope this will give readers a clearer understanding of the recent development in the relevant fields. At the same time, we are mindful of not going into too much detail about individual factors so as not to depart too far from the main study focus on robots' emotion displays. However, in the future discussion section, we will incorporate a broader discussion over the relevant factors in human-robot cooperation (such as misinformation and trust) thanks to these helpful suggestions provided by Reviewer 1 at this stage.

- Page 9, line 176-182:

<u>"An extensive body of literature on interpersonal PD games has used both experiments and data simulation to model and theorise on the emergence and evolution of human cooperative behaviours (Axelrod & Hamilton, 1981; Embrey et al., 2018; Rapoport & Chammah, 1967). With mathematical modelling, more recent research has provided considerable insights into the mechanisms and factors supporting or hampering cooperation across various social dilemma situations (e.g., in dyads and in groups) (Bravo et al., 2012; Ito & Tanimoto, 2018; Kopp et al., 2018; Perc et al., 2017)."</u>

- Page 9-10, line 188-193:

<u>"Recent research on human-robot PD games has provided preliminarily insights into the impacts of reciprocity (Sandoval et al., 2016), trust (Paeng et al., 2016), dialogic verbal reactions (Maggioni & Rossignoli, 2021), and a Nao robot's emotion displays (Kayukawa et al., 2017) on HRI. Yet, the preliminary evidence raises more questions than answers at this stage, especially with respect to the effects of robots' emotion displays in PD games."</u>

**<u>R1-3:</u>**

**It would also improve the paper if the figure captions would be made more self contained. In addition to what is shown, one could also consider a sentence or two saying what is the main message of each figure, where**

**applicable.**

Authors' response:

We thank Reviewer 1 for the advice to improve our figure captions, and have now added more description in the following figures (underlined text shows the amendment made to the captions):

- (page 14) **Figure 2.** Setup and apparatus. (A) Illustration of the experimental setup. During the experiment, participants will play games with the robot situated in front of them on a desk, and make game responses by tapping the cubes on the desk. The payoff matrix and real-time game outcomes will be shown by a monitor before them. (B) The blue Cozmo (Botz) and the red Cozmo (Roxon) used in the experiment. (C) The interactive cubes that players tap to make game decisions.

- (page 19) **Figure 5**. Experimental design. (A) The order and game rounds planned for the four blocks. Participants will firstly familiarise themselves with the game rules in the practice block, and play with a non-expressive Cozmo in the baseline block (as a measure of their cooperative disposition). Finally, they will play with Roxon and Botz (one programmed to be sad and the other to be angry) in turn in emotion block 1 and 2. (B) Payoff matrix design. (C) Emotion manipulation of the robots in emotion block 1 and 2. The main manipulation of the robot's sad and angry emotional displays happens after a human player chooses to keep coins, but the robot decides to share. The robots' emotion manipulation for the rest of three game outcomes remains the same across emotion block 1 and 2.

- (page 25) **Figure 6.** Power curve for finding an interaction between the robots' emotion and people's cooperative predisposition. Each data point is noted by (sample size, power). The result of simulation suggests that 90% power can be achieved if the sample size reaches 180 (participants).

- (page 26) **Figure 7.** Power curve for the main effect of Cozmo's emotion. Each data point is noted by (sample size, power). The result shows that sufficient statistical power (90% power) is already achieved when we have 40 samples (or more).

R1-4:

**The abstract does not communicate any conclusions or findings, only what has been done and what the implications of this might be. As a reader, I would certainly be curious to learn whether the display of negative emotions hinders cooperation or not, and whether positive emotions promote cooperation.**

Authors' response:

We thank Reviewer 1 for the suggestions regarding the abstract. We have now revised it to include a short description of methods and expected results. However, since the data collection will only be carried out after we receive an in-principle acceptance from *Cognition and Emotion* (given the manuscript is submitted for Stage 1 Registered Report review), we are not yet able to state the effects of robots' emotion displays, but only our predictions concerning them.

- Page 2, line 7-14:

"Participants will play iterated prisoner's dilemma games with a non-expressive robot (as a measure of cooperative baseline), followed by an angry, and a sad robot, in turn. Order of sad and angry robot opponents will be counterbalanced. Based on the Emotion as Social Information model, we expect that participants with higher cooperative predispositions will become less cooperative when a robot opponent displays anger, and more cooperative when sadness is displayed. Contrarily, according to this model, participants with lower cooperative predispositions should become more cooperative towards an angry robot and less cooperative toward a sad robot."

R1-5:

**Staying with the abstract, I am not sure that the rather lengthy background and referencing to COVID-19 is needed in that much detail. There are numerous other reasons why robots are likely to play an ever more important role in our lives. And sentences like "As the world is likely to embrace a "new**

**normal" after COVID-19, including remote education, increased working from home..." will surely be perceived as highly contentious by many. Maybe it is true and maybe it is not, we will see, but this has not place being stated like this in a paper that has very little to do with COVID-19 in the first place. I would strongly recommend a rewrite there.**

Authors' response:

We thank Reviewer 1 for raising the questions concerning the abstract. We have now included more descriptions about our methods and result predictions in the abstract. Hopefully, the revised version provides a better outline of the proposed research, and we wish to further update the abstract after conducting data collection and analyses, if an in-principle acceptance is received.

**R1-6:**

**The title could then also be more factual, like "The display of negative emotions hinders human-robot cooperation in prisoner's dilemma games".**

Authors' response:

We thank Reviewer 1 for the suggestion of a more informative title. Again, as the manuscript is still at stage 1 registered report review and we haven't carried out our data collection and result analyses, we are reluctant to change the paper's title to reflect findings that are only expected at this stage. However, if the regulations of registered reports in *Cognition and Emotion* permits a change in title, we are more than happy to change the current title to a more factual one after robust results are found.

## Reviewer 2

**The topic of building cooperation between humans and robots is highly significant and studying the role of emotion in accomplishing this is timely. The study builds on prior work with emotional agents, but research with robots - and of this type in particular - is missing.**

**The hypotheses are motivated based on some of the prior work, though it's not entirely clear to me that those with different cooperative orientations will behave differently given certain emotions (what would be the explanation from appraisal theory). Nevertheless, the hypotheses will be appropriately tested through the proposed procedure.**

**The materials (e.g., robot emotion displays) were appropriately selected and validated. The experimental design and task are also well motivated from prior work.**

**The sample size seems appropriate and the sampling plan reasonable.**
**The main and exploratory analyses are adequate to test the hypotheses and provide insight on this topic.**
**I look forward to the results.**

Authors' comments:

We are grateful for Reviewer 2's positive comments and the general agreement on our research plan proposed here. Indeed, we are motivated to gain a deeper understanding in human-robot cooperation via psychological emotion theories, and wishing the eventual findings could shed light on both psychological fields and robotics development.

# References

Abubshait, A., Beatty, P., McDonald, C., Hassall, C. D., Krigolson, O., & Wiese, E. (2020). *A win-win situation: Does familiarity with a social robot modulate feedback monitoring and learning?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/6z75t

Adam, H., & Brett, J. M. (2015). Context matters: The social effects of anger in cooperative, balanced, and competitive negotiation situations. *Journal of Experimental Social Psychology*, *61*, 44–58. https://doi.org/10.1016/j.jesp.2015.07.001

Agrawal, S., & Williams, M.-A. (2018). Would You Obey an Aggressive Robot: A Human-Robot Interaction Field Study. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 240–246. https://doi.org/10.1109/ROMAN.2018.8525615.

Alves-Oliveira, P., Kuster, D., Kappas, A., & Paiva, A. (2016). Psychological Science in HRI: Striving for a more Integrated Field of Research. *2016 AAAI Fall Symposium Series*, 4.

Andrighetto, G., Capraro, V., Guido, A., & Szekely, A. (2020). *Cooperation, Response Time, and Social Value Orientation: A Meta-Analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/cbakz

Anvari, F., & Lakens, D. (2019). *Using anchor-based methods to determine the smallest effect size of interest*. [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/syp5a

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, *52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Axelrod, R. (1984). *The evolution of cooperation*.

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*, *211*(4489), 1390–1396.

Baillie, L., Breazeal, C., Denman, P., Foster, M. E., Fischer, K., & Cauchard, J. R. (2019). The Challenges of Working on Social Robots that Collaborate with People. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3290607.3299022

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes and Intergroup Relations*. https://doi.org/10.1177/1368430209105040

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barratt, D., Rédei, A. C., Innes-Ker, Å., & van de Weijer, J. (2016). Does the Kuleshov Effect Really Exist? Revisiting a Classic Film Experiment on Facial Expressions and Emotional Contexts. *Perception*, *45*(8), 847–874. https://doi.org/10.1177/0301006616638595

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 591–594. https://doi.org/10.1109/ROMAN.2004.1374827

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects: Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bell, R., Mieth, L., & Buchner, A. (2017). Separating conditional and unconditional cooperation in a sequential Prisoner's Dilemma game. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0187952

Belpaeme, T. (2020). Advice to New Human-Robot Interaction Researchers. In *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Nature.

Benoit, K. (2011). Linear Regression Models with Logarithmic Transformations. *London School of Economics*, *22*(1), 23–36.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Berg, J., Lottermoser, A., Richter, C., & Reinhart, G. (2019). Human-Robot-Interaction for mobile industrial robot teams. *Procedia CIRP*, *79*, 614–619. https://doi.org/10.1016/j.procir.2019.02.080

Besel, L. D. S., & Yuille, J. C. (2010). Individual differences in empathy: The role of facial expression recognition. *Personality and Individual Differences*, *49*(2), 107–112. https://doi.org/10.1016/j.paid.2010.03.013

Bhargava, S., & Chakravarti, A. (2009). Empowered Consumers=Benevolent Consumers? The Effects of Priming Power on the Appeal of Socially Responsible Products. *NA - Advances in Consumer Research*, *36*, 831–832.

Bierman, D., & Jolij, J. J. (2020). Dealing with the Experimenter Effect. *Journal of Scientific Exploration*, *34*(4), 703–709. https://doi.org/10.31275/20201871

Bland, A. R., Roiser, J. P., Mehta, M. A., Schei, T., Sahakian, B. J., Robbins, T. W., & Elliott, R. (2017). Cooperative behavior in the ultimatum game and

Prisoner's dilemma depends on players' contributions. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2017.01017

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Bourne, P. E., Polka, J. K., Vale, R. D., & Kiley, R. (2017). Ten simple rules to consider regarding preprint submission. *PLOS Computational Biology*, *13*(5), e1005473. https://doi.org/10.1371/journal.pcbi.1005473

Bravo, G., Squazzoni, F., & Boero, R. (2012). Trust and partner selection in social networks: An experimentally grounded model. *Social Networks*, *34*(4), 481–492. https://doi.org/10.1016/j.socnet.2012.03.001

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, *42*(3–4), 167–175. https://doi.org/10.1016/S0921-8890(02)00373-1

Broadbent, E. (2017a). Interactions with Robots: The Truths We Reveal About Ourselves. *Annual Review of Psychology*, *68*, 627–652.

Broadbent, E. (2017b). Interactions with Robots: The Truths We Reveal About Ourselves. In *SSRN*. https://doi.org/10.1146/annurev-psych-010416-043958

Cahapay, M. B. (2020). Rethinking Education in the New Normal Post-COVID-19 Era: A Curriculum Studies Perspective. *Aquademia*, *4*(2), ep20018. https://doi.org/10.29333/aquademia/8315

Capraro, V. (2013). A Model of Human Cooperation in Social Dilemmas. *PLoS ONE*, *8*(8), e72427. https://doi.org/10.1371/journal.pone.0072427

Capraro, V., Jagfeld, G., Klein, R., Mul, M., & de Pol, I. van. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, *9*(1), 11880. https://doi.org/10.1038/s41598-019-48094-4

Capraro, V., Jordan, J. J., & Rand, D. G. (2015). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports*, *4*(1), 6790. https://doi.org/10.1038/srep06790

Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of The Royal Society Interface*, *18*(175), rsif.2020.0880, 20200880. https://doi.org/10.1098/rsif.2020.0880

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*(May), 1–9. https://doi.org/10.3389/fnhum.2012.00103

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.1162/003355302760193904

Charness, G., Rigotti, L., & Rustichini, A. (2016). Social surplus determines cooperation rates in the one-shot Prisoner's Dilemma. *Games and Economic Behavior*, *100*, 113–124. https://doi.org/10.1016/j.geb.2016.08.010

Chaudhuri, A., Sopher, B., & Strand, P. (2002). Cooperation in social dilemmas, trust and reciprocity. *Journal of Economic Psychology*. https://doi.org/10.1016/S0167-4870(02)00065-X

Chaudhury, B., Hortensius, R., Hoffmann, M., & Cross, E. S. (2020). *Tracking Human Interactions with a Commercially-available Robot over Multiple Days: A Tutorial* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fd3h2

Ciardo, F., Beyer, F., De Tommaso, D., & Wykowska, A. (2020). Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, *194*, 104109. https://doi.org/10.1016/j.cognition.2019.104109

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*(02), 139–153. https://doi.org/10.1017/S0140525X0350005X

Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F. S., & Paiva, A. (2016). Just follow the suit! Trust in human-robot interactions during card game playing. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*. https://doi.org/10.1109/ROMAN.2016.7745165

Côté, S., Kraus, M. W., Cheng, B. H., Oveis, C., van der Löwe, I., Lian, H., & Keltner, D. (2011). Social power facilitates the effect of prosocial orientation on empathic accuracy. *Journal of Personality and Social Psychology*, *101*(2), 217–232. https://doi.org/10.1037/a0023171

Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: Applying neurocognitive insights to human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180024. https://doi.org/10.1098/rstb.2018.0024

Cross, E. S., & Ramsey, R. (2021). Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. *Trends in Cognitive Sciences*, *25*(3), 200–212. https://doi.org/10.1016/j.tics.2020.11.009

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2018.0034

Currie, L. Q., & Wiese, E. (2019). Mind Perception in a Competitive Human-Robot Interaction Game. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1957–1961. https://doi.org/10.1177/1071181319631284

Darling, K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 770–775. https://doi.org/10.1109/ROMAN.2015.7333675

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679–704. https://doi.org/10.1098/rstb.2006.2004

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. a., & Blow, M. (2009). KASPAR - a minimally expressive humanoid robot for human-robot interaction research. *Applied Bionics and Biomechanics*, *6*(3), 369–397. https://doi.org/10.1080/11762320903123567

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10.*

Davis, M. H. (1983a). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113.

Davis, M. H. (1983b). Empathic Concern and the Muscular Dystrophy Telethon: Empathy as a Multidimensional Construct. *Personality and Social*

*Psychology Bulletin*, *9*(2), 223–229. https://doi.org/10.1177/0146167283092005

De Jong, D., Hortensius, R., Hsieh, T.-Y., & Cross, E. S. (2021). Empathy and Schadenfreude in Human–Robot Teams. *Journal of Cognition*, *4*(1), 35. https://doi.org/10.5334/joc.177

de Melo, C. M., Carnevale, P., & Gratch, J. (2010). The influence of emotions in embodied agents on human decision-making. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6356 LNAI*, 357–370. https://doi.org/10.1007/978-3-642-15892-6_38

de Melo, C. M., Carnevale, P., & Gratch, J. (2011). The Effect of Expression of Anger and Happiness in Computer Agents on Negotiations with Humans. *10th International Conference on Autonomous Agents and Multiagent Systems AAMAS 2011*, *Aamas*, 937–944. https://doi.org/10.1016/j.jclepro.2016.12.062

de Melo, C. M., Carnevale, P., & Gratch, J. (2012). The effect of virtual agents' emotion displays and appraisals on people's decision making in negotiation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-33197-8-6

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*(1), 73–88. https://doi.org/10.1037/a0034251

de Melo, C. M., Gratch, J., & Carnevale, P. J. (2014a). Humans vs. Computers: Impact of Emotion Expressions on People ' s Decision Making. *IEEE*

*Transactions on Affective Computing*, *1*(2), 1-11. https://doi.org/10.1109/TAFFC.2014.2332471

de Melo, C. M., Gratch, J., & Carnevale, P. J. (2014b). The importance of cognition and affect for artificially intelligent decision makers. *Proceedings of the 28th Conference on Artificial Intelligence*, 336–342.

de Melo, C. M., Marsella, S., & Gratch, J. (2019). Human cooperation when acting through autonomous machines. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1817656116

de Melo, C. M., & Terada, K. (2019). Cooperation with autonomous machines through culture and emotion. *PLOS ONE*, *14*(11), e0224758. https://doi.org/10.1371/journal.pone.0224758

Debruine, L. M., & Barr, D. J. (2019). Understanding mixed effects models through data simulation. *PsyArXiv*.

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920965119. https://doi.org/10.1177/2515245920965119

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Elliot, A. J., Jury, M., & Murayama, K. (2018). Trait and perceived environmental competitiveness in achievement situations. *Journal of Personality*, *86*(3), 353–367. https://doi.org/10.1111/jopy.12320

Embrey, M., Fréchette, G. R., & Yuksel, S. (2018). Cooperation in the Finitely Repeated Prisoner’s Dilemma*. *The Quarterly Journal of Economics*, *133*(1), 509–551. https://doi.org/10.1093/qje/qjx033

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F. (2017). An experimental psychological perspective on social robotics. *Robotics and Autonomous Systems*, *87*, 363–371. https://doi.org/10.1016/j.robot.2016.08.029

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*. https://doi.org/10.1007/s12110-002-1012-7

Fiebich, A. (2018). Three Dimensions in Human-Robot Cooperation. *Robophilosophy/TRANSOR*, 147–155.

Field, A. P., & Wright, D. B. (2011). A Primer on Using Multilevel Models in Clinical and Experimental Psychopathology Research. *Journal of Experimental Psychopathology*, *2*(2), 271–293. https://doi.org/10.5127/jep.013711

Fischer, K. (2021). Tracking Anthropomorphizing Behavior in Human-Robot Interaction. *ACM Transactions on Human-Robot Interaction*, *11*(1), 1–28. https://doi.org/10.1145/3442677

Fisher, R., & Katz, J. (2000). Social Desirability Bias and the Validity of Self-Reported Values. *Psychology and Marketing*, *17*, 105–120. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, *8*(15), 27.

Fox, J., & Weisberg, S. (2018). Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals. *Journal of Statistical Software*, *87*(9). https://doi.org/10.18637/jss.v087.i09

Franken, I. H. A., & Muris, P. (2005). Individual differences in decision-making. *Personality and Individual Differences*, *39*(5), 991–998. https://doi.org/10.1016/j.paid.2005.04.004

Fraune, M. R. (2020). Our Robots, Our Team: Robot Anthropomorphism Moderates Group Effects in Human–Robot Teams. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01275

Frennert, S., & Östlund, B. (2014). Review: Seven Matters of Concern of Social Robots and Older People. *International Journal of Social Robotics*, *6*(2), 299–310. https://doi.org/10.1007/s12369-013-0225-8

Frijda, N. H. (1986). *The emotions*. Cambridge University Press.

Frith, C. D., & Singer, T. (2008). The role of social cognition in decision making. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1511), 3875–3886. https://doi.org/10.1098/rstb.2008.0156

Frith, U., & Frith, C. (2001). The Biological Basis of Social Interaction. *Current Directions in Psychological Science*, *10*(5), 151–155. https://doi.org/10.1111/1467-8721.00137

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*(6), 933–948. https://doi.org/10.1037/a0029709

Gasser, G. (2021). The Dawn of Social Robots: Anthropological and Ethical Issues. *Minds and Machines*, *31*(3), 329–336. https://doi.org/10.1007/s11023-021-09572-9

George, J. M., & Dane, E. (2016). Affect, emotion, and decision making. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2016.06.004

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–179. https://doi.org/10.1006/jtbi.2000.2111

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 55–60. https://doi.org/10.1109/ROMAN.2003.1251796

Graaf, M. M. A. de, Allouch, S. B., & van Dijk, J. A. G. M. (2016). Long-Term Acceptance of Social Robots in Domestic Environments: Insights from a User's Perspective. *In: AAAI 2016 Spring Symposium on "Enabling Computing Research in Socially Intelligent Human–Robot Interaction: A Community-Driven Modular Research Platform", Palo Alto, CA, USA*, 96–103.

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.12504

Grossman, R. B., Zane, E., Mertens, J., & Mitchell, T. (2019). Facetime vs. Screentime: Gaze Patterns to Live and Video Social Stimuli in Adolescents with ASD. *Scientific Reports*, *9*(1), 12643. https://doi.org/10.1038/s41598-019-49039-7

Guerin, B. (1986). Mere presence effects in humans: A review. *Journal of Experimental Social Psychology*, *22*(1), 38–77. https://doi.org/10.1016/0022-1031(86)90040-5

Gunnthorsdottir, A., Houser, D., & McCabe, K. (2007). Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization*, *62*(2), 304–315. https://doi.org/10.1016/j.jebo.2005.03.008

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, 493–500. https://doi.org/10.1109/ROMAN.2016.7745163

Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology*, *14*(2), 233–238. https://doi.org/10.1016/j.conb.2004.03.010

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*. https://doi.org/10.1177/0018720811417254

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional Contagion. *Current Directions in Psychological Science*, *2*(3), 96–100. https://doi.org/10.1111/1467-8721.ep10770953

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., & Sagerer, G. (2009). Understanding Social Robots. *2009 Second International Conferences on Advances in Computer-Human Interactions*, 169–174.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The Weirdest People in the World. *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social Cognition in the Age of Human–Robot Interaction. *Trends in Neurosciences*, S0166223620300734. https://doi.org/10.1016/j.tins.2020.03.013

Henschel, A., Laban, G., & Cross, E. S. (2021). What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports*, *2*(1), 9–19. https://doi.org/10.1007/s43154-020-00035-0

Hess, U., Kafetsios, K., Mauersberger, H., Blaison, C., & Kessler, C.-L. (2016). Signal and Noise in the Perception of Facial Emotion Expressions: From Labs to Life. *Personality and Social Psychology Bulletin*, *42*(8), 1092–1110. https://doi.org/10.1177/0146167216651851

Hoegen, R., van der Schalk, J., Lucas, G., & Gratch, J. (2018). The impact of agent facial mimicry on social behavior in a prisoner's dilemma. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 275–280. https://doi.org/10.1145/3267851.3267911

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*, 721. https://doi.org/10.3389/fpsyg.2020.00721

Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: The scope and limits of attributing socialness to artificial agents: Socialness attribution and artificial agents. *Annals of the New York Academy of Sciences*, *1426*(1), 93–110. https://doi.org/10.1111/nyas.13727

Hortensius, R., Hekele, F., & Cross, E. S. (2018). The Perception of Emotion in Artificial Agents. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(4), 852–864. https://doi.org/10.1109/TCDS.2018.2826921

Hortensius, R., Kent, M., Darda, K. M., Jastrzab, L., Koldewyn, K., Ramsey, R., & Cross, E. S. (2021). Exploring the relationship between anthropomorphism and theory-of-mind in brain and behaviour. *Human Brain Mapping*, *42*(13), 4224–4241. https://doi.org/10.1002/hbm.25542

Houser, D., & Kurzban, R. (2002). Revisiting Kindness and Confusion in Public Goods Experiments. *The American Economic Review*, *92*(4), 1062–1069.

Hsieh, T.-Y., Chaudhury, B., & Cross, E. S. (2020). *Human-robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/q6pv7

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

Innes, J. M., & Morrison, B. W. (2020). Experimental Studies of Human–Robot Interaction: Threats to Valid Interpretation from Methodological Constraints Associated with Experimental Manipulations. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00671-8

Ishiguro, H. (2006). Android science: Conscious and subconscious recognition. *Connection Science*. https://doi.org/10.1080/09540090600873953

Ito, H., & Tanimoto, J. (2018). Scaling the phase-planes of social dilemma strengths shows game-class changes in the five rules governing the evolution of cooperation. *Royal Society Open Science*, *5*(10), 181085. https://doi.org/10.1098/rsos.181085

Jamaludin, S., Azmir, N. A., Mohamad Ayob, A. F., & Zainal, N. (2020). COVID-19 exit strategy: Transitioning towards a new normal. *Annals of Medicine and Surgery*, *59*, 165–170. https://doi.org/10.1016/j.amsu.2020.09.046

Janssen, M. A. (2008). Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents. *Journal of Economic Behavior & Organization*, *65*(3-4), 458–471. https://doi.org/10.1016/j.jebo.2006.02.004

Jost, C., Pévédic, B. L., Belpaeme, T., Bethel, C., Chrysostomou, D., Crook, N., Grandgeorge, M., & Mirnig, N. (2020). *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Nature.

Kahn, P. H., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2004). Robotic Pets in the Lives of Preschool Children. *Interaction Studies*, *7*(3), 405–436. https://doi.org/10.1075/is.7.3.13kah

Kayukawa, Y., Takahashi, Y., Tsujimoto, T., Terada, K., & Inoue, H. (2017). Influence of emotional expression of real humanoid robot to human decision-making. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015598

Kim, S. (Sam), Kim, J., Badu-Baiden, F., Giroux, M., & Choi, Y. (2021). Preference for robot service or human service in hotels? Impacts of the COVID-19 pandemic. *International Journal of Hospitality Management*, *93*, 102795. https://doi.org/10.1016/j.ijhm.2020.102795

Kjell, O. N. E., & Thompson, S. (2013). Exploring the impact of positive and negative emotions on cooperative behaviour in a Prisoner's Dilemma Game. *PeerJ*, *1*(2000), e231. https://doi.org/10.7717/peerj.231

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kompatsiari, K., Pérez-Osorio, J., De Tommaso, D., Metta, G., & Wykowska, A. (2018). Neuroscientifically-Grounded Research for Improved Human-Robot Interaction. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3403–3408. https://doi.org/10.1109/IROS.2018.8594441

Konrath, S., Corneille, O., Bushman, B. J., & Luminet, O. (2014). The Relationship Between Narcissistic Exploitativeness, Dispositional Empathy, and Emotion Recognition Abilities. *Journal of Nonverbal Behavior*, *38*(1), 129–143. https://doi.org/10.1007/s10919-013-0164-y

Kopelman, S., Rosette, A. S., & Thompson, L. (2006). The three faces of Eve: Strategic displays of positive, negative, and neutral emotions in negotiations. *Organizational Behavior and Human Decision Processes*, *99*(1), 81–101. https://doi.org/10.1016/j.obhdp.2005.08.003

Kopp, C., Korb, K. B., & Mills, B. I. (2018). Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to ‘fake news’. *PLOS ONE*, *13*(11), e0207383. https://doi.org/10.1371/journal.pone.0207383

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE*, *3*(7). https://doi.org/10.1371/journal.pone.0002597

Kwak, S. S., Kim, Y., Kim, E., Shin, C., & Cho, K. (2013). What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. *2013 IEEE RO-MAN*, 180–185. https://doi.org/10.1109/ROMAN.2013.6628441

Lakens, D. (2014a). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2014b). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.2023

LeBeau, B. (2019). *Power Analysis by Simulation using R and simglm* [Preprint]. https://doi.org/10.17077/f7kk-6w7f

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, *64*(10), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

Lee, M., Ahn, H. S., Kwon, S. K., & Kim, S. (2018). Cooperative and Competitive Contextual Effects on Social Cognitive and Empathic Neural Responses. *Frontiers in Human Neuroscience*, *12*, 218. https://doi.org/10.3389/fnhum.2018.00218

Lee, S. A., & Liang, Y. (Jake). (2016). The Role of Reciprocity in Verbally Persuasive Robots. *Cyberpsychology, Behavior, and Social Networking*, *19*(8), 524–527. https://doi.org/10.1089/cyber.2016.0124

Lefkeli, D., Ozbay, Y., Gürhan-Canli, Z., & Eskenazi, T. (2021). Competing with or Against Cozmo, the Robot: Influence of Interaction Context and Outcome on Mind Perception. *International Journal of Social Robotics*, *13*. https://doi.org/10.1007/s12369-020-00668-3

Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics*, *5*(2), 291–308. https://doi.org/10.1007/s12369-013-0178-y

Leite, I., Pereira, A., Martinho, C., & Paiva, A. (2008). Are emotional robots more fun to play with? *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 77–82. https://doi.org/10.1109/ROMAN.2008.4600646

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin*, *137*(5), 834–855. https://doi.org/10.1037/a0024244

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. (R package version 1.4.7.) [Computer software]. https://CRAN.R-project.org/package=emmeans

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Lim, V., Rooksby, M., & Cross, E. S. (2020). Social Robots on a Global Stage: Establishing a Role for Culture During Human–Robot Interaction. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00710-4

Liu, O., Rakita, D., Mutlu, B., & Gleicher, M. (2017). Understanding human-robot interaction in virtual reality. *2017 26th IEEE International Symposium on*

*Robot and Human Interactive Communication (RO-MAN)*, 751–757. https://doi.org/10.1109/ROMAN.2017.8172387

Lyons, J. B., Wynne, K. T., Mahoney, S., & Roebke, M. A. (2019). Trust and human-machine teaming: A qualitative study. *In Artificial Intelligence for the Internet of Everything*, 101–116.

Magee, J. C., & Langner, C. A. (2008). How personalized and socialized power motivation facilitate antisocial and prosocial decision-making. *Journal of Research in Personality*, *42*(6), 1547–1559. https://doi.org/10.1016/j.jrp.2008.07.009

Maggioni, M. A., & Rossignoli, D. (2021). If it Looks like a Human and Speaks like a Human ... Dialogue and cooperation in human-robot interactions. *Dialogue and Cooperation in Human-Robot Interactions*. http://arxiv.org/abs/2104.11652

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*. https://doi.org/10.1016/j.futures.2017.03.006

Manstead, A. S. R., & Fischer, A. H. (2001). Social appraisal: The social world as object of and influence on appraisal processes. In *Series in affective science. Appraisal processes in emotion: Theory, methods, research*. https://doi.org/10.1002/mde.1235

Martin, R., McKenzie, K., Metcalfe, D., Pollet, T., & McCarty, K. (2019). A preliminary investigation into the relationship between empathy, autistic like traits and emotion recognition. *Personality and Individual Differences*, *137*, 12–16. https://doi.org/10.1016/j.paid.2018.07.039

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of*

*Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mobbs, D., Weiskopf, N., Lau, H. C., Featherstone, E., Dolan, R. J., & Frith, C. D. (2006). *The Kuleshov Effect: The influence of contextual framing on emotional attributions*. 12.

Moisan, F., ten Brincke, R., Murphy, R. O., & Gonzalez, C. (2018). Not all Prisoner's Dilemma games are equal: Incentives, social preferences, and cooperation. *Decision*. https://doi.org/10.1037/dec0000079

Mokros, A., Menner, B., Eisenbarth, H., Alpers, G. W., Lange, K. W., & Osterheider, M. (2008). Diminished cooperativeness of psychopaths in a prisoner's dilemma game yields higher rewards. *Journal of Abnormal Psychology*, *117*(2), 406–413. https://doi.org/10.1037/0021-843X.117.2.406

Montagne, B., Kessels, R. P. C., De Haan, E. H. F., & Perrett, D. I. (2007). The Emotion Recognition Task: A Paradigm to Measure the Perception of Facial Emotional Expressions at Different Intensities. *Perceptual and Motor Skills*, *104*(2), 589–598. https://doi.org/10.2466/pms.104.2.589-598

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Munafò, M. R. (2016). Open Science and Research Reproducibility. *Ecancermedicalscience*, *10*(ed56). https://doi.org/10.3332/ecancer.2016.ed56

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis,

J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

Murphy, R. O., & Ackermann, K. A. (2014). Social Value Orientation: Theoretical and Measurement Issues in the Study of Social Preferences. *Personality and Social Psychology Review*, *18*(1), 13–41. https://doi.org/10.1177/1088868313501745

Murphy, R. O., & Ackermann, K. A. (2015). Social preferences, positive expectations, and trust based cooperation. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2015.06.001

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring Social Value Orientation. *Ssrn*, *6*(8), 771–781. https://doi.org/10.2139/ssrn.1804189

Nishio, S., Ogawa, K., Kanakogi, Y., Itakura, S., & Ishiguro, H. (2018). Do robot appearance and speech affect people's attitude? Evaluation through the ultimatum game. In *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids*. https://doi.org/10.1007/978-981-10-8702-8_16

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2008). Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward tobots. *IEEE Transactions on Robotics*, *24*(2), 442–451. https://doi.org/10.1109/TRO.2007.914004

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., & Lakens, D. (2014). Registered Reports A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*, 137. https://doi.org/10.1027/1864-9335/a000192

Novak, D., Nagle, A., Keller, U., & Riener, R. (2014). Increasing motivation in robot-aided arm rehabilitation with competitive and cooperative gameplay. *Journal of NeuroEngineering and Rehabilitation*, *11*(1), 64. https://doi.org/10.1186/1743-0003-11-64

Odekerken-Schröder, G., Mele, C., Russo-Spena, T., Mahr, D., & Ruggiero, A. (2020). Mitigating loneliness with companion robots in the COVID-19 pandemic and beyond: An integrative framework and research agenda. *Journal of Service Management*, *31*(6), 1149–1162. https://doi.org/10.1108/JOSM-05-2020-0148

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Open Science Collaboration. (2017). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 1–21). NY: Wiley.

Osborne, M. (2004). *An introduction to game theory* (Vol. 3). New York: Oxford university press.

Paeng, E., Wu, J., & Jr, J. C. B. (2016). Human-Robot Trust and Cooperation Through a Game Theoretic Framework. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 4246–4247.

Parsons, T. D. (2015). Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences.

*Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00660

Pender, J. (2018). Preparing for a Robot Future? Social Professions, Social Robotics and the Challenges Ahead. *Irish Journal of Applied Social Studies*, *18*(1). https://doi.org/10.21427/D7472M

Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S., & Szolnoki, A. (2017). Statistical physics of human cooperation. *Physics Reports*, *687*, 1–51. https://doi.org/10.1016/j.physrep.2017.05.004

Perc, M., Ozer, M., & Hojnik, J. (2019). Social and juristic challenges of artificial intelligence. *Palgrave Communications*, *5*(1), 61. https://doi.org/10.1057/s41599-019-0278-x

Perugia, G., Paetzel-Prüsmann, M., & Castellano, G. (2020). *On the Role of Personality and Empathy in Human-Human, Human-Agent, and Human-Robot Mimicry*. 120–131.

Peshkovskaya, A. G., Babkina, T. S., Myagkov, M. G., Kulikov, I. A., Ekshova, K. V., & Harriff, K. (2017). The socialization effect on decision making in the Prisoner's Dilemma game: An eye-tracking study. *PLOS ONE*, *12*(4). https://doi.org/10.1371/journal.pone.0175492

Phillips, M. R., McAuliff, B. D., Bull Kovera, M., & Cutler, B. L. (2000). 'Double-blind photoarray administration as a safeguard against investigator bias': Correction to Phillips et al. (1999). *Journal of Applied Psychology*, *85*(2), 304–304. https://doi.org/10.1037/h0087870

Pipitone, A., Geraci, A., D'Amico, A., Seidita, V., & Chella, A. (2021). Robot's Inner Speech Effects on Trust and Anthropomorphic Cues in Human-Robot Cooperation. *ArXiv:2109.09388 [Cs]*. http://arxiv.org/abs/2109.09388

Pletzer, J. L., Balliet, D., Joireman, J., Kuhlman, D. M., Voelpel, S. C., Van Lange, P. A. M., & Back, M. (2018). Social Value Orientation, Expectations, and Cooperation in Social Dilemmas: A Meta-Analysis. *European Journal of Personality*, *32*(1), 62–83. https://doi.org/10.1002/per.2139

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199.

Pothos, E. M., Perry, G., Corr, P. J., Matthew, M. R., & Busemeyer, J. R. (2011a). Understanding cooperation in the Prisoner's Dilemma game. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2010.05.002

Pothos, E. M., Perry, G., Corr, P. J., Matthew, M. R., & Busemeyer, J. R. (2011b). Understanding cooperation in the Prisoner's Dilemma game. *Personality and Individual Differences*, *51*(3), 210–215. https://doi.org/10.1016/j.paid.2010.05.002

Prochazkova, E., & Kret, M. E. (2017). Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2017.05.013

R Core Team. (2020). *R: a language and environment for statistical computing [Internet]* (4.0.0) [Computer software]. Foundation for Statistical Computing.

Ramsey, R. (2021). A Call for Greater Modesty in Psychology and Cognitive Neuroscience. *Collabra: Psychology*, *7*(1), 24091. https://doi.org/10.1525/collabra.24091

Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1108243108

Rand, D. G., Newman, G. E., & Wurzbacher, O. M. (2014). Social Context and the Dynamics of Cooperative Choice. *Journal of Behavioral Decision Making*.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2013.06.003

Rapoport, A. (1967). A note on the "index of cooperation" for Prisoner's Dilemma. *Journal of Conflict Resolution*. https://doi.org/10.1177/002200276701100108

Rapoport, A., & Chammah, A. M. (1967). Prisoner's Dilemma: A Study in Conflict and Cooperation. *American Political Science Review*. https://doi.org/10.1017/s000305540013240x

Rick, S., & Loewenstein, G. F. (2008). The Role of Emotion in Economic Behavior. In *Handbook of emotions* (3rd ed., pp. 138–158). http://www.ssrn.com/abstract=954862

Riddoch, K. A., & Cross, Emily. S. (2021). "Hit the Robot on the Head With This Mallet" – Making a Case for Including More Open Questions in HRI Research. *Frontiers in Robotics and AI*, *8*, 603510. https://doi.org/10.3389/frobt.2021.603510

Riek, L. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Riek, L. D., Adams, A., & Robinson, P. (2011). Exposure to Cinematic Depictions of Robots and Attitudes Towards Them. *IEEE Conference on Human-Robot Interactions, Workshop on Expectations and Intuitive Human-Robot Interaction (Vol. 6)*.

Robert, L. (2018). Personality in the human robot interaction literature: A review and brief critique. *Robert, LP (2018). Personality in the Human*

*Robot Interaction Literature: A Review and Brief Critique, Proceedings of the 24th Americas Conference on Information Systems*.

Roseman, I. J., & Smith, C. A. (2001). Appraisal theory. In *Appraisal processes in emotion: Theory, methods, research* (pp. 3–19). Oxford University Press.

Rosenthal-von der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., Brand, M., & Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior*, *33*, 201–212. https://doi.org/10.1016/j.chb.2014.01.004

Ruijten, P. A. M., Haans, A., Ham, J., & Midden, C. J. H. (2019). Perceived Human-Likeness of Social Robots: Testing the Rasch Model as a Method for Measuring Anthropomorphism. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-019-00516-z

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (Ro-MAN 2015)*. https://doi.org/10.1145/2696454.2696497

Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner's Dilemma and the Ultimatum Game. *International Journal of Social Robotics*, *8*(2), 303–317. https://doi.org/10.1007/s12369-015-0323-x

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science*, *318*(5850), 598 LP – 602.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered

Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007468. https://doi.org/10.1177/25152459211007467

Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, *78*, 102253. https://doi.org/10.1016/j.joep.2020.102253

Schrempf, O. C., Hanebeck, U. D., Schmid, A. J., & Worn, H. (2005). A novel approach to proactive human-robot cooperation. *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, 555–560. https://doi.org/10.1109/ROMAN.2005.1513838

Schwab, K. (2016). *The Fourth Industrial Revolution*. Crown.

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., & Young, J. E. (2015). Poor Thing! Would You Feel Sorry For a Simulated Robot?: A Comparison of Empathy toward a Physical and a Simulated Robot. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. https://doi.org/10.1145/2696454.2696471

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, *46*(1), 561–584. https://doi.org/10.1146/annurev.ps.46.020195.003021

Stavrova, O., & Meckel, A. (2017). Perceiving emotion in non-social targets: The effect of trait empathy on emotional contagion through art. *Motivation and Emotion*, *41*(4), 492–509. https://doi.org/10.1007/s11031-017-9619-5

Stock-Homburg, R. (2021). Survey of Emotions in Human–Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-021-00778-6

Stroessner, S. J., & Benitez, J. (2018). The Social Perception of Humanoid and Non-Humanoid Robots: Effects of Gendered and Machinelike Features.

*International Journal of Social Robotics*. https://doi.org/10.1007/s12369-018-0502-7

Stubbs, K., Hinds, P. J., & Wettergreen, D. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, *22*(2), 42–50. https://doi.org/10.1109/MIS.2007.21.

Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., & Hoekstra, R. A. (2013). Empathy and emotion recognition in people with autism, first-degree relatives, and controls. *Neuropsychologia*, *51*(1), 98–105. https://doi.org/10.1016/j.neuropsychologia.2012.11.013

Sung, J., Christensen, H. I., & Grinter, R. E. (2009). Robots in the wild: Understanding long-term use. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction - HRI '09*, 45. https://doi.org/10.1145/1514095.1514106

Swanson, D. L. (1996). Neoclassical Economic Theory, Executive Control, and Organizational Outcomes. *Human Relations*, *49*(6), 735–756. https://doi.org/10.1177/001872679604900602

Syrdal, D. S., Dautenhahn, K., Koay, K., & Walters, M. L. (2009). The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and Emergent Behaviour and Complex Systems*. https://doi.org/10.1157/13126291

Tan, X. Z., Vázquez, M., Carter, E. J., Morales, C. G., & Steinfeld, A. (2018). Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 169–177. https://doi.org/10.1145/3171221.3171247

Terada, K., & Takeuchi, C. (2017). Emotional expression in simple line drawings of a robot's face leads to higher offers in the ultimatum game. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2017.00724

Torta, E., Van Dijk, E., Ruijten, P. A. M., & Cuijpers, R. H. (2013). The ultimatum game as measurement tool for anthropomorphism in human-robot interaction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-02675-6_21

Tsiourti, C., Weiss, A., Wac, K., & Vincze, M. (2019). Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In)Congruence on Emotion Recognition and Attitudes Towards Robots. *International Journal of Social Robotics*, *11*(4), 555–573. https://doi.org/10.1007/s12369-019-00524-z

Tulk, S., & Wiese, E. (2018). *Social Decision Making with Humans and Robots: Trust and Approachability Mediate Economic Decision Making* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/4aj8v

Van der Putte, D., Boumans, R., Neerincx, M., Rikkert, M. O., & De Mul, M. (2019). *A Social Robot for Autonomous Health Data Acquisition Among Hospitalized Patients: An Exploratory Field Study.* 658–659. https://doi.org/10.1109/HRI.2019.8673280.

Van Dijk, E., Van Beest, I., Van Kleef, G. A., & Lelieveld, G. J. (2018). Communication of anger versus disappointment in bargaining and the moderating role of power. *Journal of Behavioral Decision Making, January*, 632-643. https://doi.org/10.1002/bdm.2079

Van Dijk, E., Van Kleef, G. A., Steinel, W., & Beest, I. (2008). A Social Functional Approach to Emotions in Bargaining: When Communicating

Anger Pays and When it Backfires. *Journal of Personality and Social Psychology*, 600–614.

Van Kleef, G. A. (2009). How emotions regulate social life. *Current Directions in Psychology*, *18*(3), 184–188. https://doi.org/10.1111/j.1467-8721.2009.01633.x

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004). The Interpersonal Effects of Emotions in Negotiations: A Motivated Information Processing Approach. *Journal of Personality and Social Psychology*, *87*(4), 510–528. https://doi.org/10.1037/0022-3514.87.4.510

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. In *Advances in Experimental Social Psychology* (Vol. 42, pp. 45–96). Elsevier. https://doi.org/10.1016/S0065-2601(10)42002-X

Van Kleef, G. A., De Dreu, C. K. W., Pietroni, D., & Manstead, A. S. R. (2006). Power and Emotion in Negotiation: Power Moderates the Interpersonal Effects of Anger and Happiness on Concession Making,”. *European Journal of Social Psychology,* 557–581.

Van Lange, P. A. M., Joireman, J., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125–141. https://doi.org/10.1016/J.OBHDP.2012.11.003

van Straten, C. L., Peter, J., & Kühne, R. (2020). Child–Robot Relationship Formation: A Narrative Review of Empirical Research. *International Journal of Social Robotics*, *12*(2), 325–344. https://doi.org/10.1007/s12369-019-00569-0

Villani, V., Capelli, B., & Sabattini, L. (2018). Use of Virtual Reality for the Evaluation of Human-Robot Interaction Systems in Complex Scenarios. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 422–427. https://doi.org/10.1109/ROMAN.2018.8525738

Viola, T. W., Niederauer, J. P. O., Kluwe-Schiavon, B., Sanvicente-Vieira, B., & Grassi-Oliveira, R. (2019). Cocaine use disorder in females is associated with altered social decision-making: A study with the prisoner's dilemma and the ultimatum game. *BMC Psychiatry*, *19*(1), 211. https://doi.org/10.1186/s12888-019-2198-0

Vuoskoski, J. K., & Eerola, T. (2012). Empathy contributes to the intensity of music-induced emotions. *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC)*, 1112–1113.

Walters, M. L., Dautenhahn, K., te Boekhorst, R., Koay, K. L., Kaouri, C., Woods, S., Nehaniv, C., Lee, D., & Werry, I. (2005). The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, 347–352. https://doi.org/10.1109/ROMAN.2005.1513803

Wu, J., Paeng, E., Linder, K., Valdesolo, P., & Boerkoel, J. C. (2016). Trust and Cooperation in Human-Robot Decision Making. *The 2016 AAAI Fall Symposium*, *16*(1), 110–116. https://doi.org/10.1111/j.1835-2561.2006.tb00045.x

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1693), 20150375. https://doi.org/10.1098/rstb.2015.0375

Xu, M., David, J. M., & Kim, S. H. (2018). The Fourth Industrial Revolution: Opportunities and Challenges. *International Journal of Financial Research*, *9*(2), 90. https://doi.org/10.5430/ijfr.v9n2p90

Yang, G.-Z., J. Nelson, B., Murphy, R. R., Choset, H., Christensen, H., H. Collins, S., Dario, P., Goldberg, K., Ikuta, K., Jacobstein, N., Kragic, D., Taylor, R. H., & McNutt, M. (2020). Combating COVID-19—The role of robotics in managing public health and infectious diseases. *Science Robotics*, *5*(40), eabb5589. https://doi.org/10.1126/scirobotics.abb5589

Yu Ogura, Aikawa, H., Shimomura, K., Kondo, H., Morishima, A., Hun-ok Lim, & Takanishi, A. (2006). Development of a new humanoid robot WABIAN-2. *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, 76–81. https://doi.org/10.1109/ROBOT.2006.1641164

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–1647. https://doi.org/10.1037/a0037679

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-014-0267-6