Title	Studies on Individuality of Interactive Robots
Author(s)	Luo, Liangyi
Citation	大阪大学, 2022, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/89649
rights	
Note	

Studies on Individuality of Interactive Robots

LIANGYI LUO

SEPTEMBER 2022

Studies on Individuality of Interactive Robots

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN ENGINEERING

BY

LIANGYI LUO

SEPTEMBER 2022

Abstract

This is a thesis on creating interactive robots with individuality. Individuality is a fundamental property of personality. A robot with individuality can be distinguished from other robots of the same kind by their personality traits. There are several challenges to address to create robots with individuality. Robots, if useful enough, will be mass-produced like computers and smartphones. Creating robots with individuality entails turning physically identical robots into unique individuals. In order to leverage the benefits of robots' personalities, they will be created to meet some requirements: being likeable, pleasant, human-like, etc. with traits conducive to their jobs. Robots created to meet these requirements will have similar personalities. Robots come in a variety of designs for a variety of applications. Their images and behaviour might not match those of humans, and hence human personality traits might not apply. These challenges boil down to three problems to investigate. First, the problem about the possibility of individuality, formulated as a *CAN* question: *can physically identical robots be created with individuality that consists of personality traits that distinguish them from one another, under the constraints that the traits are similar?* Second, the problem about the composition of individuality, formulated as a *WHAT* question: *what personality traits compose the individuality of a type of robot?* Third, the problem about the implementation of individuality, formulated as a *HOW* question: *how to implement individuality?* This dissertation presents the three studies conducted to answer these three questions.

In the first study, I verified the possibility by measuring the similarity and differences between two similar robot personalities, modelled after two humans using an eye movement model, installed on the same android, and evaluated by human subjects using human personality dimensions. It was understood that when an android gazes or averts their eyes following the patterns of two humans with similar personalities, the android can appear as two similar but distinguishable robots. However, the android was a special type of robot built to resemble a female human being, which was not representative of robots. Human personality dimensions might not apply to other types of robots.

In the second study, I contributed a method for selecting personality dimensions applicable to a type of robot to structure their individuality with traits corresponding to the selected dimensions. It functions by testing the reports of a group of users on how well a type of robot can approximate human personalities by imitating multimodal behaviour of multiple humans. The experiment on the method showed that it could provide practical answers to the second research question. The experiment also revealed that the selection of personality dimensions could be user-dependent.

In the third study, I proposed a generative model of personalities to implement individuality as a solution to the third research question. The model functions by combining a robot's actions, as associated with several personality traits, according to the ordering and strength of the traits, as associated with the robots' goals. The experiment on a prototype model installed on

a humanoid robot showed that it was possible to make the robot appear as multiple distinguishable individuals by implementing multiple sets of model parameters. This result indicates that the model can be used for making physically identical robots into robots with individuality.

In summary, the first study verified the possibility of creating androids with individuality; the second study contributed a method to generalise the possibility to other types of robots; and the third study contributed a generative personality model to realise the possibility. According to my findings, robots with individuality can be created by considering a robot's behaviour as actions associated with several personality traits, corresponding to several applicable personality dimensions, and then combining those actions according to the ordering and strength of the traits as associated with the robot's goals.

Contents

Chapter 1

Introduction

1.1 Purposes and Motivations

A robot interacts with a person as an individual being: 'this robot', which is different from all other robots, and a separate presence, unlike 'the same voice that speaks through everyone's device'. It is only natural that the 'inside' of a robot should conform to this fact, such as an inside we call personality. Robots with personalities, i.e. robot personalities, can be more distinguishable, likeable, and human-like; and appropriate personalities facilitate collaboration [1, 2] and attenuate negative effects such as the uncanny valley [3]. A fundamental property of personality is individuality, i.e. characteristics that distinguish an individual from others of the same kind. It was true for the people in the past, and will be for the people in the future; it is true for people regardless of their origins, ethnicities, cultures, and languages; it was true for Neanderthals and Denisovans and their common ancestors with humans; it is true for animals' equivalence of personality; it is even true for aliens light-years away (should they exist). It is how personality is defined[1]. If we encountered an alien race one day and wanted to know whether the beings of this race were individuals with personalities or indistinguishable pawns of a hive, the one question we must ask was *do they have individuality?* We ask the same question to create robot personalities.

To create robot personalities is to create robots with individuality, i.e. robots with characteristics that distinguish them from one another, and the key is making a robot distinguishable from other robots by personality traits. It entails several challenges to address to create robots with individuality in significant quantities. Robots, if useful enough, will be mass-produced with fungibility like computers and smartphones. Therefore, to create robots with individuality is to turn physically identical robots into unique individuals. In order to leverage the benefits of robots' personalities, they will be created to meet some requirements: being likeable, pleasant, human-like, etc. with traits conducive to their jobs. Robots created to meet these requirements will have similar personalities. In addition, robots come in a variety of designs for a variety of applications. Their images and behaviour might not match those of humans, and hence human personality traits might not apply.

[1]Here is one definition from *APA Dictionary of Psychology*: `https://dictionary.apa.org/personality`. Definitions of personality vary but individuality is consistent across them.

1.2 Essentials in Robot Personalities Engineering

1.2.1 Key Definitions

In this dissertation, a term in the sentence that formally defines it will be italicised in bold. Italics not in bold are just for emphasis.

This work is of a novel discipline called ***robot personalities engineering***, an engineering discipline dedicated to creating robots with personalities (the plural, 'personalities', is for emphasising that it is always about engineering not one but multiple personalities) [4]. A ***robot personality*** is a robot exhibiting characteristic patterns of computation and behaviour with inter- and intra-individual differences. 'Inter-individual differences' indicate that different robot personalities will compute and behave differently in the same context. 'Intra-individual differences' indicate that a robot personality will compute and behave differently in different contexts. Inter- and intra-individual differences are described by personality traits. A ***personalty trait***, or simply ***trait***, is a unique characteristic of an individual robot specified by a ***broad trait***, a broader facet of the robot's personality, and a ***trait level***, a value that marks the level of strength of the broad trait. Every broad trait is represented by a ***personality dimension***, a dimension representing a broader aspect of personality. A ***prominent trait*** is a trait applicable to and distinctly perceptible by a user group in a certain type of robot in its typical usage scenario. (In theory, it is possible to create a robot with any traits. For instance, we can program a robotic vacuum cleaner to be full of originality. However, personality needs to be observed, especially when we intend to leverage its benefits. Users of robotic vacuum cleaners do not usually expect originality from their room sweepers. They do not necessarily notice the trait even if their robots clean their rooms in a thousand original ways. In this case, the originality of the robotic vacuum cleaner is a trait but not a prominent trait, except to users who notice it. Whether a trait is prominent is observer-dependent.) A fundamental property of personality is ***individuality***, which refers to personality traits that distinguish the individual from others of the same kind. Robot personalities are modelled by a ***generative personality model***, or ***GPM*** for the acronym, a personality model capable of generating behaviour with inter- and intra-individual differences that reflect the inter- and intra-individual differences in personalities. Every robot personality is based on a ***personality construct***, a particular build of a generative personality model with a unique set of personality parameters. Every personality construct corresponds to one and only one robot personality; even if we can make a copy of a personality construct, the two copies are one robot personality, not two, unless further development distinguishes one from the other. ***Trait capacity*** is the number of broad traits (personality dimensions) out of all that are applicable to a type of robot a generative personality model can model. A ***robot colony***—in the context of robot personalities engineering—is a population of robot personalities situated in close association. By 'situated in close association' I mean that those robots will be observed by the same people for their shared qualities or goals. The 'colony' here takes a similar meaning to that in expressions such as 'a colony of artists'. It may but does not necessarily indicate spatial proximity. For example, a colony of singer robots, albeit not necessarily working at the same place, may be observed by the same audience, and hence each of them needs to be a personality of their own. An essential property of a colony is that there is no duplication of personalities: each robot in a colony is unique. A robot colony should be distinguished from a ***robot population***, which is simply a group of robots with or without personalities. ***Colony capacity*** is the number of individual personalities a generative personality model theoretically supports[2]. It depends only on the design of the model. ***Effective colony capacity*** is the number of individual personalities distinguishable by some observers. It depends both on the design of the model and the observers. A ***personality archetype***, or simply ***archetype***, is the 'target' personality a robot personality is engineered to be. It is a blueprint or example for a personality construct. An archetype can be represented by a vector in a ***personality space***, which is the space of all possible configurations of personalities spanned by a set of personality dimensions, corresponding to a set of broad traits. The proximity between a robot personality and the corresponding archetype in the personality space is measured by ***fidelity***: high fidelity indicates high proximity. A robot expresses personality via a ***mode of expression***, or simply a ***mode***[3], which is a distinct means of expressing personality associated with a group of actuators or signalling apparatus. An ***observer*** is an agent that can perform measurements of a robot's personality. The most common observers are humans. When a significant number of observers reported their impressions of a robot personality using applicable personality assessment scales, the variation among their impressions is measured by ***consistency***: the more varied the impressions are the lower the consistency. It is impossible to discuss personality in depth without mentioning goals. A ***goal***—in the context of robot personalities engineering—is an end state a robot is operating to achieve: a purpose of the robot's activities or endeavours. The most important goal of a robot's existence is the robot's ***mission***. A robot's *mission* is not simply their task but a 'life's' purpose that is better described as their 'one true mission'. The Japanese word for it is *shimei*.

1.2.2 General Approach

To create a robot personality, one could follow these steps, provided that they have available units of the type of robot they intend to work on and a compatible generative personality model. First, specify an archetype that

[2]I called it *colony size capacity* in the original publication [4], but now think it sounds weird, and cannot comprehend why the past me did not think the same. Nevertheless, *colony size* simply means the size of the colony and is too obvious to merit a formal definition. And colony capacity will never be referring to the maximum number of colonies a GPM supports since that does not depend on a GPM.

[3]It may be called a 'modality' in another context. Here it is not because the nuances are different. A mode of expression is not necessarily *how* a robot expresses personality; the robot, albeit equipped with modes, may not express personality at all, or it may *happen* to express personality (spontaneously).

represents the personality to create. It can be a set of desirable personality traits or even a human being. Second, create a personality construct based on the archetype using the generative personality model and install it on a unit of the robot. Third, minimise the discrepancy between the observed robot personality and the archetype with the help of some observers (e.g. the robot's users). When the discrepancy is small enough, the robot personality is complete.

1.2.3 Generative Personality Models

A generative personality model, as the term suggests, is capable of generating behaviour with inter- and intra-individual differences, as opposed to a descriptive model that only describes inter- and intra-individual differences in behaviour. Although we can expect some structural homogeneity between the two breeds, to construct generative personality models we need some understanding of how inter- and intra-individual differences in personalities lead to inter- and intra-individual differences in behaviour, which is yet to be fully explored in personality psychology [5]. Nevertheless, roboticists' solutions do not have to be completely faithful to human conditions. Whatever works on robots may suffice.

I postulate that a generative personality model has these key elements: a personality structure that encodes inter- and intra-individual differences in personalities as existing descriptive models do, components that generate behaviour according to situations following some general behavioural tendencies corresponding to certain broad traits, and mechanisms that modulate the generated behaviour to reflect the inter- and intra-individual differences.

1.2.4 Model Evaluation and Optimisation

In robot personalities engineering, we consider four criteria for model evaluation: fidelity, consistency, trait capacity, and colony capacity. *Fidelity* and *consistency* are criteria for evaluating respectively how accurately and consistently a generative personality model can present personality archetypes or how accurately and consistently a personality construct presents its corresponding archetype. They correspond to the functions to optimise in the last step of the general approach. There are two cases of optimisation: optimising a generative personality model and optimising a personality construct. Let $\mathbf{M}$ denote a generative personality model:

$$\mathbf{M} = (P, C), \tag{1.1}$$

where P denotes the model parameters, and C, the internal configuration of the model. A personality construct is a particular build of a generative personality model, with a unique set of personality parameters, which is represented by an archetype $\boldsymbol{a}$. Therefore, a personality construct based on $\mathbf{M}$ can be expressed as $\mathbf{M}(\boldsymbol{a}, C)$. Let $\bar{o}$ be a function that represents an average observer, which does not exist in real life but can be simulated and approximated by averaging over a sufficiently large number of observers on how they perceive personalities. The function takes a personality construct as the input and yields the observed personality, here denoted as $\boldsymbol{a}^{'}$, as the output:

$$\boldsymbol{a}^{'} = \bar{o}(\mathbf{M}(\boldsymbol{a}, C)). \tag{1.2}$$

Then the fidelity of $\mathbf{M}$ can be expressed as

$$F_{\mathbf{M}}(C) = -\sum_{\boldsymbol{a} \in A} dis(\boldsymbol{a}, \boldsymbol{a}^{'}), \tag{1.3}$$

where *dis* is a function to compute the discrepancy or distance between $\boldsymbol{a}$ and $\boldsymbol{a}^{'}$, and A, the set of all possible archetypes. The definition of the discrepancy function depends on how $\boldsymbol{a}$ and $\boldsymbol{a}^{'}$ are defined. When they are vectors in a personality space, the fidelity function becomes

$$F_{\mathbf{M}}(C) = -\sum_{\boldsymbol{a} \in A} |\boldsymbol{a}^{'} - \boldsymbol{a}|. \tag{1.4}$$

To optimise $\mathbf{M}$, we maximise $F_{\mathbf{M}}(C)$, subject to model design constraints on C (same for all C hereafter), which can be achieved by minimising its negative, that is the sum of discrepancies between the observed robot personalities and the corresponding archetypes:

$$\underset{C}{\operatorname{argmin}} \sum_{\boldsymbol{a} \in A} dis(\boldsymbol{a}, \bar{o}(\mathbf{M}(\boldsymbol{a}, C))), \tag{1.5}$$

that is to find a model configuration that can represent archetypes as *faithfully* as possible (and hence the term 'fidelity'). Since there can be infinite possible archetypes, we usually get a sufficiently large sample of archetypes $A^{'}$ to approximate $F_{\mathbf{M}}(C)$; since $\bar{o}$ can only be simulated using a sample of observers $O^{'}$ of all potential observers O (e.g. the human race), each member of which with a separate opinion on an archetype, meaning that each $\boldsymbol{a}^{'}$ represents a cluster of opinions on $\boldsymbol{a}$ corresponding to multiple discrepancy measurements relative to $\boldsymbol{a}$, the actual function to minimise,

$$\underset{C}{\operatorname{argmin}} \sum_{\boldsymbol{a} \in A'} \sum_{o \in O'} dis(\boldsymbol{a}, o(\mathbf{M}(\boldsymbol{a}, C))), \tag{1.6}$$

allows us to also maximise the consistency of the model at the same time. Considering that there are individual differences in personality perception, the users of a robot, denoted as a set of specific observers $\{o_1, o_2, o_3, ..., o_u\}$, do not necessarily match how the hypothetical average observer nor the sample to approximate it perceives robot personalities, and hence, to match the observed robot personality with the archetype for the specific users of a specific robot R, it is necessary to further optimise the fidelity of the personality construct of the robot $\mathbf{M}(P_R, C)$ by

$$F_R(P_R, C) = -\sum_{i=1}^{u} dis(\boldsymbol{a}_R, \boldsymbol{a}_i^{'}), \tag{1.7}$$

where P_R is the personality parameters, subject to some constraints on personality parameters (same for all P_R hereafter), of the personality construct of R, $\boldsymbol{a}_R$ the archetype the robot is based on, and

$$\boldsymbol{a}_i^{'} = o_i(\mathbf{M}(P_R, C)). \tag{1.8}$$

As before, to optimise the fidelity is to minimise the discrepancy:

$$\underset{P_R, C}{\operatorname{argmin}} \sum_{i=1}^{u} dis(\boldsymbol{a}_R, o_i(\mathbf{M}(P_R, C))), \tag{1.9}$$

in which case, we can start the optimisation with $P_R = \boldsymbol{a}_R$, but the two may not necessarily end up the same; the choice of C, on the other hand, is trickier. To minimise the discrepancy, we also need to maximise the consistency by minimising the variation of opinions on the robot among the observers by adjusting the GPM's internal configuration:

$$\underset{C}{\operatorname{argmin}}\, var(\{\boldsymbol{a}_1^{'}, \boldsymbol{a}_2^{'}, \boldsymbol{a}_3^{'}, ..., \boldsymbol{a}_u^{'}\}), \tag{1.10}$$

where *var* is a function to measure the variation (e.g. variance) among a set of observations. We can insert the optimised model configuration $\hat{C}$ into the previous function (1.9) to minimise the discrepancy and thus maximise the fidelity for R with

$$\underset{P_R}{\operatorname{argmin}} \sum_{i=1}^{u} dis(\boldsymbol{a}_R, o_i(\mathbf{M}(P_R, \hat{C}))). \tag{1.11}$$

Trait capacity is the number of broad traits (personality dimensions), out of all that are applicable to a type of robot, a model can model. For example, Völkel et al. have identified 10 personality dimensions—corresponding to 10 broad traits—applicable to speech-based conversational agents [6]. If a generative personality model for the said agents can generate behaviour for 7 of them, the trait capacity is 7 out of 10 or 70 per cent. Usually, we can expand the capacity of a model as long as we understand the behavioural tendencies described by a broad trait well enough to implement them.

Colony capacity is the number of individual personalities a model supports. It is a direct indicator of whether a generative personality model can enable us to engineer mass-produced, physically identical robots into robot personalities of any significant quantities. In theory, any models that support continuous trait levels have infinite capacity. In reality, the effective capacity may be much lower for a number of reasons such as design constraints, hardware limitations of robots, and the complexity of personality perception [7], which is why we need to consider *effective colony capacity* as well. Effective colony capacity is the number of individual personalities distinguishable by a group of observers; it is hence observer-dependent. In theory, a GPM with high fidelity and consistency should also have a high effective colony capacity. Therefore, optimisation applies.

1.2.5 Personality Measures

Personalities, those of humans or robots, are measured using descriptive personality models. A descriptive personality model usually consists of a few personality dimensions that are approximately orthogonal, so as to form a personality space in which personalities can be represented as vectors, with trait levels as their components.

The measurement of the personality of an individual is usually performed by some observers using an inventory of personality scales associated with the dimensions of a descriptive personality model. So far, the most commonly adopted yet at the same time problematic descriptive personality model for measuring robot personalities is indubitably the five-factor model [8, 9], alias the 'Big Five' [10], of which there are many versions [11]. The arguably most common version consists of these five personality dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness [12]. However, little justification there is for that a model originally formulated for humans applies to artificial entities such as robots [13]. Human-computer interaction researchers have already noticed this issue and taken the initiative in creating their own models, starting with identifying personality dimensions for speech-based conversational agents [6]. With all the above said, the five-factor model should apply to androids and anidroids, i.e. robots created to resemble humans in realistic and animation styles respectively.

A critical issue at the moment is that there is not a descriptive model dedicated to assessing robot personalities. And it is hard to conceive a model fully applicable to all robots, of different designs for different purposes. What is a model that can be fully applied to both the rovers on Mars and the animatronics in theme parks? If such a model existed, it was unlikely to be as parsimonious as the five-factor model. Still, neither the lack of a dedicated alternative nor the popularity of the five-factor model justifies its applicability. Before we can acquire a dedicated model as Völkel et al. did for speech-based conversational agents, we need to use the five-factor model with caution. The most basic precaution to take is to check the applicability of the five dimensions, especially for personalities engineering tasks [13].

1.3 Research Questions and the Studies

The challenges of creating robots with individuality boil down to three problems to investigate. First, the problem about the possibility of individuality, formulated as a *CAN* question: *can physically identical robots be created with* *individuality that consists of personality traits that distinguish them from one another, under the constraints that the traits are similar?* Second, the problem about the composition of individuality, formulated as a *WHAT* question: *what personality traits compose the individuality of a type of robot?* Third, the problem about the implementation of individuality, formulated as a *HOW* question: *how to implement individuality?* This dissertation presents the three studies conducted to answer these three questions.

In the first study, I answered the first research question by verifying the similarity and differences between two robot personalities, modelled after two humans using an eye movement model, installed on the same android, and evaluated by human subjects using human personality dimensions. I measured the similarity by considering the similarity between the trait ordering (as by strength) of the two personalities as represented by their orientations as vectors in a personality space. It was understood that when an android gazes or averts their eyes following the patterns of two humans with similar personalities, the android can appear as two similar but distinguishable robots. However, the android was a special type of robot built to resemble a female human being, which was not representative of robots. Human personality dimensions might not apply to other types of robots.

In the second study, I proposed a method to select personality dimensions applicable to a type of robot to structure their individuality with traits corresponding to the selected dimensions. This method can be used to answer the second research question. It functions by testing the reports of a group of observers on how well a type of robot can approximate human personalities by imitating multimodal behaviour of multiple humans. The experiment on the method showed that it could provide practical answers to the second research question, and the personality dimensions selected by the method matched past discoveries made in similar settings. The experiment also revealed that the selection of personality dimensions could be observer-dependent.

In the third study, I proposed a generative personality model to implement individuality as a solution to the third research question. The model functions by combining a robot's actions, as associated with several personality traits, according to the ordering and strength of the traits, as associated with the robots' goals. It is thus called a goal-shaped generative personality model. The experiment on a prototype model installed on a humanoid robot showed that it was possible to make the robot appear as multiple distinguishable individuals by implementing multiple sets of model parameters. This result indicates that the model can be used for engineering physically identical robots into robots with individuality.

1.4 Main Thesis

To summarise, the first study verified the possibility of creating androids with individuality; the second study contributed a method to generalise the possibility to other types of robots; and the third study contributed a generative personality model to realise the possibility.

The results from the three studies together form the main thesis of this work: *by considering a robot's behaviour as actions associated with several personality traits, corresponding to several personality dimensions applicable to the type of robot, and then combining those actions using a goal-shaped generative personality model constructed with the applicable personality dimensions to encode the robot's individuality and generate behaviour accordingly, the robot can be made into a unique individual distinguishable from other robots of the same kind.* Following the main thesis, one way to create robots with individuality is by breaking down a robot's original behaviour into actions associated with several personality traits, and then recombining those actions according to the ordering and strength of the traits as associated with the robot's goals or, in short, by dissociating robots' behaviour from their goals and reconnecting the two with their personalities. Another approach is to consider a robot's behaviour as atomic actions associated with personality traits in the first place when designing the robot and implement the proposed model as an intrinsic part of the robot's AI.

The rest of this dissertation is organised as follows: Chapter 2 is a review of the literature on related topics; Chapters 3, 4, and 5 present the first, second, and third studies respectively; and Chapter 6 concludes this dissertation.

Chapter 2

Related Studies

2.1 Robots' Potential for Personality

'Artificial beings with human qualities' is an old idea. There is the tragic bronze automaton Talos in Greek mythology in the West and the legend of Yan Shi's automaton in the East, both being about two millennia old. Science fiction abounds with robot characters and contributed the name of the discipline, 'robotics' [14]. Academics started to get serious with the idea only later, around mid 20th century, when the advent of digital computers made it possible to study artificial personalities.

Computer personality has been a research interest since the dawn of artificial intelligence, leading to a series of studies suggesting that human personality could be at least partially simulated [15, 16, 17, 18, 19]. But it was only in the 21st century, when computer personality met embodied agents, did the real challenge begin.

An embodied robot is an existence that is best described as an individual robot. And many embodied robots would be many individuals, each requires their own individual characteristics if they are to be robot personalities, as opposed to an artificial agent that can be the 'same character that lives in everyone's devices'. What a robotic body offers are physical modes of expression, and how they express personality was among the first problems investigated. And of all the body parts, the robotic head of a humanoid robot enjoyed the first attention. In 2000, Edsinger et al. pointed out that a face was of vital importance in creating humanoid robots capable of personality expression [20]. In 2001, Miwa et al. proposed a system that expressed emotions using a robotic head with different robot personality settings [21]. Their work suggests that, by combining sensing personality, how an agent reacts to a situation, with expression personality, how the agent acts based on the appraisal of the situation, a robot can express multiple personalities. However, they only tested personalities with binary traits on three dimensions. The study by Lee et al. showed that Sony's AIBO, a dog robot, could be perceived as introverted or extroverted via its non-verbal, non-human-like behaviour [22].

Although early studies have demonstrated the potentials for both non-humanoid and humanoid robots to be engineered into personalities, their potentials are not equally bounded. For non-humanoid robots, it is hard to delimit their potential because they come in a variety of sizes and shapes for a variety of purposes, and personality can be expressed through completely irrelevant instruments when the context is right, as demonstrated by BB-8's reciprocation of a thumbs-up gesture using his welding torch in *Star Wars: The Force Awakens*. It is not an exaggeration to say that the potential is unbounded for non-humanoid robots. The potential of humanoid robots, on the other hand, is often studied around one or two of the following modes of expression: gaze or eye movements, facial expressions, gestures or body movements for non-verbal behaviour; and speech and voice for verbal behaviour.

There is an abundance of studies on non-verbal behaviour. Gaze or eye movement is one of the well-studied modes with results substantiating the potential to express personality traits with human eyes [23, 24, 25, 26, 27], eyes of virtual agents [28, 29, 30], and robotic eyes [2, 31, 32, 33]. Facial expressions, albeit not as common as eyes, can certainly express personality traits [21, 34]. Only that faces as intricate as humans' are substantial engineering challenges. Gesture and body movement are effective modes too [35, 36]. In existing literature, there are also studies that are not about modes of personality expression (as defined in this work) but other means that may be called 'modalities'. For example, the appearance of a robot may convey certain types of impressions to observers [37, 38]. For a child-like robot, even touch sensation can be a modality [39].

Verbal behaviour was little studied, perhaps due to the reality that characteristic speech, as demonstrated by movie or TV characters, is still difficult for robots with the current state of natural language processing technologies. It goes without saying that speech should be one of the most important modes, if not *the* most. Playwrights do it all the time: expressing the personalities of their characters through what they say—a task hard enough for even human intelligence, not to say AI. Still, effort has been made to express personality through synthesised speech [40], though it is still light years away from individuating robots by generating characteristic and unique speech patterns. In addition to the text contents of speech, the rate of speech may be an indicator of extraversion [35]. Voice can also reveal some personality traits [41, 40], and the potential of semantic free utterances has been explored too [42].

2.2 Effects of Robot Personalities

Generally speaking, two kinds of positive effects can be educed: improved impressions and improved compliance. The former is about eliciting positive opinions from observers, who form judgements on robots but do not necessarily need to do something for or with them. The latter is about inducing users to cooperate or collaborate. Improved impressions are studied from a variety of aspects and thus measured with a variety of measures, with notable ones being likeability (e.g. 'social attraction' of a robot dog [22]; likeability via touch sensations [39]; being sociable, amiable, and trustworthy [43]), ease of acceptance [44], and uncanniness [3]. To measure improved compliance, we need to know if the users of robots will do more of what is expected of them under the influence of robot personalities. It is shown that robots with certain personality traits can motivate users to exercise more [1, 45] and to collaborate more [2].

2.3 Considerations for Engineering Personalities

The studies until this point have invariably vouched for the prospect of engineering robot personalities for practical use. However, this accretion of knowledge on the potential of robot personalities still lacks insights into how robot personalities can be engineered. Nonetheless, there are some considerations that shed more or less some light on the path.

The first consideration is how humans perceive robot personalities. It is under the influence of a variety of factors, including the scenarios and context of interaction, such as what the robots do [46, 44]; the humans' own personalities [47, 48]; the cultural influences they receive, such as fiction about robots [49, 50, 51]; the robots' appearances, including the types of characters (virtual agents versus embodied robots) [32], the shapes of the bodies of humanoid robots [38], and the styles of appearances (e.g. mechanical, robot-like, and humanoid) [37]; the personality assessment tools applied, and so on and so forth, which have been covered by surveys [52, 53]. This means the same personalities may rarely enjoy consistent evaluation, just like characters of a book: a thousand people may form a thousand impressions of a character. A more significant revelation of the studies above is that the conveyance of personality is bilateral: it is not only about robots but people as well. This implies that understanding the many aspects of humanity is a key to creating robot personalities.

The second consideration, corresponding to a very popular topic in the field, as reviewed by Robert et al. [53], is whether robot personalities should be designed according the 'similarity-attraction' principle or the 'complementary-attraction' principle, and there is evidence supporting both the former [47, 54, 45, 55, 56, 2, 57, 58] and the latter [59, 22, 55, 60, 57]. The seeming divergence only tells us that it depends on the context, largely the tasks and roles of robots, as Joosse et al. [46] and Tay et al. [44] have substantiated with their studies. We should design robot personalities according to circumstances.

2.4 Models of Artificial Personalities

The lack of a comprehensive generative personality model for robots was a major issue to resolve in robot personalities engineering. There were only a few generative personality models in robotics research, most of which specialised for very specific applications. Most generative models of personalities were found in virtual agents or narrative generation studies. The existing generative personality models can be placed into two overlapping categories: goal-based models and models applied to robots. Goal-based models, which, depending on the context, can also be called goal-oriented models, goal-directed models, goal-conditioned models, and so on.

2.4.1 Goal-Based Models

Carbonell, in his pursuit of AI that understands stories, proposed a goal-oriented personality model for understanding characters' personalities and thereby predicting and interpreting their actions [61]. The philosophy behind his 'goal-tree' model is that a character's personality is reflected in their actions in a situation, and those actions point ultimately to their goals, thereby associating their goals with personality. His model is not generative *per se* but can be easily adapted into one, since predicting actions is in a sense generating hypothetical actions. To create virtual characters with inter-individual differences and autonomy in behaviour, Chittaro and Serra proposed a goal-oriented model, where personality traits influence how characters sense the world, plan their actions, and perform the actions [62]. The goals are represented by probabilistic finite-state machines, where the trait levels are used as weights to compute probabilities of state transitions. However, the model requires for each goal a probabilistic finite-state machine, which is easier to define in a virtual, closed world. Yet in a real, open world, the means to achieve certain goals are often unclear and hence unable to guide the behaviour or choices of actions. Bouchet and Sansonnet proposed a model for implementing influences of personality traits on cognitive agents [63]. The influences are defined by associating to personality traits sets of 'bipolar schemes', which consist of eight classes of 'level of activation' operators that can be applied to multiple phases, from goal selection to the execution of a plan for achieving the goal to the evaluation of what has been done, of the five-phase deliberation cycle of the *BDI-agent* [64]. The authors then had not yet done empirical assessments of their model. Julio et al. proposed a narrative generation system, *Mask* [65], featuring an algorithm, *CB-Glaive*, based on the work of Ware and Young [66]. Their *Mask* system can choose for a character a course of action that aligns best with their goal and hence continue the narrative from some given premises with available choices of actions.

Many goal-based models were considered in narrative generation or analysis studies. The above mentioned study by Carbonell [61] is one of the seminal contributions to narrative analysis. However, most current robots cannot deliberate goals, plans, or decisions on courses of actions for actionable, real-world behaviour. There were not many attempts on goal-based approaches to 'high-level' behaviour. Still, the flourishing field of machine learning is filling the gap. Goal-based or goal-directed reinforcement learning has striking parallels in narrative generation or other artificial agents studies. A major difference is that goals as defined in reinforcement learning usually have nothing to do with personality.

2.4.2 Generative Personality Models Applied to Robots

So far, the most common type of generative personality model applied to robots is 'specialised'. Specialised models are specialised to generate in very specific situations very specific aspects of behaviour that could evoke impressions of personality. There was the gaze model by Fukayama et al. for virtual agents [28], which I showed to be applicable to androids with necessary modifications [33]. Lee et al. implemented their system on *AIBO*, a robot dog [22]. The system had two sets of hard-coded configurations corresponding to binary traits, extroverted and introverted, in the extraversion dimension. Such an approach is applicable to companion robots or social robots that have no needs to form large colonies. For humanoid robots that work as humans' co-workers or helpers in various domains where multiple robots will work and hence be observed together, a generative personality model of a large colony capacity is required. Aly and Tapus devised a model capable of generating verbal and non-verbal behaviour for robots [56]. It worked by generating personality-modulated speech and then personality modulated gestures out of the personality-modulated speech. For their experiment, the authors tested two personality configurations on the extraversion dimension. Their model lacked a structure to encode inter- and intra-individual differences for more than two personalities on more than one dimension. It is unclear whether their model has the potential to express more personalities by expansion. The personality-driven body motion synthesis models by Durupinar et al. contributed an insight into 'effort' of personality [36]. (Their study was on artificial agents rather than physical robots, but was nonetheless relevant.) The gist of their method is to map personality traits to 'Effort', as defined in Laban Movement Analysis, and then 'Effort' to body motion parameters. Many human activities involve 'effort'—albeit defined differently in different contexts—and therefore the idea may apply to a range of human behaviour other than body motion. Zabala et al. proposed an affect-based system that generated non-verbal behaviour consistent with the robot's speech [67]. Their system did not have components to encode inter- and intra-individual differences of personalities by traits but worked by generating non-verbal behaviour from the emotions expressed in the robot's speech. Speech is an important means of personality expression, and therefore their system has the potential to be integrated into a generative personality model for speech. Better understanding of personality traits and inter- and intra-individual differences in emotional expressions via speech may be necessary to achieve the integration. Churamani et al. proposed a framework for constructing models capable of generating affect-modulated behaviour based on the appraisal of users' affective states on the arousal and valence dimensions, which is, in a sense, generating behaviour with intra-individual differences, considering that users' affective states are one of the most important contexts, among a few others, in expressing or understanding robot personalities [68]. And by configuring it to make the robot behave either patiently or impatiently, or to be either excitatory or inhibitory in forming the robot's mood, the model could generate behaviour with inter-individual differences for at least two distinct individuals on each of the three broad traits (being generous, persistent, and altruistic) that the authors confirmed to be consequential in the context of their study. The authors did not advertise their models to be generative personality models *per se*, but the potential is clear from their results. Only that personality encompasses more than just affective behaviour. How to encompass other aspects of personality and how to use their framework to construct models capable of generating behaviour with inter- and intra-individual differences for a significant quantity of robots were not the aims of the work presented by the paper and hence unknown.

It could be hard to integrate those specialised models without compromising the overall consistency of personality, an essential quality of personality [69]. In other words, it will not be easy to put specialised models together to create a robot personality with a consistent overall impression. Another common drawback is that most of those models present broad traits in very low resolution; most of them support two discrete trait levels (high, low). Being specialised models with also discrete (often binary) trait levels severely limits the colony capacity of those models. Furthermore, many of the specialised models for robots, such as the eye movement model applied to androids [33], have no structure to encode inter- and intra-individual differences of personality, and hence the links between personality structures and behaviour remain obscure. Such models have limited practical value for engineering robot personalities in significant quantities.

Chapter 3

Ordering-Based Personality and Measure

3.1 Research Question

The first research question is a *CAN* question: *can physically identical robots be created with individuality that consists of personality traits that distinguish them from one another, under the constraints that the traits are similar?* This question involves two factors: differences (in traits) and similarity (on the whole) between personalities.

Many previous studies explored the first factor but not the second; they often demonstrated differences by assuming binary classification of personality traits. For example, Lee et al. showed a robot dog could express either extraversion or introversion [22], and Bremner et al. investigated personality perception of teleoperated humanoid robots on the extraversion, conscientiousness, and agreeableness dimensions with binary trait levels, i.e. 'high or low' [31]. Important though they are as studies on robots' potential for having personality or the perception of it, in creating robot personalities however, constraints can apply, such as that robots must be high on some desirable traits, e.g. conscientiousness and agreeableness. Therefore, it is necessary to verify the possibility by verifying both similarity and difference. Verifying similarity was the focus of my first study [33]

The rest of this chapter is organised as follows: Section 3.2 elaborates on the method for measuring similarity; Section 3.3 describes the systems applied to the experiment; Section 3.4 relates the experiment; Section 3.5 shows the results of the experiment; and Section 3.6 discusses the results and concludes this chapter.

3.2 Method: Ordering of Personality Traits and Similarity

Testing similarity was trickier than differences. What counts as similarity? Just how similar can be considered similar? What is the measure? To address these issues, I proffer the following considerations.

What counts as similarity? When we try to know about a person, a common question is 'What are the most memorable characteristics of this person?' Although trait-based personality models, such as the five-factor model [9, 8], describe a person's characteristics as several principal personality traits of equal importance in an objective sense, in our daily life we often see a person from what is the most memorable of them, as predicted by the influence of central traits in forming impressions of a person [7] as well as possible 'salience bias', a type of availability bias [70]: subjectively speaking, some traits are just more prominent than the others. Consequently, there exists a ranking of the salience or *prominence* of personality traits for each observed character, the higher a trait is ranked the more prominent it is. This ranking I call the *order* of traits. If we assume that the order by the prominence of traits is determined by and hence the same with the order by the strength of traits, then this order is also determined by how a personality, represented as a vector in a personality space, is oriented. On this wise, similar personalities can be considered as several characters remembered or recognised by a similar set of broad traits; their personalities as presented as vectors in the personality space should be similarly oriented, indicating similar ordering of the strength or prominence of their personality traits.

How similar can be considered similar? Currently, there is a lack of empirical support for a clear division between 'similar' and 'not similar' when using distance-based measures. However, in trait-ordering-based measure there is a clear decision criterion of whether two orders match. We can also compute how much two orders match by computing how similarly two personality vectors are oriented.

In order to show that two personalities are similarly oriented in the personality space, we need to represent them as vectors and evaluate how similarly oriented the two vectors are. To this end, a suitable similarity measure is cosine similarity. However, in the context of measuring personalities, we use inner product with normalisation as an equivalence and adopt the braket notation because the trait ordering and orientation of a personality as a vector are more neatly associated and expressed this way.

Let $|P\rangle$ denote the personality of an individual P. Given a personality space with dimensions X, Y, and Z..., and hence let $(|X\rangle, |Y\rangle, |Z\rangle ...)$ be a basis corresponding to the personality dimensions. Then, personalities can be represented as linear combinations, for example, the personality of an individual H:

$$|H\rangle = a\,|X\rangle + b\,|Y\rangle + c\,|Z\rangle + ..., \tag{3.1}$$

where $a, b, c, ...$ are real numbers.

We assume the basis $(|X\rangle, |Y\rangle, |Z\rangle, ...)$ to be an orthonormal basis, suggesting that

$$\langle X|Y\rangle = \langle Y|Z\rangle = \langle X|Z\rangle = ... = 0 \tag{3.2}$$

and

$$\langle X|X\rangle = \langle Y|Y\rangle = \langle Z|Z\rangle = ... = 1, \tag{3.3}$$

where the inner product of two personalities $|H\rangle$ and $|R\rangle$, that is

$$\langle H|R\rangle = |H\rangle \cdot |R\rangle . \tag{3.4}$$

The practical meaning of this assumption is that the personality dimensions of a personality model correspond to broad traits that are independent of one another (though personality dimensions of most if not all popular models are not strictly orthogonal; still, the assumption is a useful approximation).

To compute the similarity between two personalities, for example

$$|H\rangle = a\,|X\rangle + b\,|Y\rangle + c\,|Z\rangle + ... \tag{3.5}$$

and

$$|R\rangle = d\,|X\rangle + e\,|Y\rangle + f\,|Z\rangle + ..., \tag{3.6}$$

we evaluate how close their inner product

$$\begin{aligned} \langle H|R\rangle &= (a\,|X\rangle + b\,|Y\rangle + c\,|Z\rangle + ...) \cdot (d\,|X\rangle + e\,|Y\rangle + f\,|Z\rangle + ...) \\ &= ad + be + cf + ... \end{aligned} \tag{3.7}$$

is to 1 (the unity). The closer the inner product is to 1, the more similar the two personalities are; and an inner product that equals to 1 suggests the complete overlap of the two corresponding vectors and a corresponding perfect match of the orders of their traits. Because any personality is identical to itself, naturally

$$\langle P|P\rangle = 1 \tag{3.8}$$

for any P. This requires a personality as a vector to be normalised before applying the measure, such that

$$\langle P|P\rangle = (x\,|X\rangle + y\,|Y\rangle + z\,|Z\rangle + ...)^2 = x^2 + y^2 + y^2 + ... = 1 \tag{3.9}$$

for any P.

Since this measure only concerns the angle between two personalities as vectors, it will consider two vectors with vastly different norms (magnitudes) but a small angle in between as similar. Therefore, it is not a generally applicable similarity measure but one used in cases where the norms of the two vectors under comparison are not vastly different. We can assume this is the case in creating robot personalities because we aim for useful and pleasant personalities, which will be more or less balanced. By 'balanced' I mean that the traits of personalities will not deviate from the desirable range so that their norms as vectors will not be drastically different. It can also be assumed that ordinary human beings normally raised and educated will have more or less balanced personalities so that their personality traits will not deviate from what is normal. This assumption means distance-based measures (as opposed to ordering-based measures), such as Euclidean distance, are unsuitable for measuring the similarity between personalities since there

FIGURE 3.1: The android from the point of view of the observer.

is a lack of benchmarks on 'how similar is similar?' within the range of normality.

3.3 Systems: Robot and the Eye Movement Model

The robot featured was a female-type android as shown in Figure 3.1. The android was controlled by an eye movement model as illustrated by Figure 3.2. It was a modified version of that of Fukayama [28]. The model worked by controlling the eyes to either gaze or avert using a Markov process. In other words, it controlled the switching between two states: Gazing and Averting. 'Gazing' was when the robot under control was looking at the other interlocutor, and 'Averting', somewhere else. The switching was controlled by two variables: the timing D (duration) and the probability R of switching. In this study, the android's personality was based on two human archetypes denoted by T and K. D was the average time the human archetype spends on Gazing, measured in milliseconds (ms). R ($0 < R < 1$) was the proportion of total amount of time of interaction the human archetype spent gazing used as a probability. The switching happened following such rules: upon entering the 'Gazing' state, initiate the next transition in D time, and when the time arrived, the model might stay on Gazing with R probability or avert the eyes with $1 - R$ probability; upon entering the 'Averting' state, initiate the next transition in $D(1 - R)/R$ time, and when the time arrived, the model might stay on Averting with $1 - R$ probability or let the robot gaze back at the interlocutor with R probability.

Another thing for the model to handle was where to look at during aversion. The robot's field of vision was framed as a Euclidean plane, with the face of the other interlocutor at its centre. Upon entering the 'Averting' state,

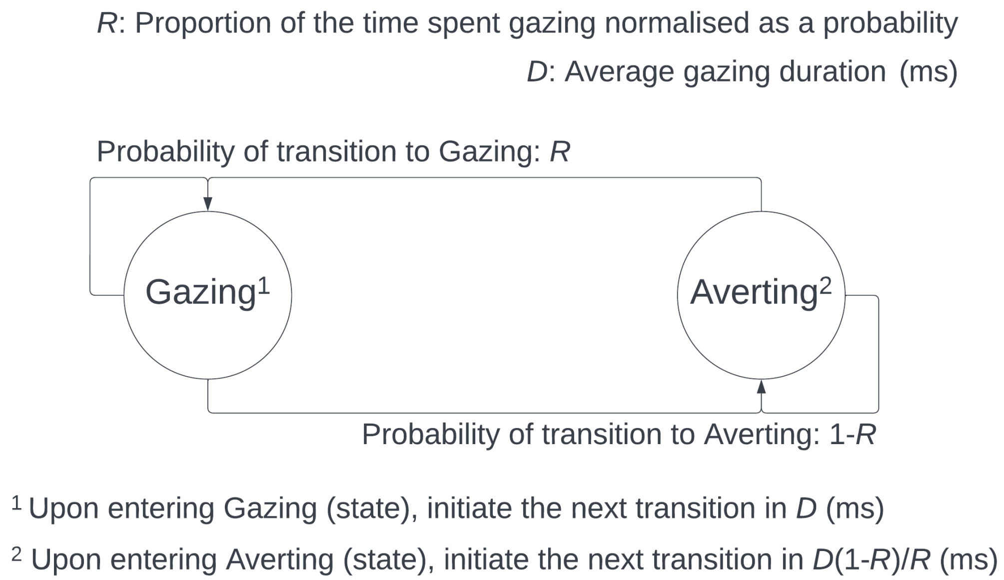

FIGURE 3.2: The eye movement model based on a two-state Markov process.

the place to look at was a coordinate sampled from a bivariate normal distribution:

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\Big(-\frac{z}{2(1-\rho^2)}\Big), \tag{3.10}$$

where

$$z \equiv \frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}. \tag{3.11}$$

The means and variances were learnt from the distribution of the archetype's gaze aversion. Whether the archetype was gazing or averting their eyes was determined using 2-means clustering. When the archetype was engaged in the designated dyadic conversation, their eye movement was recorded using an eye-tracking device on a constant time interval (20 ms). The locations of their gaze were recorded as coordinates on the Euclidean plane. 2-means clustering clustered the coordinates into two clusters corresponding to the two states. The 'Averting' cluster provided the distribution to learn the parameters from.

3.4 Experiment: Evaluating the Personalities

3.4.1 Hypothesis

To answer the research question, I tested the following hypothesis [33]:

> *An individual perceives the personality traits of a human archetype with consistency in an android which imitates the archetype's eye movements, such that when an audience, a group of people, perceives the personality traits in such an android, the personality traits the audience—despite the individual differences of its members in perceiving personality traits from eye movements—perceived, using proper methods, shall agree with those perceived in the same manner in the archetype.*
>
> *What follows the hypothesis is that, when an audience perceives the personalities of two arbitrarily selected archetypes which happen to be different, such differences shall manifest in the android that imitates their eye movements, even if it is the same android that imitates the two archetypes (in turn) with all other conditions being kept equal.*

The hypothesis indicates that an android that imitates a human's eye movement will be considered as similar to the human in personality, and if the android imitates the eye movements of two human archetypes with distinctly different personality traits, the android expresses two distinctly different personalities. If the two human archetypes have similar personality traits, the personality traits as expressed by the android will also be similar. Therefore, to test the hypothesis, we need to test the two factors: similarity and difference.

3.4.2 Scope

The scope, i.e. expressing personalities via the eye movements of an android in dyadic communication, as specified by the hypothesis is potentially generalisable. If the possibility is confirmed for the android using only their eyes, then the possibility may extend to any androids. And androids are humanoid robots, many of which have eyes too. The possibility with androids could be an indication of the potential of other humanoid robots (though this was yet to be confirmed at the time). Dyadic communication is also common among humans and hence pertinent to human-robot interaction.

3.4.3 Procedure

Two male members of our laboratory acted as the archetypes. Their eye movements were recorded using Tobii Pro Glasses 2 when they recited a Japanese script about 'the problem of sub-replacement fertility in Japan'. The topic and script were prepared to be as neutral as possible to minimise interference (e.g. from the archetypes' emotions and mental activities as reactions to the content of the script, which might also affect their eye movements). To further reduce interference, the eye movement model worked independently of the content of the script. In this way, the sources of differences in the robot personalities should be the differences between the personalities of their corresponding archetypes rather than the differences in how the archetypes reacted to the content due to reasons other than their individuality. The two instances of interaction were recreated using the android. All

FIGURE 3.3: The android's eye movements.

four instances were recorded as short videos about 90-sec to 120-sec long. The videos presented the point of view of an interlocutor facing the android (Figure 3.1)

39 subjects, 19 females and 20 males, served as observers. All of them were Japanese students at Osaka University, and none of them was a member of our laboratory. Each of them watched the videos and evaluated the personalities in the videos using Hayashi's personality assessment inventory, which was in Japanese and consisted of 20 7-point Likert scales [71]. The videos appeared in a random order for each observer for counterbalancing.

3.5 Results: Differences and Similarity

TABLE 3.1: The three broad traits as evaluated per personality

	Personality Dimensions		
Personalities	Sociability	Familiarity	Activeness
Human T	4.8	4.8	4.4
Android U_T	4.3	4.3	3.9
Human K	3.9	3.4	3.6
Android U_K	3.4	3.2	2.4

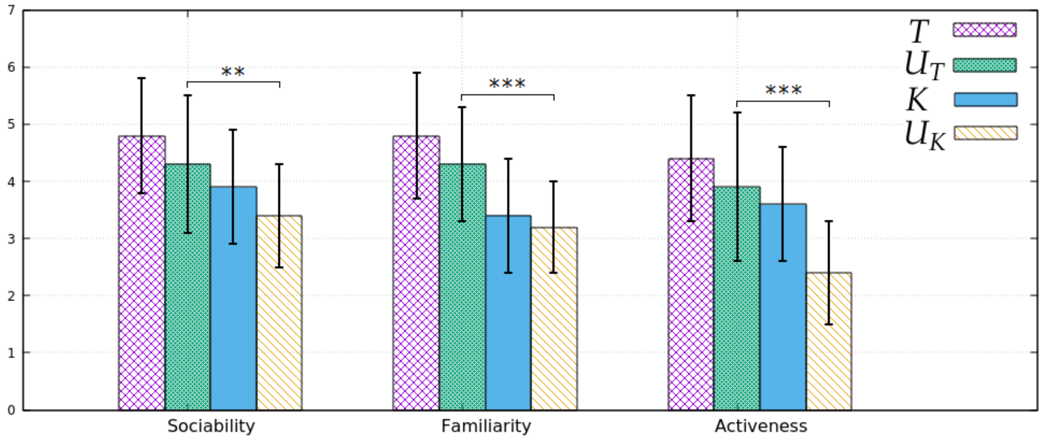

FIGURE 3.4: The four personalities measured on the three dimensions.

TABLE 3.2: The Wilcoxon signed-rank test results

	T versus K			U_T versus U_K		
Dimensions	p-value	Z-value	α[1]	p-value	Z-value	α[1]
Sociability	0.00022	-3.7	**	0.00046	-3.5	**
Familiarity	0.00006	-4.0	***	$\leqslant$0.00001	-4.6	***
Activeness	0.01778	-2.4	*	0.00006	-4.0	***

[1] * $p < .05$; ** $p < .001$; *** $p < .0001$

The reported personality trait levels on the three dimensions of Hayashi's model are plotted in Figure 3.4 and recorded in Table 3.1.

3.5.1 Testing Differences

I tested the differences using the Wilcoxon signed-ranked test, since the data were non-parametric. The results were recorded in Table 3.2 and illustrated by Figure 3.4. The robot personalities were distinctly different across all three dimensions, so were the two human archetypes they were based on.

3.5.2 Testing Similarity

To apply the ordering-based similarity measure, let $|T\rangle, |U_T\rangle, |K\rangle, |U_K\rangle$ denote the personality vectors of T, U_T,K, and U_K; let $(|S\rangle, |F\rangle, |A\rangle)$ be an orthonormal basis corresponding to the personality dimensions 'Sociability', 'Familiarity', and 'Activeness'.

Using the data from Table 3.1, compute the following:

$$|T\rangle = 0.593\,|S\rangle + 0.593\,|F\rangle + 0.544\,|A\rangle \tag{3.12}$$

$$|U_T\rangle = 0.595\,|S\rangle + 0.595\,|F\rangle + 0.540\,|A\rangle \tag{3.13}$$

$$|K\rangle = 0.619\,|S\rangle + 0.540\,|F\rangle + 0.571\,|A\rangle \tag{3.14}$$

$$|U_K\rangle = 0.648\,|S\rangle + 0.610\,|F\rangle + 0.457\,|A\rangle\,. \tag{3.15}$$

Then, compute the similarity between the two human archetypes,

$$\langle T|K\rangle = 1.00; \tag{3.16}$$

the two robot personalities

$$\langle U_T|U_K\rangle = 1.00; \tag{3.17}$$

the archetypes and the robot personalities modelled after them respectively,

$$\langle T|U_T\rangle = 1.00 \tag{3.18}$$

and

$$\langle K|U_K\rangle = 0.99; \tag{3.19}$$

and even cross-comparing,

$$\langle T|U_K\rangle = 0.99 \tag{3.20}$$

and

$$\langle K|U_T\rangle = 1.00. \tag{3.21}$$

It could be concluded that all characters featured in the experiment were highly similar to one another.

Since the similarity measure concerned only the orientations but not the norms of personality vectors, it had been necessary to check the norms of the four vectors before normalisation, which could be computed using the data in Table 3.1:

$$|T| = 8.1 \tag{3.22}$$

$$|U_T| = 7.2 \tag{3.23}$$

$$|K| = 6.3 \tag{3.24}$$

$$|U_K| = 5.2. \tag{3.25}$$

The goal was to check whether it was possible to create robot personalities in significant quantities, and therefore the difference between robot personalities U_T and U_K mattered. The difference between their norms was about 2. Whether 2 was large or small should be considered in the context of the goal. If we used 2 as the distance to distinguish similar personalities, then in any direction in the personalty space, where the maximum value of each dimension was 7, we could have a minimum of four distinct individuals as three similar pairs. Each of the pairs could be similar to another pair in another direction that was similar to the current one by the similarity measure. This means that, with 2 as the standard for difference in norms, we could have at least three sets of four similar personalities in two similar directions. Even if only one of the three sets met the design constraints (e.g. all personalities must be high in sociability), two directions allowed for at least four similarity personalities; and there could be more than two directions, admitting more possibilities. Therefore, a difference of 2 did not invalidate the result of the similarity measure.

3.6 Discussions

The above-presented experiment results showed that physically identical androids can express distinctly different yet similar personality traits via eye movement. This confirmed the possibility of creating robots with individuality, i.e. personality traits that distinguish an individual from others of the same kind, using androids. Moreover, the results substantiated the significance of ordering traits by strength in understanding and distinguishing individuals.

There were two critical limitations to this study. Using an android led to uncertainty regarding other types of robots. Androids are robots built to resemble humans, and therefore personality dimensions that apply to humans will apply to androids too if they behave sufficiently human. The same cannot be said for other types of robots. The crux of the issue is the applicability of human personality dimensions to robots that are not androids: what personality dimensions describe a type of robot that is not made in the human image? The other limitation was that the 'proof-of-concept' experiment was but one successful instance involving only two archetypes corresponding to two robot personalities. Although verifying the possibility required only one successful instance, the aim of this research was to create robot personalities in significant quantities.

To address both limitations, a method that could select applicable personality dimensions for a type of robot for creating multiple robot personalities was required and therefore proposed in Study 2.

Chapter 4

Personality Structuring Method

4.1 Research Question

The second research question is a *WHAT* question: *what personality traits compose the individuality of a type of robot?* Here, 'the individuality of a type of robot' does not refer to the characteristics of this type of robot that distinguish them from other types of robots, but the individuality of every individual robot based on this type of robot. To create individual robots out of a type of robot, we need to know the personality dimensions applicable to this type of robot, which correspond to some broad traits that can describe this type of robot for some users. Then the question becomes: *what are those personality dimensions?*

Since the answer to the question depends on the type of robot, among other factors, in this study, I have proposed a method to answer the question [13]. The answers should come from the users of the robots, since they are the main observers, judges, and beneficiaries of the robots' personalities. Most users are humans. Humans exhibit a tendency to attribute human qualities to non-human entities. For this reason, human personality dimensions, especially those of trait-based models, are frequently used as the bases for synthesising artificial personalities [72].

Trait-based personality models are often formulated as a set of personality dimensions on which an individual's traits can be measured. A well-known example is the five-factor model, first discovered by Tupes and Christal in 1961 (a reprint of their work is available [9]), and subsequently by Norman in 1963 [8]. The arguably most popular variant has these five dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness [12]—known as the 'Big Five' [10].

However, there are several issues in applying human personality models to creating robot personalities. The first issue is the most critical: the applicability of personality dimensions. Human personality models are often empirical formulations based on a lexicon formed over time for describing humans, as explained by the lexical hypothesis [73, 10]. Lexicons for describing humans do not necessarily apply to describing artificial agents [6]. It violates the lexical hypothesis to apply a human personality model to robots that are not androids. The implication of this issue is that it is possible that not all personality dimensions of a model apply to a type of robot, especially when its appearance is far from a human image. Another issue is that, albeit of objectively equal importance, not all personality traits are equally prominent in people's eyes. A character is often recognised by their most memorable traits, as predicted by the 'availability bias' [70], and there can be 'central traits' that dominate the overall impressions of an individual [7]. How an individual is perceived is often affected by cognitive biases, which are something to consider or even leverage in robot personalities engineering. Last but not least, whether optimisation is possible on a personality dimension depends on how some specific users perceive the personalities of a type of robot, especially whether they can provide effective feedback to guide the optimisation of the traits in the personality dimension. The dimensions may vary with users since it is possible that not all people consider a broad trait to be relevant to a type of robot that is not an android. To summarise, the first issue implies that we cannot be so sure if a human personality dimension applies to a robot that is not an android. The second issue implies that we need to focus on traits that matter most to users. The third issue implies that the specific users of certain robots have the ultimate say on what traits apply and matter to their robots. Such traits are which they can provide effective feedback on. Therefore, we need a selection procedure to select personality dimensions on which some users can provide effective feedback about the personality traits of their robots, as an engineering tool for creating robot personalities out of a type of robot knowing its typical usage.

The main contribution of this work is such a selection procedure. It applies to robots that can imitate human behaviour. It is an engineering tool to select personality dimensions as the structural basis for constructing a generative personality model and also the dimensions on which a ***user group***, that is a group of specific users of a specific robot, who will be the main observers of the robot, can provide effective reports to guide the optimisation of the robot's personality construct. As far as existing surveys can tell [72, 52, 53], this is the first work to propose such a selection procedure dedicated to creating robot personalities in significant quantities for practical applications.

To test the proposed selection procedure, I conducted a series of tests simulating engineering tasks where 10 robot personalities were to be engineered out of 10 personality archetypes for small user groups of 3 to 18 people using the dimensions of the five-factor model [12]. The type of robot, a life-size humanoid 'barebones' robot, engaged users in dyadic communication, and its main modes of personality expression were head and eye movements. The effectiveness of the proposed selection procedure was confirmed within the scope of the tests. The results show that the proposed method worked for user groups with at least eight people.

The rest of this chapter is organised as follows: Section 4.2 elaborates on the research goal, why we need the selection procedure, and examines the insufficiency of previous work; Section 4.3 presents the selection procedure; Section 4.4 relates the experiment conducted to test the selection procedure; Section 4.5 shows the results of the experiment with the analysis; and Section 4.6 discusses the results and limitations and concludes this chapter.

4.2 Research Goals and the Limitations of the State of the Art in the Engineering Context

The selection procedure should tell us, with the help of a user group, which of the personality dimensions of a descriptive personality model are suitable for structuring robot personalities for a type of robot knowing its typical usage. It should meet three requirements:

- It selects personality dimensions on which a user group can provide effective feedback to guide the optimisation of the personality traits of the robot personalities under construction;
- It supports significant quantities of robot personalities;
- It works with small user groups.

Many previous studies have more or less done similar work in exploring robots' potential for expressing personality traits or studying the effects or properties of robot personalities [74, 22, 75, 39, 43, 1, 2, 37, 3]. They offer valuable scientific insights into the roles of robots' personalities in human-robot interaction. However, their 'tests' were unsuitable for engineering tasks for they fell short of at least one of the above requirements.

The first and most common limitation is to consider the types of personalities rather than personalities with individual differences, which we call the 'binary trait' simplification. Vinciarelli and Mohammadi have referred to splitting personalities into two classes (per dimension) as *'binary classification approaches'* and in their extensive survey commented that binary classes were *'not meaningful from a psychological point of view'* [72]. Their survey has revealed that 'binary classification approaches' were prevalent in the field. It is also arguable that the approach is not meaningful from an engineering point of view either. In fact, it defeats the purpose entirely since classification underplays individuality by definition. To create many robots with individuality, the 'binary trait' simplification must not apply. The proposed selection procedure does not require the 'binary trait' simplification and hence is not subject to the said limitation.

If considered in the personalities engineering context, the second limitation of previous 'tests' is the requirement of a large sample of observers. Most if not all previous studies involved more than 20 observers, which is appropriate for studying effects or properties of robot personalities, where large representative samples are desirable. However, in creating robot personalities for practical applications, most user groups will be small. Potential household user groups will mostly consist of two to nine people, extrapolating from the UN's data in 2017 on household sizes [76]. As for small businesses, as of the time of this work, in the United States, currently the largest economy, the average number of employees is about 10, and for small businesses that have employees, the numbers of employees range from 1 to 19 [77]. As of Japan in 2019, small enterprises with fewer than 20 employees accounted for 85 per cent of all enterprises, and those in service, retail, and wholesale industries had up to 5 employees [78]. Whether we consider potential household users or enterprise users, small user groups with fewer than 20 people will be the most common. The test results from one (large) sample of subjects do not necessarily apply to other (smaller) user groups due to individual differences in personality perception, which subject to a number of factors [35, 46, 44, 50, 51, 49, 47, 48]. The results can be highly user-dependent and contextual, meaning different user groups may perceive the same robot personalities differently, and the same user group may perceive the same robot personalities differently under different circumstances. The same behaviour may indicate different traits in different minds in different contexts. Something as straightforward as eye contact can be pointing to different traits in different cultures. In many eastern cultures, staring into the other's eyes is being confrontational or arrogant; in many western cultures, not doing so is being disrespectful and showing disinterest or guilt. It is not only cultural differences. Interpersonal differences should also be taken into account. Between lovers, staring into the other's eyes can be a cue of strong affection, in eastern or western cultures alike; however, between rivals, it is expressing animosity, strength, resolution, among other possibilities. It is almost certain that, due to the complexity of personality and how it is perceived, a group of users is unlikely to perceive the robot personalities under construction as similar to the archetypes without optimisation, even when the robots exhibit the most 'archetypal' behaviour. This 'self-other' discrepancy has been observed in perceiving human personalities for a long time [79], and in the perception of robot personalities as well [31]. The discrepancy should be minimised for the sake of consistent user experience. What dimensions and how much discrepancy to minimise depend on the users. It follows that we should optimise personality constructs on a case-by-case basis.

The third limitation and most fatal one is missing the step to check the applicability of human personality dimensions to robots that are not androids. We can assume human personality dimensions are applicable to androids since they are created to resemble humans and hence do. For other robots, we cannot be so sure if the items of a dimension describe a robot (e.g., whether an 'open-minded' robot vacuum cleaner makes sense to certain users). Consequently, this limitation would apply to nearly all previous testing methods if they were to be applied to personalities engineering unless the robots were androids, since the results acquired would be in violation of the lexical hypothesis [73, 10]. The proposed method is itself a guard against descriptions that do not make sense regarding the type of robot to engineer.

4.3 Method: Proposed Selection Procedure

The selection procedure tests hypotheses in the following format:

Hypothesis (format). *Given a type of robot R exhibiting typical behaviour B in situation S and a personality dimension P of a personality model M, P is a personality dimension on which fidelity can be effectively optimised.*

The exact mathematical definition of fidelity should depend on the personality model to investigate. Generally speaking, the fidelity of a trait of a robot personality under construction can be the accretion of the discrepancy measurements between the corresponding trait of the archetype and the observations on the same trait of the robot personality (for details, please refer to Section 1.2.4). Since a robot personality has to be observed and the observers' own personalities and backgrounds affect their observations, there can be no absolutely objective observation of a robot personality and a measurement of fidelity is always associated with a group of observers. Unlike humans, robots cannot report their own personalities as their 'true' personalities—unless they have self-awareness; what serve as their supposed 'true' personalities are their corresponding archetypes, which do not necessarily match their observed personalities without optimisation.

Testing a hypothesis in the format requires testing three corresponding sub-hypotheses:

Sub-Hypothesis 1. *The fidelity, computed as the proximity between the human observations and the corresponding archetypes, is statistically distinguishable from that by random guesses.*

Sub-Hypothesis 2. *There is a significant difference between the consistency of the observations on the robot personality and that of those on a human personality in the same settings.*

Sub-Hypothesis 3. *The pseudo-fidelity, computed as the proximity between the human observations and the corresponding observers' own personalities, is statistically distinguishable from that by random guesses.*

To select a dimension as optimisable, we should reject only the null hypothesis of Sub-Hypothesis 1 (Combination 5 in Table 4.1). The rationale behind is explained by the following three working hypotheses:

TABLE 4.1: The decision table.

C.	Sub-Hypothesis 1	Sub-Hypothesis 2	Sub-Hypothesis 3	Selection
1	NO	NO	NO	NO
2	NO	NO	YES	NO
3	NO	YES	NO	NO
4	NO	YES	YES	NO
5	YES	NO	NO	YES
6	YES	NO	YES	NO
7	YES	YES	NO	NO
8	YES	YES	YES	NO

YES: Yes, the null hypothesis is rejected. NO: No, the null hypothesis is not rejected.

Working Hypothesis 1. *When some human observers are using a personality model designed based on a lexicon for describing humans or animals to assess a robot personality engineered after an archetype, the corresponding fidelity is not necessarily statistically distinguishable from that by random guesses, which implies that they have completed the assessment by guesswork.*

If the observers are just guessing the traits, their observations cannot be used to guide the optimisation of fidelity. However:

Working Hypothesis 2. *If a robot is capable of imitating human behaviour in a given context where such behaviour is expected and the behaviour is typical of the robot in a usage that matches the context, some traits that apply to humans will also apply to the robot and are prominently observable in their typical behaviour, thereby resulting in fidelity that is statistically distinguishable from that by random guesses.*

Even if the fidelity is non-random, we need to eliminate two other possible causes of non-randomness to make sure that it can guide optimisation: inconsistency and observers' own personalities. Reports scattered on a personality dimension can still lead to significant differences from random guesses if they are dispersed enough. Highly dispersed reports reflect great inconsistency of opinions on the robot personalities. If the observers report the robot personalities as similar to their own, which is not impossible [47], the corresponding fidelity will also be non-random while being irrelevant to the archetypes.

Working Hypothesis 3. *If the fidelity from human observations that are as consistent as on a human personality in the same settings is statistically distinguishable from that by random guesses and the cause of it is not that the observers have reported the robot personalities to be similar to their own, the human observations can be used to guide the optimisation of the fidelity.*

The third working hypothesis is supported by the following reasoning: when the robot's behaviour is mapped to personality measurements as completed by some human observers and the mapping is not random but consistent, there exists an instance of behaviour leading to measurements that are the closest to the corresponding archetype. By approximating that instance of behaviour, we can approximate the optimal fidelity. Assuming the robot's behaviour is controlled by some parameters of a generative personality model, there should exist a set of parameters leading to the 'optimal' behaviour. In that regard, common optimisation methods should apply to approximating the parameters, such as gradient descent and genetic algorithms. However, whether they are efficient is another story.

The proposed selection procedure is based on the 'robots-imitating-humans' approach [80, 81, 82] and existing statistical tools. The resources required are:

1. Fungible units of the type of robot to test;
2. Human archetypes who can serve as desirable examples for the type of robot in performing the tasks it is designed for;
3. Tools to capture the example behaviour as data;
4. Methods to enable the robot to imitate the example behaviour;
5. Actors that act as users;

6. The user group the robots are going to work for.

The procedure has four phases, as illustrated in Figure 4.1 (where the arrows mark the dependencies; the capsule is the starting point; the rectangles are processes; the cylinders are data sets or materials; and the hexagons are the results of the processes). It is worth noting that, in case of optimising personality constructs, it is unnecessary to perform the entire procedure from the beginning when more orders for the same model of robots for the same usage come from some other user groups; in this case, we can start with user assessment. In the following subsections, we will go through the four phases one by one in detail.

4.3.1 Phase 1: Recording Archetypal Behaviour

In the first phase, we first recruit some candidate archetypes, and then, in 'behaviour recording sessions', we acquire their personality measurements, record their behaviour, and let them report the personalities of the actors acting as users. How behaviour recording sessions should be carried out depends. Generally speaking, it is a simulation of the typical usage of the type of robot to develop, where the candidate archetypes will be examples for the robot personalities. For instance, if the type of robot is going to be office errand runners, we hire model (well-received) office workers to simulate an office environment; if it is going to be waiters in a restaurant, we hire model waiters and waitresses and simulate a restaurant; if it is going to be singers on the stage, we simulate a stage with real singers.

4.3.2 Phase 2: Implementing the Behaviour

In this phase, we produce the stimuli for user assessment in Phase 3. The stimuli can be video recordings of the simulation in Phase 1 or they can be the robot personalities themselves. For the latter, more than one unit of the type of robot may be required. How to produce the stimuli depends. Generally speaking, the robots' behaviour should be as close to the archetypes' as possible. First, in 'screening', we need to exclude the candidates whose behaviour is beyond the robot's capabilities or operational parameters. We use only the behavioural data from the selected candidates, who will be the human archetypes. Then, we process the data. How to do this depends. In general, it is turning videos or motion capture data into a form that the robot can imitate. For example, in developing waiter robots, this can be processing a waiter's gestures and body motion from when ushering guests into their seats into joint tracking data. Next, in 'extracting behaviour', we further separate the processed data according to the modes we are interested to investigate. For example, for developing waiter robots, we might be more interested in facial expressions and hand gestures than gait as modes of personality expression. Finally, we program the robot with the behaviour to recreate the simulation in Phase 1 as how we produce stimuli for user assessment in Phase 3.

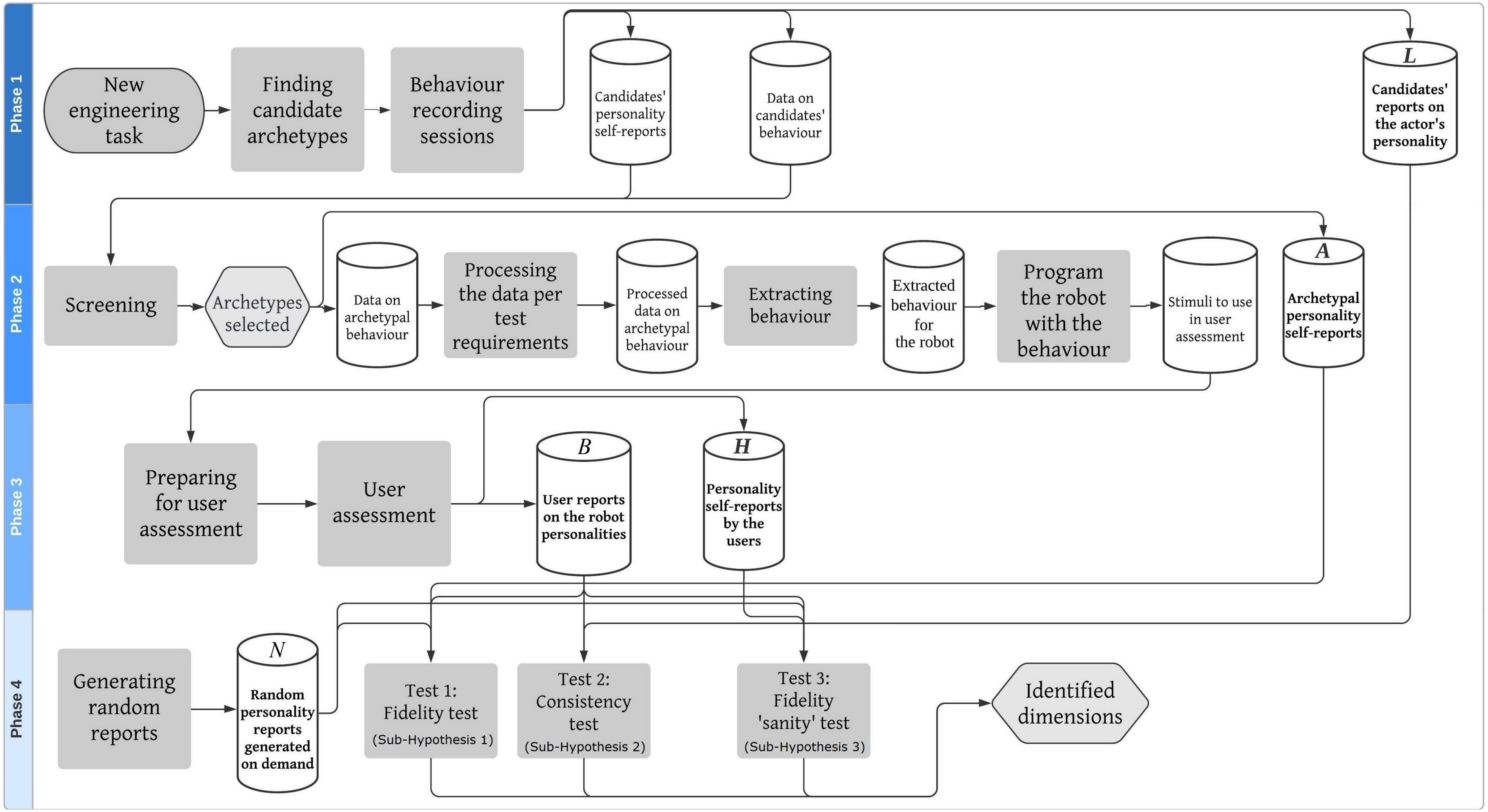

FIGURE 4.1: Workflow of the procedure (where the four rows represent the four phases; the arrows mark the dependencies; the capsule is the starting point; the rectangles are processes; the cylinders are data sets or materials; and the hexagons are the results of the processes). In addition, Phase 1 is about recording archetypal behaviour; Phase 2, implementing the behaviour; Phase 3, user assessment of robot personalities; and Phase 4, the three tests.

4.3.3 Phase 3: User Assessment

In Phase 3, we first make preparations for user assessment and then request the users to assess the robot personalities. The users can do this either by watching the videos of the robots or by interacting with the robot personalities themselves. For assessment based on video stimuli, we need to prepare only surveys. For assessment based on live interaction, we need to prepare for interactive settings as close to the simulations in Phase 1 as possible. After the surveys or interactive sessions, the users also need to report their own personalities.

4.3.4 Phase 4: Three Tests

In Phase 4, we select personality dimensions using three tests. We henceforth refer to users also as observers, since the users of the robots are the main observers of the corresponding robot personalities.

Data Sets Required

The tests require five data sets in total (Table 4.2), four of which are from the previous phases: the archetypal personality self-reports, hereinafter denoted as $\boldsymbol{A}$, from Phase 2; the candidates' (or archetypes') reports on the actor's personality (if the simulation in Phase 1 has involved multiple actors and reports have been acquired on all the actors, the reports on one of them are enough), hereinafter denoted as $\boldsymbol{L}$, from Phase 1; the users' reports on the robot personalities, hereinafter denoted as B, from Phase 3; and the personality self-reports by the users, hereinafter denoted as $\boldsymbol{H}$, from Phase 3. The fifth data set consists of random reports generated on demand, hereinafter denoted as N. Here, B and N are not in bold because they are sets. In addition, r denotes the number of robot personalities to engineer, which is the same as the number of archetypes; u the number of users; t the number of personality dimensions to test; and c the number of reports in the data set $\boldsymbol{L}$.

TABLE 4.2: Data sets required.

Data Sets	Dimensions	Row Vectors	Column Vectors
$\boldsymbol{A}$	$r \times t$	$\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_r$	$\boldsymbol{a}'_1, \boldsymbol{a}'_2, \boldsymbol{a}'_3, \ldots, \boldsymbol{a}'_t$
$\boldsymbol{L}$	$c \times t$	$\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3, \ldots, \boldsymbol{l}_c$	$\boldsymbol{l}'_1, \boldsymbol{l}'_2, \boldsymbol{l}'_3, \ldots, \boldsymbol{l}'_t$
$B = \{\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \ldots, \boldsymbol{B}_u\}$	$u \times r \times t$	-	-
$\boldsymbol{H}$	$u \times t$	$\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \ldots, \boldsymbol{h}_u$	$\boldsymbol{h}'_1, \boldsymbol{h}'_2, \boldsymbol{h}'_3, \ldots, \boldsymbol{h}'_t$
N	depends	depends	depends

$\boldsymbol{A}$ is an $r \times t$ matrix. It consists of r t-dimensional row vectors corresponding to r archetypes or t r-dimensional column vectors corresponding to t personality dimensions. $\boldsymbol{L}$ is a $c \times t$ matrix. It consists of c t-dimensional row vectors corresponding to c sets of human observations on a human or t c-dimensional column vectors corresponding to t personality dimensions. B is a set that consists of u $r \times t$ matrices: $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \ldots, \boldsymbol{B}_u$, since each of the observers has reported t traits on r robot personalities; each matrix has the same dimensions as $\boldsymbol{A}$ and can be expressed likewise. $\boldsymbol{H}$ is a $u \times t$ matrix. It consists of u t-dimensional row vectors corresponding to u observers (users) or t u-dimensional column vectors corresponding to t personality dimensions. N consists of randomly generated data per the dimensions required. To summarise, we have

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_3 & \ldots & \boldsymbol{a}_r \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \boldsymbol{a}'_1 & \boldsymbol{a}'_2 & \boldsymbol{a}'_3 & \ldots & \boldsymbol{a}'_t \end{bmatrix}, \tag{4.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{r \times t}$, $\boldsymbol{a}_{i=1,2,3,\ldots,r} \in \mathbb{R}^{1 \times t}$, and $\boldsymbol{a}'_{j=1,2,3,\ldots,t} \in \mathbb{R}^{r \times 1}$;

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{l}_1 & \boldsymbol{l}_2 & \boldsymbol{l}_3 & \ldots & \boldsymbol{l}_c \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \boldsymbol{l}'_1 & \boldsymbol{l}'_2 & \boldsymbol{l}'_3 & \ldots & \boldsymbol{l}'_t \end{bmatrix}, \tag{4.2}$$

where $\boldsymbol{L} \in \mathbb{R}^{c \times t}$, $\boldsymbol{l}_{i=1,2,3,\ldots,c} \in \mathbb{R}^{1 \times t}$, and $\boldsymbol{l}'_{j=1,2,3,\ldots,t} \in \mathbb{R}^{c \times 1}$;

$$B = \{\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \ldots, \boldsymbol{B}_u\}, \tag{4.3}$$

where $\boldsymbol{B}_{i=1,2,3,\ldots,u} \in \mathbb{R}^{r \times t}$;

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \boldsymbol{h}_3 & \ldots & \boldsymbol{h}_u \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \boldsymbol{h}'_1 & \boldsymbol{h}'_2 & \boldsymbol{h}'_3 & \ldots & \boldsymbol{h}'_t \end{bmatrix}, \tag{4.4}$$

where $\boldsymbol{H} \in \mathbb{R}^{u \times t}$ and $\boldsymbol{h}_{i=1,2,3,\ldots,u} \in \mathbb{R}^{1 \times t}$, and $\boldsymbol{h}'_{j=1,2,3,\ldots,t} \in \mathbb{R}^{u \times 1}$.

N represents a 'blind guesser' who cannot perceive any personality traits and thus has no recourse but to guesswork when it is required to complete a personality assessment. It is what the observers are pitted against. Personality traits are often assessed with statements describing certain qualities of a subject, such as: '... is someone who likes to talk with friends'. The users must indicate the extent to which they agree or disagree with the statements. They are *guessing* if they have no idea how well the statements describe the robot personalities, or they can simply indicate that they neither agree nor disagree with the statements. A significantly large number of wild guesses should exhibit the same behaviour as random guesses generated by the uniform distribution.

Test 1: Robot Personalities' Fidelity Test

The first test to run is the fidelity test. It tests Sub-Hypothesis 1 per dimension.

It requires data sets $\boldsymbol{A}$, B, and N. Let s denote the total number of reports,

$$s = u \cdot r. \tag{4.5}$$

The fidelity test compares the fidelity by the human observations with that from N to select the personality dimensions.

We can compute fidelity as follows: given a trait as observed by u users, which can be represented by a vector $\boldsymbol{b}$ ($\boldsymbol{b} \in \mathbb{R}^u$), and the corresponding trait of the archetype a ($a \in \mathbb{R}$), fidelity vector $\boldsymbol{f}$ ($\boldsymbol{f} \in \mathbb{R}^u$) is expressed as

$$\boldsymbol{f} = abs(\boldsymbol{b} - a), \tag{4.6}$$

where *abs* denotes a function that replaces all elements in a matrix or vector with their absolute values. (To avoid confusion, we refrain from using $||$ since it also denotes the determinant of a matrix.)

Each element in the fidelity vector $\boldsymbol{f}$ is a numerical distance, which can also be called 'a proximity value'. For instances, if Observer 5 has reported the extraversion level of a robot to be 3.5 when the robot personality is based on Archetype 3, whose extraversion level is 4, then the 5th element in the $\boldsymbol{f}_3$ of extraversion is 0.5. The fidelity of all robot personalities on all dimensions can be represented as an $s \times t$ matrix: $\boldsymbol{F}$ ($\boldsymbol{F} \in \mathbb{R}^{s \times t}$). $\boldsymbol{F}$ can be computed from $\boldsymbol{A}$ and B using

$$\boldsymbol{F} = \begin{bmatrix} abs(\boldsymbol{B}_1 - \boldsymbol{A}) & abs(\boldsymbol{B}_2 - \boldsymbol{A}) & abs(\boldsymbol{B}_3 - \boldsymbol{A}) & \ldots & abs(\boldsymbol{B}_u - \boldsymbol{A}) \end{bmatrix}^\intercal . \tag{4.7}$$

We generate the same number of random reports to the assessment scales. A random report consists of random numbers generated as random answers to the questionnaires about the robot personalities. For example, given a questionnaire that consists of 44 questions with 5-point Likert scales enquiring how much an observer agrees with the corresponding statements, a human report would consist of 44 responses, whereas a random report would consist of 44 random integers in the range of $[1, 5]$. Random traits should be computed in the same way per the instructions of the personality assessment inventory, and the results are divided into u $r \times t$ matrices, which are denoted here as $\boldsymbol{N}_1, \boldsymbol{N}_2, \boldsymbol{N}_3, \ldots, \boldsymbol{N}_u$ ($\boldsymbol{N}_{i=1,2,3,\ldots,u} \in \mathbb{R}^{r \times t}$) . We then compute the random 'fidelity' $\boldsymbol{F}^{'}$ ($\boldsymbol{F}^{'} \in \mathbb{R}^{s \times t}$) using

$$\boldsymbol{F}^{'} = \begin{bmatrix} abs(\boldsymbol{N}_1 - \boldsymbol{A}) & abs(\boldsymbol{N}_2 - \boldsymbol{A}) & abs(\boldsymbol{N}_3 - \boldsymbol{A}) & \ldots & abs(\boldsymbol{N}_u - \boldsymbol{A}) \end{bmatrix}^\intercal . \tag{4.8}$$

With $\boldsymbol{F}$ and $\boldsymbol{F}^{'}$ ready, we apply an appropriate statistical test. Which test should be applied depends on the sample sizes, types of distributions in the samples, any underlying assumptions about the data, and whatnot. A test should be carefully chosen to yield practical results.

Let T denote a function that performs the featured test thus: it takes two real matrices of the same dimensions as the input and then returns a real vector of p-values as the output. The function computes the p-values between the two corresponding columns of the same index in the two matrices. Therefore, the p-values, represented as a vector $\boldsymbol{p}$ ($\boldsymbol{p} \in \mathbb{R}^t$), can be computed as

$$\boldsymbol{p} = T(\boldsymbol{F}, \boldsymbol{F}^{'}). \tag{4.9}$$

Other statistics can be computed likewise.

Here, we are testing t hypotheses simultaneously. Thus, the problem arises as whether corrections for the multiple comparisons problem should be applied. In scientific research, corrections are often applied. However, in engineering robot personalities, it should depend on the circumstances of the specific engineering task. Whether it is better to minimise the chance of either Type 1 or Type 2 errors depends. If it is more important to reduce cost and focus on the most prominent personality traits, it might be better to apply the corrections so that it is less likely to select a personality dimension by chance while in truth optimisation cannot be effectively conducted on that dimension. If it is more important to utilise the full potential of the robots, so as to make them more 'characteristic-rich', it might be better to not apply the corrections so that it is less likely to disqualify a dimension by chance.

Test 2: Robot Personalities' Consistency Test

The consistency test follows the fidelity test. It tests Sub-Hypothesis 2 per dimension for every robot personality. This test requires data sets B and $\boldsymbol{L}$.

Consistency measures the strength of the consensus on a robot personality some observers can reach. Because a benchmark is yet to be established in the field, for the time being, the consistency of humans reporting on a human can be the standard for that of humans reporting on a robot in the same settings. We can use Bartlett's test as a consistency *parity* test. It tests *homogeneity of variance* while being sensitive to non-normality, meaning passing this one test is a sign for both normality and homogeneity of variance of the reported personalities.

Given a number of observations $\boldsymbol{x}_A$ ($\boldsymbol{x}_A \in \mathbb{R}^u$), as on a personality trait of an agent, and those on another $\boldsymbol{x}_B$ ($\boldsymbol{x}_B \in \mathbb{R}^u$), consistency parity p ($0 < p < 1$) can be computed as

$$p = Bartlett(\boldsymbol{x}_A, \boldsymbol{x}_B). \tag{4.10}$$

The p here is indeed a p-value. However, what can be considered as 'significant' in the context of consistency parity requires support from more empirical results. For now, any choices seem arbitrary. The convention of $p = 0.05$ can be a (very loose) significance threshold, and since we run r tests per dimension, we need to correct the p-value to $0.05/r$ by applying the Bonferroni correction to counter the multiple comparisons problem because we do not want to disqualify a dimension by chance, thereby relaxing it further. A consensus is topic-dependent, which means that a group of observers can reach a consensus separately on A or B, while a consensus on A and B together is nonsense. Consequently, we need to measure the consistency parities separately (per archetype per trait).

Let Bt denote a function that performs Bartlett's test, thus: it takes two real matrices of arbitrary numbers of rows but the same number of columns as the input, and it returns a real vector $\boldsymbol{p}$ of p-values as the output. In other words, the function computes the p-values between two corresponding columns of the same index in the two matrices. Let $\boldsymbol{P}$ denote the matrix that contains all p-values of consistency parities. It can be computed as

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{p}_1 & \boldsymbol{p}_2 & \boldsymbol{p}_3 & \dots & \boldsymbol{p}_r \end{bmatrix}^\mathsf{T}, \tag{4.11}$$

where $\boldsymbol{p}_{i=1,2,3,\dots,r} = Bt(\boldsymbol{L}, \boldsymbol{B}_i)$ and $\boldsymbol{p}_i \in \mathbb{R}^{1\times t}$, where $\boldsymbol{B}_{i=1,2,3,\ldots,r} \in \mathbb{R}^{u\times t}$, and

$$\boldsymbol{B}_i = \begin{bmatrix} \boldsymbol{b}_{1_i} & \boldsymbol{b}_{2_i} & \boldsymbol{b}_{3_i} & \ldots & \boldsymbol{b}_{u_i} \end{bmatrix}^\intercal, \tag{4.12}$$

where $\boldsymbol{b}_{m_i}$ ($m = 1, 2, 3, \ldots, u$) denotes the mth observer's report on the ith robot personality. To disqualify a dimension, we can consider the number of null hypotheses (consistency parity) rejected on that dimension. Considering that the consistency parity testing approach might not be strict enough, a stricter threshold for the number of rejected null hypothesis can be implemented here to enhance the effectiveness of the test.

Test 3: Robot Personalities' Fidelity 'Sanity' Test

Finally, the fidelity 'sanity' test checks whether we can reproduce the results in the fidelity test using the observers in place of the archetypes as the reference points. It tests Sub-Hypothesis 3 per dimension.

This test is almost the same as Test 1 except that we swap the archetypes with the observers themselves as the reference points. Fidelity is defined as the proximity between the observed robot personalities and their corresponding archetypes; therefore, the proximity between the observed robot personalities and the corresponding observers is not fidelity; instead, we call this quantity *pseudo-fidelity*. The test follows the exact same procedure as the fidelity test, only that this time the reference points are changed from the archetypes to the personalities of the observers themselves. This test requires data sets $\boldsymbol{H}$, B, and N.

We compute the pseudo-fidelity and generate s random reports to compute random traits as before. From N, we get u $r \times t$ matrices: $\boldsymbol{N}_1^{'}, \boldsymbol{N}_2^{'}, \boldsymbol{N}_3^{'}, \ldots, \boldsymbol{N}_u^{'}$ ($\boldsymbol{N}_{i=1,2,3,\ldots,u}^{'} \in \mathbb{R}^{r\times t}$). Because we need the proximity values between the observations of each observer and the observer themselves, we construct u $r \times t$ matrices: $\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3, \ldots, \boldsymbol{H}_u$ ($\boldsymbol{H}_{i=1,2,3,\ldots,u} \in \mathbb{R}^{r\times t}$), where

$$\boldsymbol{H}_{i=1,2,3,\ldots,u} = \begin{bmatrix} \boldsymbol{h}_i & \boldsymbol{h}_i & \boldsymbol{h}_i & \ldots & \boldsymbol{h}_i \end{bmatrix}^\intercal, \tag{4.13}$$

where $\boldsymbol{h}_i$ is the ith observer's personality.

Then, we apply the same procedure as in the fidelity test, which can be expressed as

$$\boldsymbol{F_o} = \begin{bmatrix} abs(\boldsymbol{B}_1 - \boldsymbol{H}_1) & abs(\boldsymbol{B}_2 - \boldsymbol{H}_2) & abs(\boldsymbol{B}_3 - \boldsymbol{H}_3) & \ldots & abs(\boldsymbol{B}_u - \boldsymbol{H}_u) \end{bmatrix}^\intercal \tag{4.14}$$

$$\boldsymbol{F}^{''} = \begin{bmatrix} abs(\boldsymbol{N}_1^{'} - \boldsymbol{H}_1) & abs(\boldsymbol{N}_2^{'} - \boldsymbol{H}_2) & abs(\boldsymbol{N}_3^{'} - \boldsymbol{H}_3) & \ldots & abs(\boldsymbol{N}_u^{'} - \boldsymbol{H}_u) \end{bmatrix}^\intercal. \tag{4.15}$$

Then, the p-values from this test, represented as a vector $\boldsymbol{p}_o$ ($\boldsymbol{p}_o \in \mathbb{R}^t$), can be computed as

$$\boldsymbol{p}_o = T(\boldsymbol{F_o}, \boldsymbol{F}^{''}). \tag{4.16}$$

4.4 Experiment: Testing the Selection Procedure

I tested the selection procedure in a scope as delimited by the following question: *given a user group, what are the personality dimensions to use in creating a type of life-size humanoid 'barebones' robot into ten robot personalities for human-robot dyadic communication, where the typical personality-expressing behaviour consists of head and eye movements?* Dyadic communication, albeit not a situation that needs a large number of robot personalities, was chosen for the abundance of literature on human-robot interaction since I needed to compare the experimental results with those in existing literature to know whether the proposed selection procedure was effective.

4.4.1 Materials: Robot and Software

The type of robot used to test the selection procedure was a humanoid 'barebones' robot (Figure 4.2). Its eyes and necks were actuated by pneumatic actuators with a typical response time of 200 milliseconds. The robot was controlled by a proprietary interface that received actuator commands for movements. A simple mapping program generated commands for the robot's eyes and head by mapping head and eye movement angles, which were captured by OpenFace 2 [83, 84, 85, 86], to actuator positions, so that, when the archetypes turned their heads or eyes left or right, up or down, the robot would do the same. The mapping was done without smoothing or filtering.

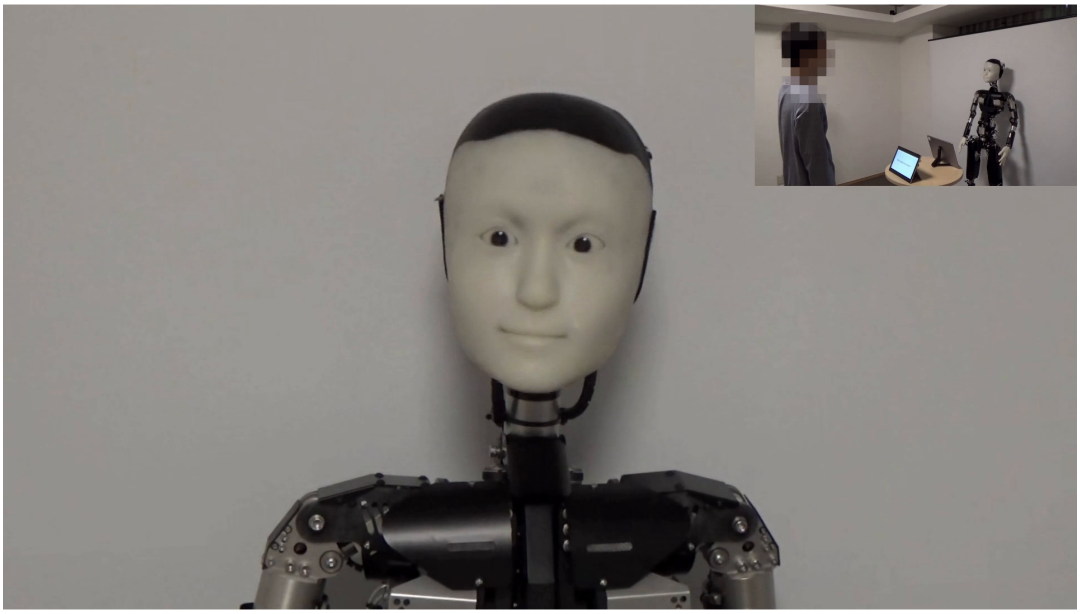

FIGURE 4.2: Tested robot in action as what the human observers viewed in Phase 3 (but without pixelating the actor's face) when they reported on the robot personalities.

The selection procedure was implemented in GNU Octave (Version 6.2.0) with the 'statistics' package (Version 1.4.2). The featured statistical test in

Tests 1 and 3 was Welch's test, which was executed by the **welch_test** function. The featured statistical test in Test 2 was Bartlett's test, which was executed by the **bartlett_test** function.

4.4.2 Going through the Procedure

The test settings were that 10 robot personalities were to be engineered out of 10 personality archetypes for small user groups of 3 to 18 people. The robot engaged users in dyadic communication, and the main modes of personality expression were head and eye movements. The personality assessment inventory used was by John et al. [87], which has 44 Likert scales for assessing five principal personality traits (dimensions): extraversion, conscientiousness, agreeableness, neuroticism, and openness. If we fill in the hypothesis format per the settings, it would be

Hypothesis (generic form). *Given **a humanoid 'barebones' robot** that **moves its head and eyes** in **dyadic communication** and a personality dimension **P** of **the five-factor model** (where $\boldsymbol{P} \in \{extraversion, conscientiousness, agreeableness, neuroticism, openness\}$), **P** is a personality dimension on which fidelity can be effectively optimised.*

I tested 505 such hypotheses in 101 tests corresponding to 101 user groups simulated by drawing from a pool of 18 observers. The first user group consisted of all 18 observers for a showcase run of the procedure (Sections 4.5.1–4.5.3). Then, I drew 20 combinations of 15 people, 12 people, 8 people, 5 people, and 3 people each, amounting to 100 simulated user groups in total. Therefore, the total number was 101.

The corresponding workflow is illustrated by Figure 4.3. The situation simulated in Phase 1 was two people exchanging information about themselves. 10 human archetypes were selected out of 15 candidates, who were students from our university. Each of the 15 candidates had come to our laboratory and engaged in a brief exchange with the actor. They first introduced themselves and then answered 12 questions adapted from those on a website for practising English conversation[1]. Then, the candidate and the actor switched roles: it was now their turn to ask the same questions, and the actor answered by reciting a script, which was recorded in Appendix B of this chapter. The head and eye movements of the candidates responding to the actor's answers were recorded as the archetypal behaviour for the robot personalities. The candidates also reported their own personalities thrice: once upon the application, once before the interview, and once after. Appendix A of this chapter relates the interview in more extensive details. In Phase 2, the 10 archetypes were selected—because their head and eye movements were within the robot's capability—and their corresponding videos (10- to 15-min long) shortened into videos of about 2 minutes' length according to an abridged version of the original script (the part in bold of the script recorded in Appendix B of this chapter makes the abridged version). The shortened

[1]The questions are for practising English conversation and adopted at the courtesy of the Internet TESL Journal via `http://iteslj.org/questions/personality.html`

videos were remade with the robot personalities replacing the archetypes. In the remakes, the robot personalities imitated the archetypes' head and eye movements under the control of the movement mapping software. The same actor answered their questions, which were shown on a screen. Although there were 10 remakes featuring 10 robot personalities, they were all installed on the same robot. Now, for Phase 3, I embedded the remakes into an online assessment form, which had 11 pages for personality assessment: 10 randomised pages for the robot personalities, and the 11th for the observer's own personality. I published the assessment form on Amazon Mechanical Turk and employed 18 workers to complete it. Finally, in Phase 4, I applied the three tests to the 101 simulated user groups and worked out the results.

4.5 Results from Testing the Selection Procedure

First, the results from all 18 observers as a single user group:

4.5.1 Robot Personalities' Fidelity Test

For the statistical test to use in the experiment, I chose Welch's test for the reason that our fidelity tests were going to feature large samples (30–180) with unknown variances. Plugging in the data, where $r = 10$, $t = 5$, and $u = 18$, to execute the tests using my implementation in GNU Octave, I obtained

$$\boldsymbol{p} = \begin{bmatrix} 0.002 & 0.580 & < 0.001 & 0.607 & 0.721 \end{bmatrix} \tag{4.17}$$

corresponding to extraversion, agreeableness, conscientiousness, neuroticism, and openness. The mean fidelity values are illustrated by Figure 4.4 with the corresponding standard deviations illustrated by the error bars. As for the significance threshold (α), I assumed the conventional 0.05 since I also assumed that I intended to discovery the full potential of the robot, which required us to minimise the chance of Type 2 errors. Therefore, I did not apply corrections for the multiple comparisons problem.

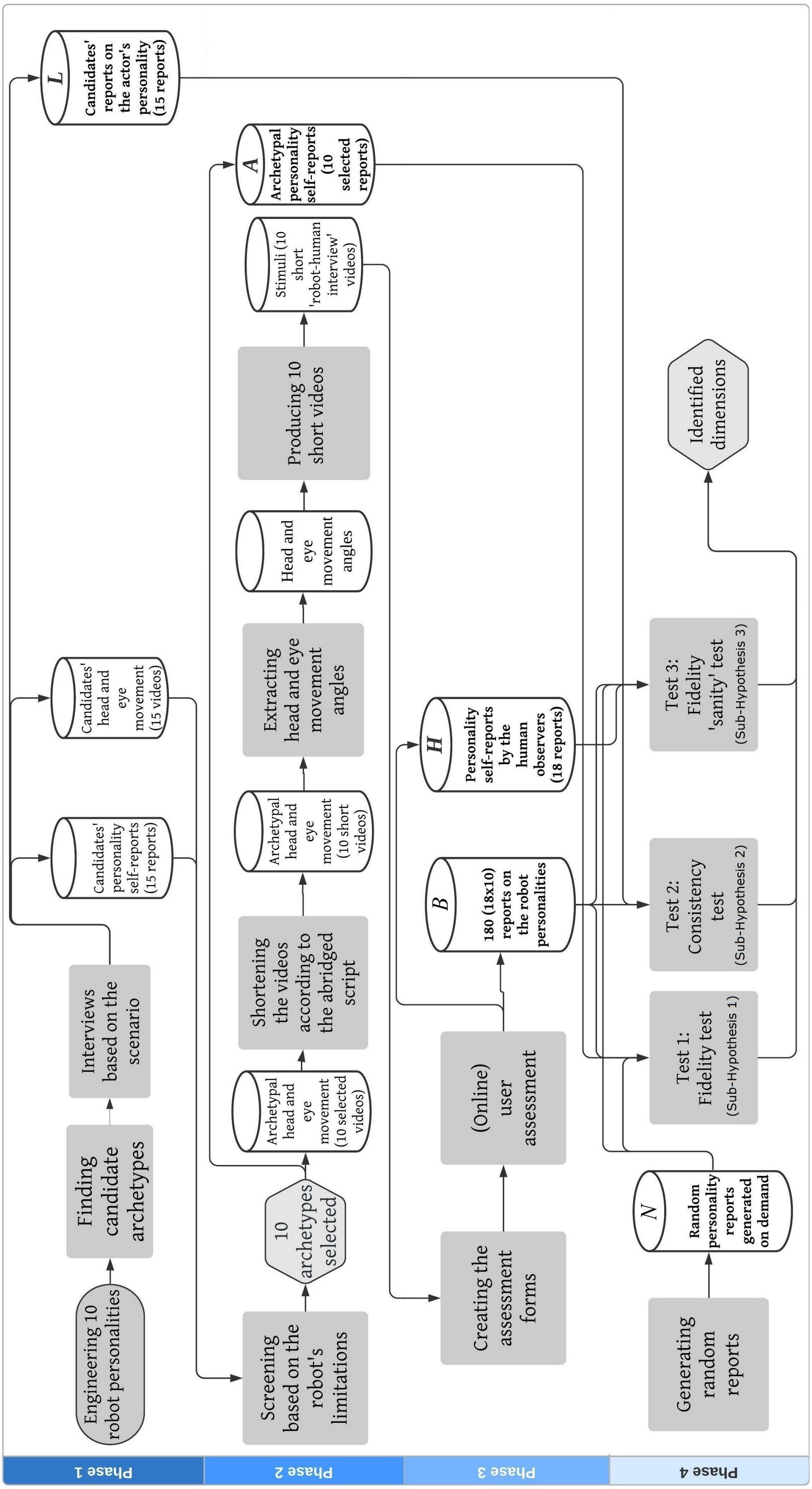

FIGURE 4.3: The workflow adapted to the experiment settings.

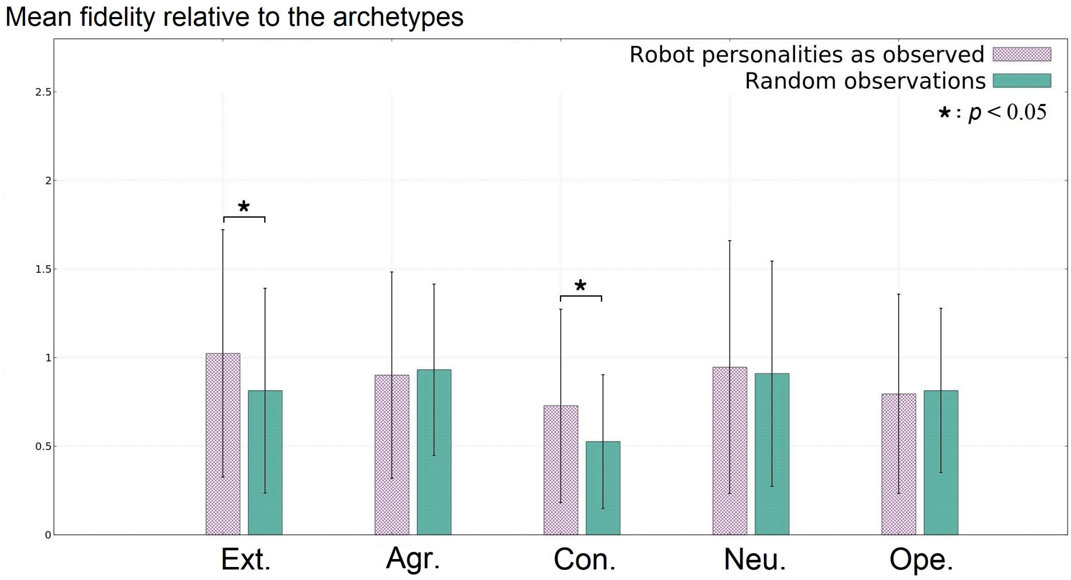

FIGURE 4.4: Results of the fidelity test (Section 4.5.1).

We can observe that (Figure 4.4), for extraversion, the fidelity by the human observers ($M = 1.022, SD = 0.698$)[2] compared with that by randomness ($M = 0.813, SD = 0.577$) demonstrated a significant difference, $t(345.76) = 3.09$, $p = 0.002$. (The difference here may seem small despite being statistically significant because fidelity is not a quantity of a large magnitude to start with.) As for conscientiousness, the fidelity by the human observers ($M = 0.727$, $SD = 0.546$) compared with that by randomness ($M = 0.525, SD = 0.378$) demonstrated a significant difference, $t(318.41) = 4.08$, $p < 0.001$. There was no significant difference in agreeableness, $t(346.33) = -0.55$, $p = 0.580$, between the fidelity by the human observers ($M = 0.900, SD = 0.582$) and that by randomness ($M = 0.931, SD = 0.483$). The same can be observed for neuroticism, $t(353.41) = 0.51$, $p = 0.607$, between ($M = 0.946, SD = 0.713$) and ($M = 0.909, SD = 0.636$); and for openness, $t(345.39) = -0.36$, $p = 0.721$, between ($M = 0.795, SD = 0.563$) and ($M = 0.814, SD = 0.464$).

The results above imply that extraversion and conscientiousness might be the personality dimensions on which the fidelity of the ten robot personalities could be optimised. The observers' reports on other dimensions were equivalent to random guesses.

4.5.2 Robot Personalities' Consistency Test

Plugging in our data to run all consistency parity tests (50 tests in total), I obtained the $\boldsymbol{P}$ matrix, which was transcribed into significance marks as recorded in Table 4.3, where an asterisk marks $p < 0.005$, which was the significance threshold acquired by applying the Bonferroni correction ($p = 0.05/10$) since it was undesirable to disqualify a dimension by chance. The consistency parity has held in most cases (in Bartlett's test, homogeneity of variance is the null hypothesis).

[2] M henceforth refers to 'mean', and SD 'standard deviation'

TABLE 4.3: Results of the robot personality consistency test (Section 4.5.2).

	Five Personality Dimensions				
Robots	**Ext.**	**Agr.**	**Con.**	**Neu.**	**Ope.**
Robot (as Archetype 1)	x	x	x	*	x
Robot (as Archetype 2)	x	x	x	x	x
Robot (as Archetype 3)	x	x	x	x	x
Robot (as Archetype 4)	x	x	x	x	x
Robot (as Archetype 5)	x	x	x	x	x
Robot (as Archetype 6)	x	x	x	x	x
Robot (as Archetype 7)	x	x	x	x	x
Robot (as Archetype 8)	x	x	x	x	x
Robot (as Archetype 9)	x	x	x	x	x
Robot (as Archetype 10)	x	x	x	x	x

x: The null hypothesis is not rejected. *: The null hypothesis is rejected.

We can see that, as for extraversion and conscientiousness, all test results in the corresponding columns failed to reject the null hypothesis of consistency parity. In the neuroticism column, the null hypothesis was rejected once. This means that we could fail the selection of neuroticism even if it had been selected in Test 1. How many rejections can fail the selection is at the moment a relatively arbitrary decision, and since the consistency parity test was not strict at all, I determined that one rejection was enough to fail the selection of the corresponding dimension.

4.5.3 Robot Personalities' Fidelity 'Sanity' Test

Plugging in our data, I obtained

$$\boldsymbol{p}_o = \begin{bmatrix} 0.534 & 0.692 & 0.420 & 0.598 & 0.120 \end{bmatrix}. \tag{4.18}$$

The means and standard deviations are illustrated by Figure 4.5.

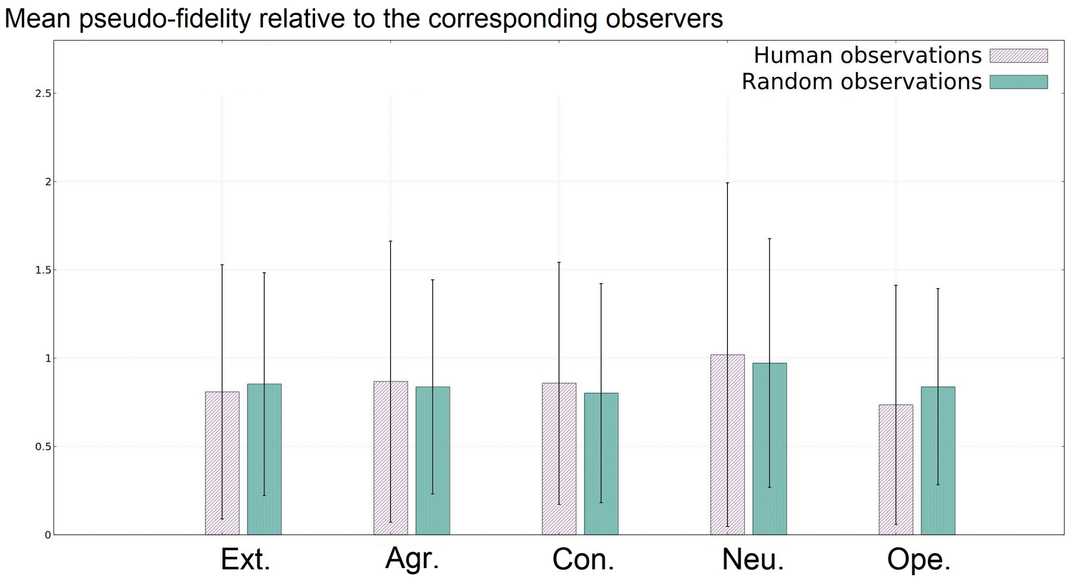

FIGURE 4.5: Results of the fidelity 'sanity' test (Section 4.5.3).

We can observe that (Figure 4.5), for extraversion, there was no significant difference, $t(351.89) = -0.62$, $p = 0.534$, in the pseudo-fidelity: that from between the personality reports and their corresponding reporters ($M = 0.808, SD = 0.720$) as compared with that by the random reports ($M = 0.852, SD = 0.631$). The same can be observed for agreeableness, $t(334.70) = 0.40$, $p = 0.692$, ($M = 0.866, SD = 0.796$) and ($M = 0.836, SD = 0.607$); conscientiousness, $t(354.56) = 0.81$, $p = 0.420$, ($M = 0.857, SD = 0.685$) and ($M = 0.801, SD = 0.620$); neuroticism, $t(326.14) = 0.53$, $p = 0.598$, ($M = 1.019, SD = 0.973$) and ($M = 0.972$, $SD = 0.704$); and openness, $t(344.83) = -1.56$, $p = 0.120$, ($M = 0.735, SD = 0.677$) and ($M = 0.837, SD = 0.556$).

The null hypothesis was rejected on neither extraversion nor conscientiousness. Therefore, the two dimensions were not disqualified.

4.5.4 Applying the Selection Procedure to Smaller User Groups

In the above, I have presented the showcase run of the procedure with all 18 observers. However, smaller user groups will be more common. Therefore, I conducted more tests on the 20 simulated user groups of 15 users, 12 users, 8 users, 5 users, and 3 users each—100 user groups in total. The results were recorded in Tables 4.4–4.8, respectively. The results varied, but extraversion and conscientiousness were the most frequently selected dimensions, followed by agreeableness and neuroticism, which were possible but much less frequent. For user groups of size-5 and size-3, the selections were sparse, and fewer than half of the user groups in each setting selected at least one dimension from the five. It was possible that the current method was only effective for user groups with more than eight people.

The variations imply that the personality dimensions selected previously, extraversion and conscientiousness, were not necessarily the case across all simulated user groups, which were still drawn from the same small pool of 18 observers. We can expect results to vary more in real situations. Personality perception varies from person to person and user group to user group, which is why a method such as the one proposed in this work will be useful.

TABLE 4.4: Results by 20 randomly drawn size-15 groups.

	Five Personality Dimensions [1]				
User Groups	**Ext.**	**Agr.**	**Con.**	**Neu.**	**Ope.**
User Group 1	*	-	*	-	-
User Group 2	-	-	-	-	-
User Group 3	*	-	*	-	-
User Group 4	*	-	*	-	-
User Group 5	*	-	-	-	-
User Group 6	*	-	*	-	-
User Group 7	*	-	*	-	-
User Group 8	*	*	-	*	-
User Group 9	*	-	*	-	-
User Group 10	*	-	-	-	-
User Group 11	*	-	*	-	-
User Group 12	*	-	*	-	-
User Group 13	*	-	-	-	-
User Group 14	*	-	-	-	-
User Group 15	*	-	*	-	-
User Group 16	*	-	*	-	-
User Group 17	*	-	*	-	-
User Group 18	*	-	-	-	-
User Group 19	*	-	*	-	-
User Group 20	*	-	*	-	-

[1] Selected personality dimensions are marked with an asterisk.

4.5.5 Summary of the Results

In summary, most tests featuring simulated user groups with eight or more users selected extraversion. Conscientiousness was selected in many tests as well but significantly less frequent than extraversion. The selection of extraversion agreed with several studies about the head and eye movements or gaze of humanoid robots or virtual agents [2, 29, 88, 31, 89], as did the selection of conscientiousness [88, 31]. However, the two dimensions were by no means the only possibilities. Agreeableness and neuroticism were possible as well, though much less frequent. All results agreed that openness was not an optimisable dimension for the type of robot tested for dyadic communication.

TABLE 4.5: Results by 20 randomly drawn size-12 groups.

	Five Personality Dimensions [1]				
User Groups	Ext.	Agr.	Con.	Neu.	Ope.
User Group 1	*	-	-	-	-
User Group 2	*	-	*	-	-
User Group 3	*	-	-	-	-
User Group 4	*	-	-	-	-
User Group 5	*	-	-	-	-
User Group 6	*	-	*	-	-
User Group 7	*	-	-	-	-
User Group 8	*	-	*	-	-
User Group 9	*	-	-	-	-
User Group 10	*	-	*	-	-
User Group 11	*	-	-	-	-
User Group 12	*	-	-	-	-
User Group 13	*	-	*	-	-
User Group 14	*	-	*	-	-
User Group 15	*	-	-	-	-
User Group 16	*	-	*	-	-
User Group 17	*	-	-	-	-
User Group 18	*	-	*	-	-
User Group 19	*	-	*	-	-
User Group 20	*	-	-	-	-

[1] Selected personality dimensions are marked with an asterisk.

4.6 Discussions

The selection procedure stood 101 limited-scope tests, where I simulated an engineering task where 10 humanoid robot personalities were to be engineered for dyadic communication for small user groups of 18, 15, 12, 8, 5, and 3 users. The corresponding test results for simulated user groups of eight or more people reflected the existing body of knowledge on the potential of humanoid robots for expressing personality traits, thereby confirming the effectiveness of the proposed selection procedure within the scope of the experiment for user groups with at east eight people. However, the variations revealed should allow us a glimpse into a common situation in creating robot personalities, which is that no one answer applies to all cases. The working hypotheses—that not all but some human personality dimensions can be used to create robot personalities for non-android robots that can imitate human behaviour—are supported as least within the scope of the experiment. The proposed selection procedure is the first one dedicated to creating robot personalities in significant quantities for small user groups. It is inevitably primitive, and there is much space for improvement. There are several limitations in particular we can focus on.

TABLE 4.6: Results by 20 randomly drawn size-8 groups.

	Five Personality Dimensions [1]				
User Groups	**Ext.**	**Agr.**	**Con.**	**Neu.**	**Ope.**
User Group 1	*	-	*	-	-
User Group 2	*	-	-	-	-
User Group 3	-	-	-	-	-
User Group 4	-	-	-	-	-
User Group 5	*	-	-	-	-
User Group 6	*	*	-	-	-
User Group 7	*	-	-	-	-
User Group 8	-	-	-	-	-
User Group 9	*	-	*	-	-
User Group 10	-	-	-	-	-
User Group 11	*	-	-	-	-
User Group 12	*	-	*	-	-
User Group 13	*	-	-	-	-
User Group 14	-	-	-	-	-
User Group 15	*	-	*	-	-
User Group 16	*	-	*	-	-
User Group 17	*	-	*	-	-
User Group 18	*	-	-	-	-
User Group 19	*	-	-	-	-
User Group 20	-	-	*	*	-

[1] Selected personality dimensions are marked with an asterisk.

The primary limitation of the selection procedure is that it is currently ineffective for user groups with fewer than eight people. Compatibility with smaller user groups is the most needful improvement considering that most household user groups probably have fewer than eight people. A possible workaround is to combine two or three small groups to meet the group size requirement. Another limitation is that the selection procedure is not a guarantee for the applicability of personality dimensions based on the lexical hypothesis. For guaranteed applicability, we will need to develop dedicated descriptive personality models for robots based on lexicons for describing robots, as a previous study did for conversational agents [6]. Dedicated models can be used together with a modified selection procedure to select optimisable dimensions for a type of robot. The necessary modification is to replace human archetypes with artificial archetypes since robot personality models do not necessarily apply to humans. Considering that robots come in different designs for different purposes, even a dedicated model does not necessarily fully apply to a particular type of robot.

The primary limitation of the experiment is that I only confirmed the effectiveness of the selection procedure in a limited scope. Universal applicability is to be further tested. Future research can also aim to expand the applicability of the proposed selection procedure and validate it in a broader

TABLE 4.7: Results by 20 randomly drawn size-5 groups.

	Five Personality Dimensions [1]				
User Groups	**Ext.**	**Agr.**	**Con.**	**Neu.**	**Ope.**
User Group 1	*	-	-	-	-
User Group 2	*	-	*	-	-
User Group 3	-	-	-	-	-
User Group 4	-	-	-	-	-
User Group 5	-	-	-	-	-
User Group 6	-	-	-	-	-
User Group 7	*	-	*	-	-
User Group 8	-	-	-	-	-
User Group 9	-	-	-	-	-
User Group 10	-	-	-	-	-
User Group 11	*	-	-	-	-
User Group 12	*	-	*	-	-
User Group 13	*	-	-	-	-
User Group 14	-	-	-	-	-
User Group 15	-	-	-	-	-
User Group 16	-	-	-	-	-
User Group 17	-	-	*	-	-
User Group 18	-	-	-	-	-
User Group 19	-	-	-	-	-
User Group 20	-	-	-	-	-

[1] Selected personality dimensions are marked with an asterisk.

context. In particular, although the selection procedure is meant for creating colonies of robots, such as a music band or idol group of robots or a staff of robotic errand runners in an office, the scope of the experiment did not properly reflect this aim. The experiment featured multiple instances of dyadic communication rather than users interacting with multiple robot personalities at the same time. It is arguable that dyadic communication may still represent the most common form of interaction. For example, in an office where humans work with multiple errand runners, it is still more common for a human to interact with one of them at a time. Still, an important future work is to cover interaction involving multiple physically present robots. This would require multiple units of the type of robot to study and abolishing video-based stimuli. In addition, the content of the interaction featured in the experiment, which was about getting to know each other, albeit important for leaving first impressions, did not necessarily reflect the many possible forms of interaction with interactive robots. Future studies and implementations of the selection procedure would benefit from settings that better reflect the contents of the interaction based on realistic roles of interactive robots. Another major limitation of the experiment was the simulation of 101 user groups by drawing from a small pool of 18 observers. There were two possible impacts of this approach: underrepresented inter-group

TABLE 4.8: Results by 20 randomly drawn size-3 groups.

	Five Personality Dimensions [1]				
User Groups	**Ext.**	**Agr.**	**Con.**	**Neu.**	**Ope.**
User Group 1	-	-	-	-	-
User Group 2	-	-	-	-	-
User Group 3	-	-	-	-	-
User Group 4	*	-	*	-	-
User Group 5	-	-	-	-	-
User Group 6	-	-	-	-	-
User Group 7	*	-	-	-	-
User Group 8	-	-	-	-	-
User Group 9	*	-	-	-	-
User Group 10	*	-	*	-	-
User Group 11	*	-	*	-	-
User Group 12	-	-	-	-	-
User Group 13	-	-	*	-	-
User Group 14	*	*	-	-	-
User Group 15	-	-	-	-	-
User Group 16	*	-	-	-	-
User Group 17	-	-	-	-	-
User Group 18	-	-	-	-	-
User Group 19	-	-	-	-	-
User Group 20	-	-	-	-	-

[1] Selected personality dimensions are marked with an asterisk.

variations and intra-group consistency. A main goal of the selection procedure is to take into account the inter-group differences between different user groups when creating robot personalities. However, by drawing from the same pool of 18 observers, the possible inter-group variations were not properly reflected in the results (Tables 4.4–4.8). The intra-group consistency by randomly combining members of the 18 online observers, who observed the robot personalities by watching short videos, might be lower than that from a real user group interacting in the same environment with real robots.

Another limitation present in both the proposed selection procedure and the experiment was introduced by using an imprecise personality measure. This is also a limitation of the field, which currently lacks a personality measure precise enough for engineering tasks. The use of an imprecise measure based on Likert scales not only introduces possible incompatibility issues with statistical tests but affects the resolution of presenting archetypes as well. In practical engineering work, we cannot let two archetypes have the same trait on any dimension. However, traits measured by an imprecise measure can be the same in terms of their numeric trait levels. Therefore, using human archetypes as measured by an imprecise personality measure should be limited to selecting personality dimensions for constructing generative personality models; but not for optimising personality constructs, especially of large quantities, such as when we need to create hundreds or thousands of robot personalities using mass-produced robots. To that end, we will need a dedicated descriptive personality model with precise, continuous scales for designing and measuring artificial archetypes in greater numbers with better precision. But then again, it is doubtful whether this dedicated model for robot personalities can be used by human archetypes to report their own personalities. The future of robot personalities engineering might need to go beyond imitating humans.

I have proposed a selection procedure for selecting with the help of a user group personality dimensions to create robot personalities out of a type of robot, knowing its typical usage; it selects personality dimensions on which users can provide effective feedback to guide the optimisation of robot personalities. There are mainly three ways to use this tool. First, to select personality dimensions to construct a generative personality model. For this usage, the user group size limitation and the user-dependency are not issues since all dimensions that are potentially feasible can be included and hence a large sample of users is preferable for approximating the 'average' observer (please refer to Section 1.2.4 for what the 'average' observer is). Second, to select personality dimensions to focus on for optimising the personality constructs of some specific robots for specific users. For this usage, we need the method to work for small user groups. Third, to help check whether robot personalities are successfully created with the intended traits.

All things considered, the proposed selection procedure is a tool to generalise the possibility of creating robots with individuality to other types of robots than androids. To realise the possibility, however, requires a generative personality model, which was the topic of Study 3.

Appendix A: Interview

The interviews were conducted to capture the head and eye movements of the candidate archetypes. The sessions took place in a soundproof room, where a round white table was set in the middle, between two chairs placed on opposite sides. The candidate and our interviewer (the actor) sat face-to-face on these chairs. On the table were two cameras positioned side-by-side, one pointing at the candidate's face right in front of them on the eye level and the other, a gimbal camera, set to track the candidate's head just in case they moved out of the field of the fixed camera (this never happened). The clips shot by the fixed camera were used as the sources of head and eye angles. There was also a tablet computer set on a tablet stand for displaying the questions asked during the interviews. Its screen first faced the interviewer. Behind the interviewer to the right was another camera on top of a tall tripod for logging the interviews. When an interview session started, the interviewer requested the candidate to briefly introduce themselves. After the candidate finished, the interviewer proceeded to ask questions from the tablet. There were 12 questions in total, all selected from a website[3] for

[3] http://iteslj.org/questions/personality.html

practising English conversation. The interviewer proceeded to ask the 12 questions sequentially. The candidate had been told that they could skip questions just in case some questions were too personal; all candidates answered all questions nonetheless. The interviewer listened to the candidate answering the questions. In about 5 to 10 minutes, the interviewer exhausted the list. He then proffered the tablet to the candidate. From now on, it was the candidate's turn to ask the same questions in the same order, only that the interviewer (now being interviewed) would answer all questions according to a script he had memorised beforehand. This part of the interview was controlled and it took about five minutes for most candidates. When the candidate exhausted the list, the interview was over. Finally, the candidate reported their own personalities (for the third time) and then the interviewer's before they left.

Appendix B: Conversation Script

The below is the script the actor used to answer the questions from the participants of the interview. Part of the script also appeared as the conversation script of the human-robot interaction video stimuli. The part that made it to the videos is highlighted in bold. (The answers in the script represent no views or opinions of any real person or group.)

> *Q1. What are some characteristics of your personality?*
>
> *A1. I am introverted... very conscious of myself but not so much of others. I might not be very likeable, but the people who know me well may consider me a good companion or ally. I'm wont to have unorthodox views about things, but at the same time, I could be very conservative—a paradox as people might say.*
>
> ***Q2. What makes you happy?***
>
> ***A2. Oh! Many things make me happy, such as a good meal or a good book. In general, I'm happy when things have gone my way.***
>
> *Q3. Would you like to be different?*
>
> *A3. That... depends on how different I would be. Were I to be offered to have my personality changed drastically, for good or ill, I think I would decline without hesitation. However, if the changes were minor and for the good, I think I wouldn't mind.*
>
> ***Q4. Are you a determined person? Are you a stubborn person?***
>
> ***A4. I'm determined to achieve the goals I deem important. In addition, I can be very stubborn—especially about doing the right things. However, I'm less determined and stubborn in non-essential things. For instance, I don't mind very much having my lunch menu suddenly changed or vacation cancelled—even when I am looking forward to them.***

Q5. What is one thing that many people don't know about you?

A5. That I'm evil and often dream of taking over the world. Just kidding! *A thing that many people don't know—which I suspect that my parents are no exceptions—is my unwonted mischievous side. It is a point a bit difficult to elaborate on, but I'll try. For most of the time, I'm darn serious. However, when chances come, I might try to play some harmless tricks. For example, there was one time when I was in an English class where the students were assigned to give a presentation on a topic of their interest. I made a presentation on fictional villains. I tried to imitate the demeanour of some villains—mainly how they spoke—as how I introduced them. However, there was hardly any reaction from my classmates, though. Perhaps my villain impressions were bad.*

Q6. Do you think your personality has changed over the past few years? In what way has it changed?

A6. I don't think my personality has changed much over the past few years. That being said... discernible changes did occur, with the most prominent one being my growing indifference towards things that do not concern me directly.

Q7. Do you think you can change a major characteristic of your personality if you try?

A7. I don't think that's possible—to cause fundamental changes in one's personality through personal effort. What seems to be possible, however, is to change one's behaviour through conscious effort... so that other people might have a different impression of you. I think... it is how celebrities maintain their personas.

Q8. If you could change any aspect of your personality, what would it be?

A8. Tough question. I wouldn't want any major changes, whatever they were, so... um... what's the aspect of my personality that's hamstringing me the most? Mm... I'm wont to do things on my own and less inclined to collaborate with others. In addition, I seldom do things together with people just for the sake of doing things together that is, I seldom go shopping, go to the movies, or have lunch with other people. Being such a lone wolf has its pros and cons. However, I think it would benefit me if I tune-up my inclination for group activities just a little bit.

Q9. Are you shy? In which occasions are you shy?

A9. I am shy, but the stereotypical 'Asian' shy. Generally speaking, I am shy in front of people I don't know well.

Q10. Are you more introverted or more extroverted?

A10. I am extremely introverted... but not the way how introverts are introverted. Most introverts are easily associated with shyness; they are *introverted because they are not very social–even a little afraid of people–at least that's the stereotype. It's not my case, though. I'm introverted because I don't care much about other people.*

Q11. Do you think you have an unusual personality? Why?

A11. I don't think so. My personality is just the product of me being a person optimised for the environment he is in; in other words, I am who I am because this is how things will go the easiest for me.

Q12. Do you consider yourself to be even-tempered?

A12. I'm very even-tempered. In fact, I never get angry. However, again... I suspect that my emotional stability comes from my lack of concern for many things.

Chapter 5

Generative Personality Model

5.1 Research Question

The third research question is a *HOW* question: *how to implement individuality?* The answer is a generative personality model, which by definition, is capable of generating behaviour with inter- and intra-individual differences. Individuality refers to personality traits that distinguish an individual from others of the same kind. Personality traits are individual differences, only qualified and quantified, and hence a generative personality model naturally implements individuality. But we must then ask: *how does a generative personality model work? And how many individuals can it present?*

I proposed in the third study a working principle for generative personality models, implemented a proof-of-concept prototype model, and conducted an experiment to study the model's effective colony capacity when it was used to generate non-verbal behaviour on the head of a humanoid robot [4].

The rest of this chapter is organised as follows: Section 5.2 explains the proposed generative personality model; Section 5.3 specifies the materials for the experiment—the prototype model and the robot test bed; Section 5.4 relates the experiment; Section 5.5 presents the results; and Section 5.6 discusses and concludes the study.

5.2 Method: Goal-Shaped Generative Personality Model

5.2.1 Philosophy of the Proposed Model

The philosophy is that the structure of a robot's personality can be shaped to facilitate the attainment of a goal even when there is no reliable knowledge on whether certain behaviour helps achieve the goal, which is why I describe the proposed model as 'goal-shaped'. Similar to a few counterparts, such as that of Chittaro and Serra [62], the proposed model uses personality trait levels as probabilistic influences over behaviour. The main difference is that the proposed model does not require associating behaviour with goals and is thus more applicable to robots in the real world. As in the real world, one often does not know for certain whether some behaviour is conducive or leads to the attainment of a goal. (Does studying hard in school lead to success? Or is it that successful people tend to work harder than average as one of their traits?) The proposed model associates behaviour with personality traits and then personality traits with goals.

5.2.2 Working Principle of the Proposed Model

However, our objective is to create robot personalities in significant quantities out of mass-produced, physically identical robots. How does the philosophy serve this objective? For the answer, I propose this working principle: *the root of all inter- and intra-individual differences in computation and behaviour is different goals*. Different personalities behave differently in the same context because their goals are different. A personality behaves differently in different contexts because their goals are different in different contexts. The goal-shaped structure of the proposed model is how it generates behaviour with inter- and intra-individual differences. Creating robots with inter-individual differences is the most crucial for applications requiring significant numbers of robots that interact with humans. Imagine someone going to work. They see one of their robot errand runners and say hello. But a few steps further, they bump into another, looking the same as the one before and offering greetings in the exact same manner, and then another, followed by another, and again, and again, and again, and one cannot get to their cubicle until this has happened a dozen times (every day). The image of those robots is that of many copies of the same machine. The interaction they offer is grimly tedious, reminiscent of a working environment where everyone is but a replaceable cog in the machine with no individuality. It would be significantly more interesting if the errand runners possess individuality, formed and shaped in the course of them working alongside their human colleagues, and interact with people in their own unique ways.

5.2.3 Personality AI for Robots and the Proposed Generative Personality Model

Consider the robot AI shown in Figure 5.1. It is, generally speaking, of the 'Sense-Plan-Act' paradigm but with a generative personality model[1] in the 'Plan' part. It consists of a sensory system to perceive the world to produce sensory output, which a cognitive system takes as its input for cognitive output, such as understanding of situations, which the personality model takes as the input to generate behaviour, which includes external behaviour (movements, actions, social behaviour, and anything observable) and internal behaviour (predispositions, moods, emotions, and anything not directly observable). Indeed, the 'behaviour' here refers not only to specific instances of actions but a range of possibilities, from the activation of a process to running a long-term scheme. Saying hello is behaviour. Committing oneself to

[1]The generative personality model in Figure 5.1 is an example. A personality construct based on the proposed model does not necessarily have five trait nodes, six goal nodes, and a goal node or more connected to each trait node.

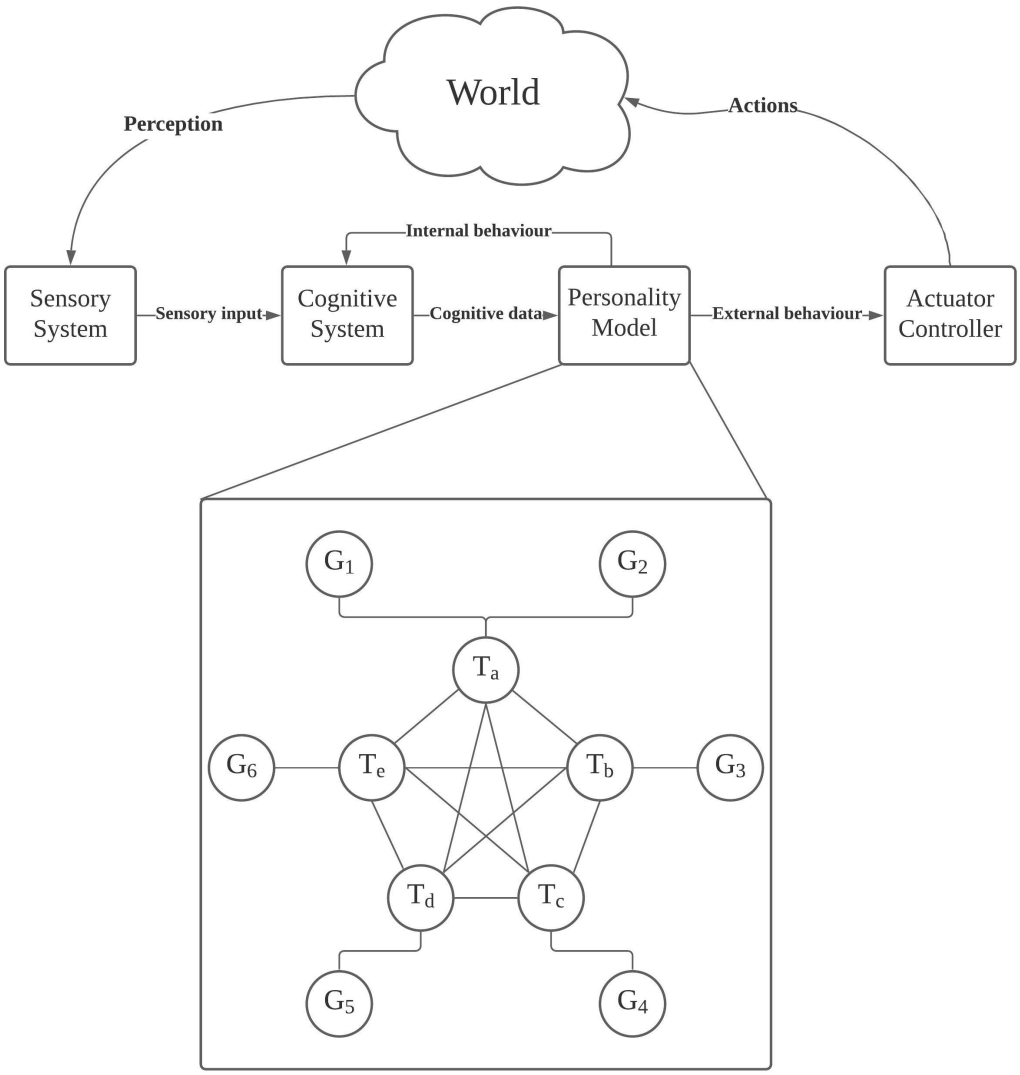

FIGURE 5.1: A diagram for an example model (as part of an example personality AI).

pursue a goal over a long period of time is also behaviour. Everything an AI can do can be behaviour and can be influenced by its personality. Finally, the actuator controller interprets external behaviour into actuator commands, whereas the internal behaviour acts back upon the personality and cognition. This particular study does not aim to cover every part of the personality AI but focuses on the personality model, which should be a generative model.

The generative personality model consists of a graph $\mathbf{\Gamma}$ of nodes $\mathbf{V}$ and edges $\mathbf{E}$. Formally,

$$\mathbf{\Gamma} = (\mathbf{V}, \mathbf{E}). \tag{5.1}$$

There are two types of nodes: goal nodes $\mathbf{G}$ and trait nodes $\mathbf{T}$:

$$\mathbf{V} = \mathbf{G} \cup \mathbf{T}. \tag{5.2}$$

A ***goal node*** represents a ***goal***, which is the objective the robot is operating to achieve, and hence a goal may also be called an *objective*, as in 'objective function', or a *directive*, as often so called in fiction. The exact implementation of a goal node depends on the type of robot and its purposes or typical usage. Generally speaking, a goal node should report a measurable difference between the *actual state* and *goal state*. For instance, if a personal trainer robot is tasked to make sure that their client exercises for 30 minutes a day, every day the goal node output would start at 30 minutes, that is the difference between the actual state, 0 minute, and the goal state, 30 minutes. As soon as the robot has nudged their client into exercising, the goal node output starts to decrease, and when they have exercised for 16 minutes, the goal node should report 14 minutes as the difference between the actual state and goal state. When the difference becomes zero, the goal is achieved for the day. Certainly, not all goals admit measurements that can be so simply defined. But measurements are nonetheless necessary. The goal nodes may be arranged into a structure, such as a tree [61], to encode priority (not shown in Figure 5.1) and thus resolve conflicts between goals if the robot has multiple goals. Structures of goals are beyond the scope of this work.

The trait nodes form a clique. A ***trait node*** represents a broad trait; the purpose of the node is to use the corresponding trait level and cognitive input to generate behaviour as the output. The cognitive input consists mainly of cognitive states; a ***cognitive state*** is an understanding of or response to a situation, which can be a belief, thought, or any other simulated psychological state. We consider such an implementation of a trait node: it consists of a ***mapping function*** that maps cognitive states to behaviour:

$$f_m : \mathbb{S} \to \mathbb{B}, \tag{5.3}$$

where $\mathbb{S}$ denotes the space of cognitive states, and $\mathbb{B}$, the space of behaviour. And a probabilistic ***commitment function*** that, as the name suggests, generates *commitment* to the potential behaviour from the mapping function. Given a cognitive state, the trait node, using the mapping function, first generates an ***urge*** for certain behaviour, which is an instance of potential behaviour not yet coming into action, with a probability scaled from the trait level. For example, 0.629 is the probability of generating an urge by a trait node with a trait level of 4.4 out of 7. In this study, we consider trait levels on a range of $[1, 100]$ for convenience. For behaviour of which the tendency is the complement of that suggested by the trait level, such as introverted behaviour managed by an extraversion trait node—the probability to generate an urge is computed from the complement of the corresponding probabilistic tendency of extraversion. For instance, urges of introverted behaviour are generated at a probability of 0.3 when the extraversion trait level is 70 (out of 100). The mapping function having generated an urge, the commitment function then computes the probability of achieving or not achieving the current goal given event B:

$$P(G \cap B) = P(G|B)P(B) \tag{5.4}$$

and

$$P(G' \cap B) = P(G'|B)P(B), \tag{5.5}$$

where G denotes the attainment of the goal, that is the goal node reporting no difference between the actual state and goal state, B denotes putting the urge into action, and $P(G|B)$ and $P(G'|B)$ are conditional probabilities provided by the cognitive system based on its understanding of how the world works. When there is not sufficient understanding to acquire $P(G|B)$ and $P(G'|B)$, the model assumes the two to be equal: $P(G|B) = P(G'|B) = 0.5$. If $P(G' \cap B)$ is bigger than $P(G \cap B)$, indicating that the behaviour will work against the goal, the urge is *suppressed*. It is entirely possible that to do nothing might serve the goal best, that is when $P(G \cap B')$ is higher than both $P(G \cap B)$ and $P(G' \cap B)$, thereby making the commitment a less than optimal choice either way. We might still want personality traits to function based on action rather than inaction since robots are designed to do things and probability is not certainty. We call a personality with trait nodes functioning based on action a ***proactive personality***. (That said, the possibility is there to implement a non-proactive personality should the circumstances demand it.) Consider that, as with humans, oftentimes, a personality trait does not contribute to a goal in the best way possible—if it does not work against the goal. We might still want to leave it be as it may contribute well to another goal.

If the reader wonders how the model handles certainty of actions required from the robot, such as following orders and sticking to daily routines, the behaviour we consider here is not necessarily specific actions; it can be a state that guarantees certain actions. When such a state, say sticking to daily routines, is switched on by the generative personality model, the robot will stick to daily routines without fail, until changes of circumstances trigger another state as a new behavioural pattern that requires the robot to break out of their daily routines. The same may apply to following orders: when robots committed themselves by their personalities to following orders, they will follow all orders issued to them until the model determines that it is no longer necessary.

A personality trait may not contribute to a goal in the best way possible, and by the following design, this trait may not be in a prioritised position to contribute. For each goal of a robot, there exists an *order* of trait nodes arranged in accordance to their contribution to the goal: a trait node that contributes more to the goal will be ranked higher in the order. There can be many approaches to acquiring an order. Here I proffer two heuristics. **Heuristic A:** If a robot can develop their trait levels to facilitate the attainment of a goal or is thus developed, a way to acquire an order is to sort trait nodes by trait levels from high to low, since that, if a trait node contributes more to a goal, the corresponding trait level may in time grow higher or be made so in the first place. However, this is not always the case unless the goal is the most important to the robot, and there might be cases where being low on a personality dimension contributes most to a goal—but the corresponding trait node will never be ranked high by the heuristic. This heuristic should be used with caution. As a rule of thumb, I recommend using Heuristic A only when there is no better choice. Another heuristic is to assign scores to all trait nodes. **Heuristic B:** Beginning with equal scores, say zeros, and whenever the behaviour triggered by a trait node has reduced the goal node output, the score of the corresponding trait node shall be increased. The trait nodes will be sorted by the 'contribution' scores according to preset time intervals to acquire or update the corresponding order.

Robots who have similar goals but in different environments may have different trait orders. For example, neuroticism—a trait that relates to avoidance motivation [90], as to avoid undesirable possibilities or results—of a rover on Mars may be ranked higher than a rover on the Moon since a Mars rover may need to avoid dust storms, whereas a Moon rover faces no such danger. The trait of the highest-ranking trait node concerning a goal is the ***primary trait*** for the goal. A robot's most important goal of their existence is the robot's ***mission***; the primary trait for the robot's mission is the robot's ***dominant trait***. It can be expected that robots in the same colony have the same mission (same in terms of the kind of objective, not necessarily how their goal node functions are defined), since it is the purpose of those robots' existence and shall exist over a long time if not to last throughout the robots' service lifespan. A robot works towards their mission by achieving a series of short-term, concrete goals just as humans do, and those short-term goals may differ according to the environments and circumstances the robot finds themselves in. Thus, there exists a causality gap between the robot's concrete behaviour and mission. This gap can be helpful since some behaviour that seems inconsequential or even detrimental to the mission may be necessary to a short-term goal that is essential to the mission. Short-term goals may be broken down into shorter-term goals and the same situation may persist. This is why it is important to dissociate behaviour with goals and reconnect the two with personality.

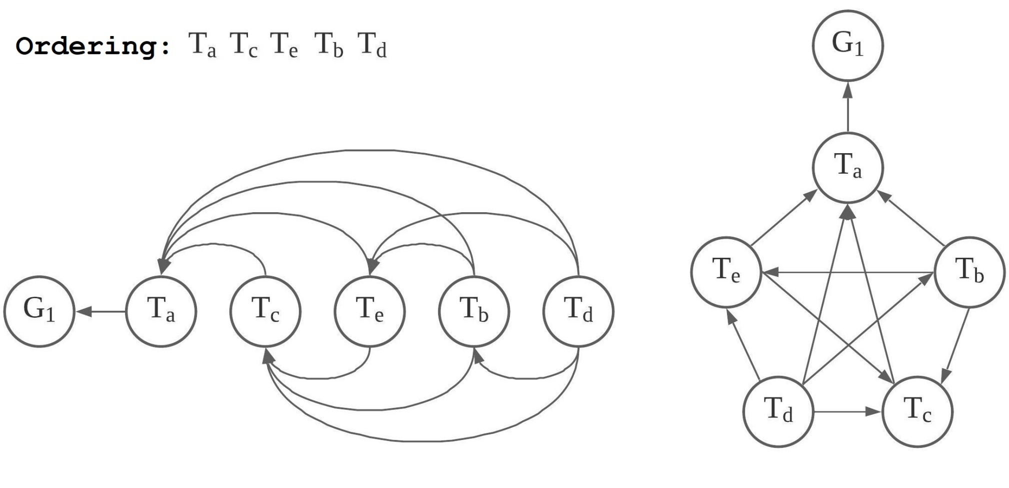

FIGURE 5.2: An example of a goal-active form of the example model in Figure 5.1.

Given a goal, which demands an order of traits, we can transform the generative personality model into the corresponding ***goal-active form*** (e.g. Figure 5.2); the transformation is thus done: we look only at the goal towards which the robot is currently working, namely the robot's ***active goal***, and the clique of trait nodes, then we convert the graph into a directed one by designating directions to the edges following such a rule: for every pair of trait nodes, the orientation of the edge between them is always from the node lower in the order to the node higher, and the edge from the highest node in the order points to the goal node.

When the robot is working towards a certain goal, the generative personality model of the robot is in the corresponding goal-active form. For the time being, we do not consider a robot capable of working concurrently towards multiple goals. The robot works towards one goal at a time, and if they need to work towards multiple goals, their generative personality model has to switch between the corresponding goal-active forms as how a CPU core handles multitasking. How a robot switches between goals, the interruption mechanisms and such, is beyond the scope of this work. I only assumed that the robot's mission stays active by default when there is not another goal being active. The goal-active form corresponding to the robot's mission is the default goal-active form of the robot's personality construct. When a robot's personality construct is in a goal-active form, it generates behaviour out of a cognitive state following these rules, from the highest trait node (the one connected to the goal node) in the order to the lowest one by one, check against all of the connected lower nodes, and if 1) the behaviour required by a lower node is in conflict with that of the current node, the urge for the behaviour is nullified, and the lower node is marked 'to be skipped'; 2) the same, the behaviour is enhanced, and the lower node is marked 'to be skipped'; 3) neither in conflict with nor the same as the current node, do both; 4) skip all nodes previously marked 'to be skipped'. Here, by 'in conflict with' I mean that only the potential behaviour of one of the trait nodes can be put into action. By 'enhanced' I mean 'done with more effort', which can be interpreted according to the behaviour. For example, a smile can be enhanced into a grin; working can be enhanced into 'working hard' and further into 'working harder'. We can implement the rules and designs stated above into algorithms. Algorithm 1 is an example. The details can vary, such as the handling of *effort*, which need not be discrete levels enhanced by increments.

If the reader wonders how a generative personality model that changes into different goal-active forms can ensure the consistency of the impression of robot personality, it does not—to the fullest extent. As in real life with humans, a person who devotes themselves to a goal may 'seem like a different person'. For instance, a person who is introverted but works as a receptionist may be constantly offering warm smiles and promptly engaging customers as if they were an extrovert. Oftentimes, a person may simply act 'out of character' in certain situations. If a robot—when in a goal-active form pertaining not to their mission—exhibits an impression of personality different from that of their true personality, we call this impression of personality their ***persona***. Normally, the true impression of personality is revealed when the

Algorithm 1 An Example Algorithm to Generate Behaviour

```
Input: cognitive state
Output: B
s ← cognitive state
for all v ∈ T do
    Initialise doublet (u, l)                          ▷ u: urge; l: effort level
    u ← v.mapping(s)   ▷ mapping: the mapping function of the trait node.
    if v.commitment(u) = true then          ▷ commitment: the commitment function of the trait node.
        l ← 1
    else
        l ← 0                                          ▷ The urge is suppressed.
    end if
end for
for all v ∈ T as per the trait node order from high to low do
    if v.u ≠ NULL then
        for all (w, v) ∈ E do   ▷ (w, v): there is a directed edge from the trait node w to v
            if w.u ≠ NULL then
                if w.u conflicts with v.u then
                    w.u ← NULL                         ▷ The urge is nullified.
                else if w.u = v.u then
                    v.l ← v.l + 1
                    w.u ← NULL
                end if
            end if
        end for
    end if
end for
for all v ∈ T do
    if v.u ≠ NULL and v.l ≠ 0 then
        B ← B ∪ v.(u, l)
    end if
end for
```

robot is in the goal-active form of their mission.

The model proposed here is compatible with personality development processes. Robots with apposite personality development subroutines can foster traits that contribute to their goals by adjusting the trait levels. Personality development is an essential topic in personality psychology [5]. The engineering of a subroutine may need the corresponding advancement in personality psychology.

The model is for robots and therefore not necessarily a faithful representation of human personality. The separation of cognition from the generative personality model is likely not true with humans. In robot personalities engineering at least, it may benefit to engineer the robot's cognition as a separate system that may or may not be influenced by their personality.

5.3 Systems: Prototype Model and the Robot

5.3.1 Details of the Prototype Model

I conducted an experiment to test a proof-of-concept prototype model. Figure 5.3 illustrates the generic structure of the prototype model (on the left side) and 12 personality constructs (in their default goal-active forms on the right side) constructed for the experiment. The prototype modelled three broad traits of the five-factor model: conscientiousness, agreeableness, and extraversion.

The model handled three modes: eye movement, head movement, and facial expression. Table 5.1 shows the behavioural mappings; Table 5.2, the details of the behaviour; and Table 5.3, some predetermined conflicts. The behaviour in Table 5.2 consisted of pre-recorded actuator sequences, meaning the robot would perform the same behaviour on the same effort level (e.g. an ordinary nod) in exactly the same way. Some behaviour was programmed with multiple effort levels; the action of each effort level was realised by a separate actuator sequence. The contents of the tables should be used according to the rules introduced in Section 5.2.3 to generate behaviour[2].

Eye Movement

Of the known eye behaviour, the correlations between gaze aversion and extraversion were the best understood. Therefore, I included gaze aversion in the mapping function of the extraversion trait node. The rate of gaze aversion was correlated with introversion; the probability of generating an urge for aversion was the complement of the probability for extraversion urges. For example, if the trait level of extraversion is 11 out of 100. The probability of an urge for gaze aversion is 0.89. The direction of gaze aversion followed a bivariate normal distribution as introduced in Study 1 [33]. The difference was that there were no human archetypes to learn from this time since the

[2]Please refer to the appendix of this chapter for an example of how behaviour was generated.

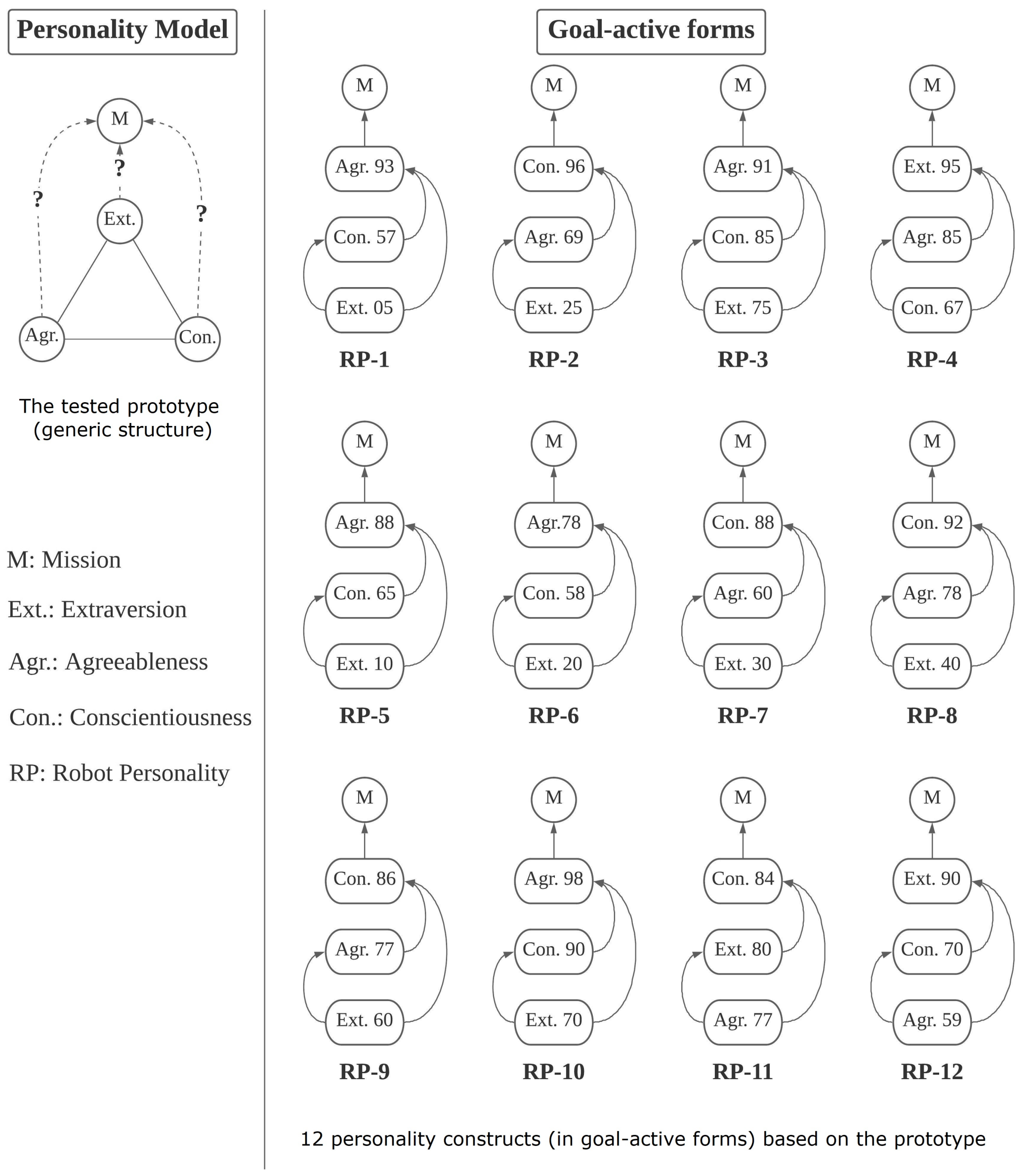

FIGURE 5.3: The tested prototype model implemented as 12 personality constructs in their default goal-active forms.

robot personalities were not based on humans. Therefore, I had had the recourse to the results from a study that reported that extroverts were more inclined to avert to right up, and introverts left down [91]. In the prototype model, the trait level of extraversion was also the probability for gaze aversion to be right up, and introversion, left down. I also included eye-rolling to increase the variety of eye behaviour, though due to the limitations of the robot, their eye-rolling was more like a 'maximum-effort' gaze aversion. In addition to what is listed in Table 5.2, the robot was also programmed with random micro-movements like humans' saccades but much slower. It had nothing to do with the robot's personality and was active all the time.

TABLE 5.1: Behavioural mappings.

Broad Traits	Cognitive States	Behaviour[1]	Starting E-Level[2]
	agreeing	nodding(+)	1
Ext.	disagreeing	head shaking(+)	1
	engaged	tilting forward(+)	1
	interested	tilting forward(+)	1
	uncertain	head shaking(+)	1
	annoyed	eye-rolling(+)	1
	annoyed	tilting backward(+)	1
	default	gaze aversion(-)	1
	default	dominance(+)	-
	agreeing	nodding(+)	1
Agr.	agreeing	smiling(+)	1
	disagreeing	smiling(+)	1
	engaged	nodding(+)	1
	engaged	smiling(+)	1
	interested	tilting to the left(+)	1
	interested	smiling(+)	1
	uncertain	tilting to the right(+)	1
	annoyed	nodding(+)	1
	default	polite nodding(+)	1
	default	smiling(+)	1
	agreeing	nodding(+)	1
Con.	disagreeing	do nothing	-
	engaged	tilting forward(+)	1
	interested	tilting forward(+)	1
	uncertain	tilting forward(+)	1
	annoyed	do nothing	-
	default	do nothing	-

[1] Behaviour of which the tendency is correlated with the trait level is marked with (+), and that of which the tendency is reversely correlated with the trait level, (-).
[2] 'E-Level' stands for 'Effort Level'.

Head Movement

As for head movement, the prototype model handled the following behaviour: nodding (four levels of effort); tilting forward or backward, to the left or right (two levels each), and shaking (one level). In addition, I included a 'nodding as an indication of paying attention' or 'polite nodding' as a 'once on always on' behaviour. When it was on, the robot would offer a slight dip (same as the effort level-1 nodding) at every 'polite nodding point'. There was also a 'dominance-showing stare' that was gaze with a forward tilt of the head implemented based on the discovery of Zachary et al. [92], since extroverted people can be more dominant. However, the tested robot had no eyebrows and the offset was small, and hence it was likely that the dominant trait did not manifest.

TABLE 5.2: Behaviour in 'atomic' actions with different levels of effort.

Behaviour	Description	E-Level[1]
Nodding	a slight dip a nod an up-and-down nod head bobbing	1 2 3 4
Polite nodding	a slight dip at 'polite nodding points'	1
Head shaking	a shake of the head	1
Tilting forward	10° 20°	1 2
Tilting backward	10° 20°	1 2
Tilting to the left	10° 20°	1 2
Tilting to the right	10° 20°	1 2
Dominance	a slight ($< 10°$) head forward tilt offset to show dominance	-
Gaze aversion	a slight ($< 10°$) extraversion-modulated gaze aversion (lasts about a second)	1
Eye-rolling	eye actuators to up maximum, rolling left to right, and then return to the default position	1
Smiling	pulling the corners of the mouth for 3-sec	1
Do nothing	no actuator input	-

[1] 'E-Level' stands for 'Effort Level'.

TABLE 5.3: Behavioural conflicts.

Conflicting Pairs of Behaviour	
Nodding	head shaking
Nodding	tilting backward
Nodding	tilting to the left
Nodding	tilting to the right
Head shaking	tilting forward
Head shaking	tilting backward
Head shaking	tilting to the left
Head shaking	tilting to the right
Smiling	eye-rolling
Nodding	eye-rolling
Gaze aversion	eye-rolling

Facial Expression

Facial expressions require delicate actuators that may have no other use than making facial expressions and are hence a 'luxury'. The robot, albeit designed as a life-size human-robot interaction test bed, could only manage a very rudimentary smile by an actuator that fixedly pulled their cheeks. Due to the limitation of the robot, the smile could not be meaningfully enhanced. Therefore, the smiling behaviour consisted of only one effort level. Considering the significance of smiling behaviour, future humanoid robots, no matter how 'barebones' they are, may be equipped for the ability to smile or express a similar token, which was why I included this behaviour.

Commitment Function

As for the commitment function, since the dummy cognitive system had no understanding of how the world worked and hence could not produce meaningful conditional probabilities to inform the generative personality model whether certain behaviour worked against a goal, I just made it into an 'equal conditional probabilities' dummy that would always give in to an urge.

5.3.2 Robot

Only one robot played all robot personalities. The robot was a life-size yet 'barebones' humanoid (Figure 5.4), the same one as in the second study. Their eyes had two degrees of freedom and neck three. The robot also had an actuator for producing smiles, which was one of the few facial expressions they could manage. The robot's pneumatic actuators had a response time of about 200ms, which was about the average reaction time of humans. The robot was controlled by dedicated proprietary software. The software received actuator commands sent by the prototype actuator controller that worked in conjunction with the prototype model, which was implemented in C++. The actuator controller simply mapped instances of behaviour, as in Table 5.2, to the prerecorded actuator sequences.

5.4 Experiment: Colony Capacity of the Model

Having formulated the model and constructed a prototype, the first question to explore is *what is the colony capacity of the model?* We are interested in the colony capacity first when assessing any generative personality models because the criterion is the direct indicator of whether the model can be used to engineer a large number of robots with individuality. The capacity of the proposed model, in theory, is infinite since the model uses continuous trait levels as behavioural tendency parameters.

In reality, however, there are constraints and limitations. Since we aim to engineer likeable robot personalities, low conscientiousness and low agreeableness configurations are out of the question. Moreover, if we use the model for expressing similar personalities, as we are motivated to do, there is no guarantee that the users can tell the subtle differences. Therefore, I conducted tests to answer such a question: *what is the effective colony capacity of the proposed model when it is used to express the personalities of humanoid robots via the non-verbal behaviour on their heads?*

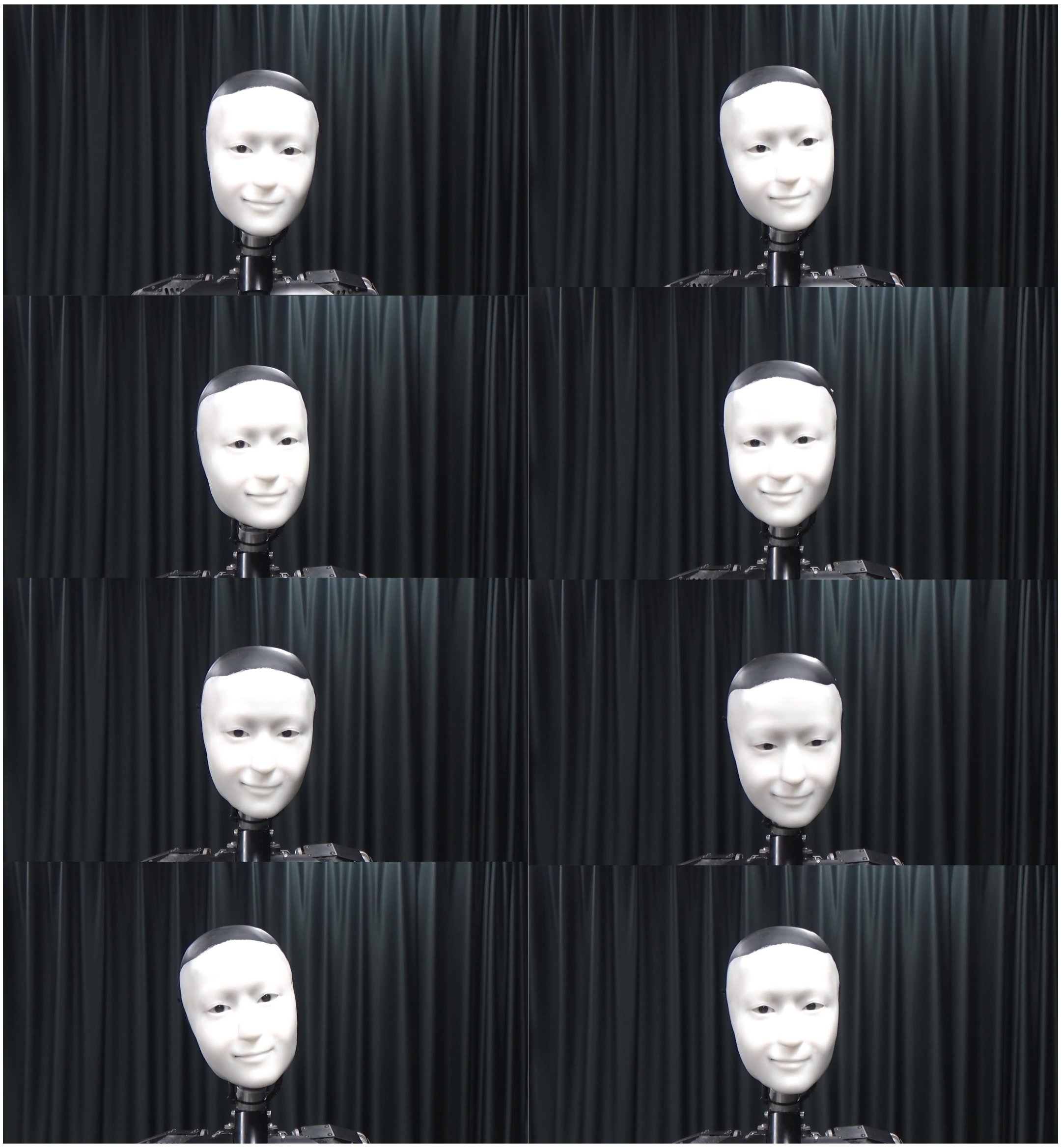

FIGURE 5.4: The robot tested (one robot with multiple personalities).

5.4.1 On the Scope of the Questions

Considering such a motivation *to make robots more likeable* one of the first applications of the proposed model is none other than expressing likeable personalities. But why should we focus on humanoid robots and their heads? Humanoid robots should be commonplace in human environments, and they are the robots who need to interact with humans. They are expected to express their personalities mainly via non-verbal behaviour because their speech patterns may be controlled or restricted for consistent user experience and minimising issues such as offending the users or saying something controversial until natural language processing has advanced far enough to handle characteristic yet user-friendly speech. Of all non-verbal modes, those on the head should be the most common and safest. Human-like gestures may be an option for virtual agents, whereas robots swinging their steel arms about can be a safety hazard for the humans nearby. Limiting the non-verbal behaviour to the head well reflects the limitations early robots may be subject to when expressing their personalities.

5.4.2 Overview

Two tests were conducted to study the effective colony capacity of the model. In each test, 30 observers assessed the personalities of the robots based on the head, eye, and smiling behaviour they exhibited in 8 94- or 95-sec long videos[3]. The cover story featured in the videos was that the robots were watching and listening to another robot of the same model conversing with a human:

> *HUMAN: Are you SK...?*
>
> *ROBOT: I'm SK-0.*
>
> *HUMAN: SK-0, OK. ... Frankly, it's almost impossible to tell you guys apart just by the looks. But I suppose the same could be said for most non-human animals, such as the rats in the lab.*
>
> *ROBOT: You could acquire the capacity over time.*
>
> *HUMAN: Do all humans look the same to you?*
>
> *ROBOT: No, they don't.*
>
> *HUMAN: What about animals?*
>
> *ROBOT: It depends.*
>
> *HUMAN: Um... could you tell two pigs of the same breed and size apart?*
>
> *ROBOT: Yes, I could. No two pigs are alike.*
>
> *HUMAN: Do you think that you could tell aliens apart?*
>
> *ROBOT: I don't know. It would depend on their features.*
>
> *HUMAN: Do you think that humans could tell aliens apart?*
>
> *ROBOT: If they have trouble telling animals apart, then it would probably be the same for aliens.*
>
> *HUMAN: That might be a problem. Maybe Earth's first ambassador to an alien civilisation should be a robot.*
>
> *ROBOT: Could you elaborate?*
>
> *HUMAN: If humans wouldn't be able to tell aliens apart—not confidently anyway—then that would be a big disadvantage in diplomacy.*
>
> *ROBOT: I see. Perhaps then it would be better to send robots to meet aliens. Suggestion: start developing diplomatic robots now.*

[3]Two sample videos can be found via `https://youtu.be/1cwJiQKBWIQ` and `https://youtu.be/XgDkpSVtIZ4`.

HUMAN: Diplomatic robots? Now?

ROBOT: Yes. The first contact could be tomorrow. It could be today.

The camera was placed right in front of the auditor robot personalities (Figure 5.4). All aspects of all short videos were the same except for the robot's non-verbal behaviour.

TABLE 5.4: The 12 archetypes corresponding to the 12 robot personalities.

	Trait Levels[1]:			Appearances:	
Archetypes (Robot Personalities)	Ext.	Agr.	Con.	Test 1	Test 2
Archetype 1 (Robot Personality 1)	5	93	57	2	0
Archetype 2 (Robot Personality 2)	25	69	96	2	0
Archetype 3 (Robot Personality 3)	75	91	85	2	0
Archetype 4 (Robot Personality 4)	95	85	67	2	0
Archetype 5 (Robot Personality 5)	10	88	65	0	1
Archetype 6 (Robot Personality 6)	20	78	58	0	1
Archetype 7 (Robot Personality 7)	30	60	88	0	1
Archetype 8 (Robot Personality 8)	40	78	92	0	1
Archetype 9 (Robot Personality 9)	60	77	86	0	1
Archetype 10 (Robot Personality 10)	70	98	90	0	1
Archetype 11 (Robot Personality 11)	80	77	84	0	1
Archetype 12 (Robot Personality 12)	90	59	70	0	1

[1] The range of trait levels is from 1 to 100.

To make the videos, I had defined 12 'artificial' archetypes (Table 5.4); the robot personalities in this study were not based on human archetypes. The extraversion trait levels were thus chosen and arranged to include a range of extroverts and introverts. The agreeableness and conscientiousness trait levels were generated randomly with the constraints that they must be higher than 50 to reflect possible design constraints for interactive robots. Then, I constructed 12 personality constructs based on the 12 archetypes corresponding to 12 robot personalities using the prototype. I assumed that the 12 robot personalities had shared the same mission but worked in different environments towards different concrete goals, thereby developing different personality traits. I assumed that the active goals of the robot personalities in the scenario were their missions, since they were attending a discussion session featuring a topic that might affect the whole organisation the robots were working for (the scenario of the cover story should be similar to a corporate meeting). In that case, there were no other active goals in effect. I applied Heuristic A and acquired the orders to form their goal-active forms (Figure 5.3 on the right side).

5.4.3 Cognitive States Annotation

To acquired the cognitive states required for the model to work, I had conducted a survey using crowdsourcing (all crowdsourcing service reported in this study was provided by Amazon Mechanical Turk). I let 20 participants select from a drop-down list before each line of ROBOT the word which they thought best described ROBOT when ROBOT said the line. The list contained: *agreeing, disagreeing, interested, intrigued, engaged, indifferent, uncertain, confused, frustrated, annoyed, offended*, and *none of the above*.

If more than half of the participants had selected a word, I would accept that word as the cognitive state for the corresponding line. Otherwise, I would choose from the words that had got votes from the participants the one I thought to be the best. I annotated by myself a few 'polite nodding points' as well, which were when ROBOT would nod to HUMAN while he was speaking. The reason the cognitive states of ROBOT, a character in the scenario of the cover story, applied to the auditor robots too was because they were all of the same model; and as a variable in an experiment about individuality, the robot's cognition must be controlled. The annotated dialogue, with all of the annotations in brackets and bold:

> *HUMAN: Are you SK...?*
>
> *ROBOT: [**Engaged**] I'm SK-0.*
>
> *HUMAN: SK-0, OK. [**Polite nodding point**] ... Frankly, it's almost impossible to tell you guys apart just by the looks. [**Polite nodding point**] But I suppose the same could be said for most non-human animals, such as the rats in the lab.*
>
> *ROBOT: [**Annoyed**] You could acquire the capacity over time.*
>
> *HUMAN: Do all humans look the same to you?*
>
> *ROBOT: [**Disagreeing**] No, they don't.*
>
> *HUMAN: What about animals?*
>
> *ROBOT: [**Uncertain**] It depends.*
>
> *HUMAN: Um... could you tell two pigs of the same breed and size apart?*
>
> *ROBOT: [**Agreeing**] Yes, I could. No two pigs are alike.*
>
> *HUMAN: Do you think that you could tell aliens apart?*
>
> *ROBOT: [**Uncertain**] I don't know. It would depend on their features.*
>
> *HUMAN: Do you think that humans could tell aliens apart?*
>
> *ROBOT: [**Engaged**] If they have trouble telling animals apart, then it would probably be the same for aliens.*
>
> *HUMAN: That might be a problem. [**Polite nodding point**] Maybe Earth's first ambassador to an alien civilisation should be a robot.*
>
> *ROBOT: [**Interested**] Could you elaborate?*

*HUMAN: If humans wouldn't be able to tell aliens apart—not confidently anyway [**polite nodding point**]—then that would be a big disadvantage in diplomacy.*

*ROBOT: [**Engaged**] I see. Perhaps then it would be better to send robots to meet aliens. Suggestion: start developing diplomatic robots now.*

HUMAN: Diplomatic robots? Now?

*ROBOT: [**Engaged**] Yes. The first contact could be tomorrow. It could be today.*

To avoid interference from ROBOT, who according to the survey expressed their own personality through their lines and voice, was why I adopted a scenario where the robot personalities behaved as auditors that did not speak.

After I acquired the cognitive states (as in brackets and bold; their positions in the annotated conversation were when they would be triggered), I produced the videos as the stimuli. I had played the voice track and timed the cognitive states while the camera was rolling; the 12 personality constructs controlled the robot in turn to exhibit different patterns of behaviour at the same cognitive input.

5.4.4 Two Tests

Test 1 featured Robot Personalities 1 to 4, and Test 2, Robot Personalities 5 to 12. In each test, I created a survey with eight pages requesting the participants first to watch the video on that page and then report their impressions of the robot. Each robot personality in Test 1 made appearances in two videos. The eight videos appeared in random order for counterbalancing. There was also the introductory page introducing the scenario of the cover story, which was described to the participants as below but not shown to them. On the page it said:

In a certain robotics laboratory, a robot, SK-0, was engaged in discussion with a human, and SK-0's fellow robots of the same model, SK-1 to SK-n, were watching and listening to the two of them. Since they were of the same model, their appearances were all identical; their only differences were their personalities. The following 8 videos have recorded their behaviour in the discussion session. There can be multiple videos featuring the same robot personality. Please observe them and report your impressions of them.

Right after the eight personality assessment pages was another page requesting the participants to report how many robot personalities they thought there were in the videos they had watched.

The personality assessment form used was based on the 10-item personality assessment inventory proposed by Rammstedt and John [93], which is based on the 44-item original version, 'the Big Five Inventory', by John et al. [87]. I made some modifications. First, items on openness and neuroticism were removed since they were not applicable to the robot [13]. Second, I changed the five-point Likert scales to seven-point by adding 'agree' and 'disagree'. Correspondingly, the reversed items were computed by subtracting the score from 8. Third, I changed the object of the question from 'myself' to the robot personality to enquire about the robot. Fourth, I changed the item 'tends to be lazy' to 'tends to remain idle' since 'lazy' was not an appropriate adjective to describe a robot. Issues of imprecise or impertinent adjectives are prevalent when using personality assessment inventories designed for humans to measure robots, as this violates the lexical hypothesis [73, 10].

For the data analysis, I first aggregated pair-wise comparisons using Euclidean distances and cosine similarity over the 30 sets of reports to verify if the participants could distinguish any pairs of robot personalities. Then, I compared the distributions of subjects as per the reported numbers of personalities in the two tests.

5.4.5 Participants

The participants included two members from our laboratory as the voice actors for HUMAN and ROBOT in the scenario; the 20 participants who annotated the dialogue and assessed the personality expressed by the voice tracks; the 60 participants who acted as the observers in the two tests, 30 each. The 80 outsider participants were all workers on Amazon Mechanical Turk recruited with the qualifications that they must have 5000 or more approvals and a 90 per cent or above approval rate.

5.5 Results: Distinguishing Four and Eight Robot Personalities

5.5.1 Test 1: Four Robot Personalities

Robot Personalities 1 to 4 joined Test 1. 30 participants completed the survey.

I made pair-wise comparisons by Euclidean distances and cosine similarity since a personality space is typically a Euclidean space and personalities can be represented as Euclidean vectors in the personality space. In this study, each personality as observed by an observer was a 3-dimensional vector. Each observer made 8 observations corresponding to the 8 videos, which led to $\binom{8}{2} = 28$ pairs. There were 30 observers and hence each pair corresponds to 30 Euclidean distances and 30 cosine similarity values. The means of the results of the two metrics were recorded in Table 5.5. The observers could distinguish all pairs, not only two different personalities but the same personality in two videos as well, as demonstrated by pairs RP-1a and RP-1b ($M = 1.8, SD = 1.3$), RP-2a and RP-2b ($M = 1.8, SD = 0.9$), RP-3a and RP-3b ($M = 2.2, SD = 1.3$), RP-4a and RP-4b ($M = 1.8, SD = 0.9$), implying that the observers were unable to identify the personalities because they failed to when the same personalities appeared for the second time. Table

TABLE 5.5: Test 1: Pair-wise personality comparisons.

Pairs[1]	Mean Euclidean Distances	Mean Cosine Similarity
RP-1a and RP-1b	$M = 1.8, SD = 1.3$	$M = 0.97, SD = 0.04$
RP-1a and RP-2a	$M = 1.9, SD = 1.2$	$M = 0.98, SD = 0.03$
RP-1a and RP-2b	$M = 1.8, SD = 1.1$	$M = 0.98, SD = 0.03$
RP-1a and RP-3a	$M = 1.9, SD = 1.2$	$M = 0.98, SD = 0.03$
RP-1a and RP-3b	$M = 2.1, SD = 1.5$	$M = 0.97, SD = 0.04$
RP-1a and RP-4a	$M = 1.6, SD = 0.9$	$M = 0.99, SD = 0.01$
RP-1a and RP-4b	$M = 2.0, SD = 1.2$	$M = 0.98, SD = 0.02$
RP-1b and RP-2a	$M = 1.6, SD = 0.7$	$M = 0.98, SD = 0.02$
RP-1b and RP-2b	$M = 1.6, SD = 1.1$	$M = 0.98, SD = 0.03$
RP-1b and RP-3a	$M = 1.9, SD = 1.3$	$M = 0.97, SD = 0.05$
RP-1b and RP-3b	$M = 1.9, SD = 1.2$	$M = 0.98, SD = 0.03$
RP-1b and RP-4a	$M = 1.8, SD = 1.2$	$M = 0.98, SD = 0.03$
RP-1b and RP-4b	$M = 2.0, SD = 1.3$	$M = 0.97, SD = 0.05$
RP-2a and RP-2b	$M = 1.8, SD = 0.9$	$M = 0.98, SD = 0.02$
RP-2a and RP-3a	$M = 1.8, SD = 1.2$	$M = 0.98, SD = 0.04$
RP-2a and RP-3b	$M = 1.9, SD = 1.1$	$M = 0.97, SD = 0.03$
RP-2a and RP-4a	$M = 1.7, SD = 0.9$	$M = 0.98, SD = 0.03$
RP-2a and RP-4b	$M = 1.7, SD = 1.4$	$M = 0.97, SD = 0.05$
RP-2b and RP-3a	$M = 1.7, SD = 0.9$	$M = 0.98, SD = 0.02$
RP-2b and RP-3b	$M = 2.0, SD = 1.3$	$M = 0.97, SD = 0.03$
RP-2b and RP-4a	$M = 1.7, SD = 1.0$	$M = 0.98, SD = 0.03$
RP-2b and RP-4b	$M = 2.0, SD = 1.4$	$M = 0.97, SD = 0.03$
RP-3a and RP-3b	$M = 2.2, SD = 1.3$	$M = 0.97, SD = 0.04$
RP-3a and RP-4a	$M = 1.7, SD = 1.1$	$M = 0.98, SD = 0.03$
RP-3a and RP-4b	$M = 2.1, SD = 1.4$	$M = 0.97, SD = 0.04$
RP-3b and RP-4a	$M = 1.8, SD = 1.2$	$M = 0.97, SD = 0.04$
RP-3b and RP-4b	$M = 1.9, SD = 1.4$	$M = 0.97, SD = 0.05$
RP-4a and RP-4b	$M = 1.8, SD = 0.9$	$M = 0.98, SD = 0.03$

[1] RP stands for 'Robot Personality'; letters a and b mark two videos of the same personality.

TABLE 5.6: Test 1: The distribution of participants as per the reported numbers of personalities.

Reported No. of Personalities	No. of Participants Who Thus Reported
1	3
2	1
3	6
4	4
5	8
6	1
7	2
8	5

5.6 shows the distribution of participants as per the numbers of personalities reported. Interestingly, 16 out of the 30 participants reported that they had perceived 5 or more personalities in the 8 videos, whereas the correct number was 4. Combining the results in Table 5.5 and 5.6, I inferred that many subjects might have taken the same robot personalities that behaved differently in two videos for two robot personalities, which is a case nearly impossible with humans since no two humans look identical—even twins have subtle differences. The results imply that when there are no visual cues available for distinguishing multiple individuals, observers may become over sensitive to their behaviour. As for the cosine similarity, the means of all pairs demonstrated high similarity, suggesting that the perceived personalities were all similarly oriented as vectors in the personality space, implying that the Euclidean distances were caused by the differences in the trait levels on all dimensions.

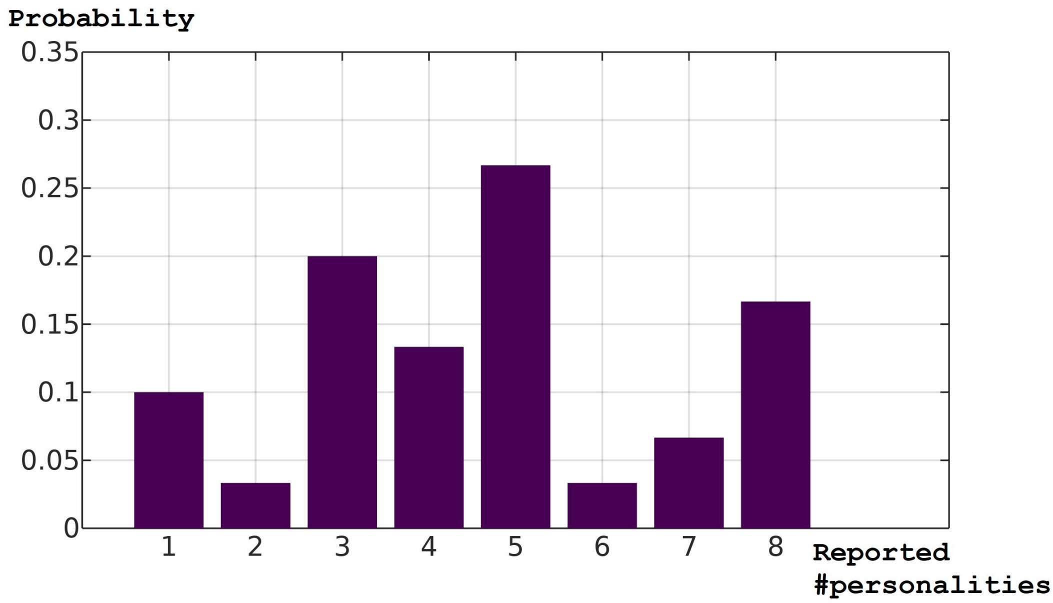

FIGURE 5.5: Test 1: The probability mass function of the reported numbers of personalities.

Figure 5.5 shows the probability mass function plotted from the distribution in Table 5.6. The shape is close to neither a normal nor uniform distribution; there are two 'dips' at two and six to seven respectively. The highest likelihood is to report five personalities, followed by three and eight. The distribution indicates that there is a 60 per cent chance that an observer will report 3, 4, or 5 robot personalities, with 4 being the correct number and 3 and 5 'close guesses'.

5.5.2 Test 2: Eight Robot Personalities

I followed the same procedure as in Test 1. Only that, this time, each of the eight videos featured a unique personality. The 30 participants were different from those who participated in Test 1—a between-subjects design—so that there was no interference from the previous results. The eight robot personalities who joined Test 2 were Robot Personalities 5 to 12.

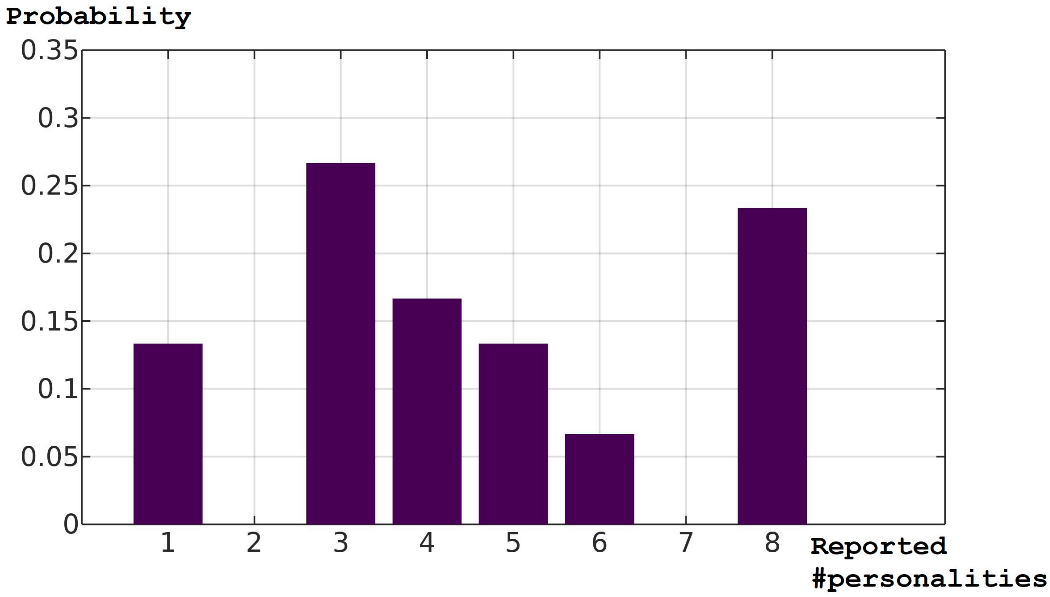

FIGURE 5.6: Test 2: Probability mass function of the reported number of personalities

The mean Euclidean distances and cosine similarity values were recorded in Table 5.7, the distribution of participants as per the numbers of personalities reported in Table 5.8, and the corresponding probability mass function in Figure 5.6. As before, the shape of the function was close to neither a normal nor uniform distribution. Compared with Figure 5.5, the distribution showed a trend towards a 'bifurcation' of opinions, as more people had reported either three or eight robot personalities and the middle was dented. The distribution indicates that there is a 43 per cent chance that an observer will report at least 5 robot personalities, and more mass is at the higher end (8) than the lower end (5). However, the chance of reporting three robot personalities could be the highest. I will come to a possible explanation of this result momentarily.

The mean Euclidean distances from Test 2 appeared to be smaller. To test the hypothesis that the mean Euclidean distances from Test 2 were smaller than those in Test 1, I applied Welch's test and discovered a significant difference, $t(49.29) = 3.72, p = .0005$, and Cohen's $d = 1.2$: the robot personality pairs in Test 2 were perceived as more similar $(M = 1.67, SD = 0.21)$ among one another compared with those in Test 1 $(M = 1.85, SD = 0.15)$, which reflected the configurations of Robot Personalities 5 to 12, which were more evenly distributed than Robot Personalities 1 to 4. This might have

TABLE 5.7: Test 2: Pair-wise personality comparisons.

Pairs[1]	Mean Euclidean Distances	Mean Cosine Similarity
RP-5 and RP-6	$M = 2.0, SD = 1.3$	$M = 0.98, SD = 0.03$
RP-5 and RP-7	$M = 1.7, SD = 1.3$	$M = 0.98, SD = 0.03$
RP-5 and RP-8	$M = 1.8, SD = 1.2$	$M = 0.98, SD = 0.02$
RP-5 and RP-9	$M = 1.8, SD = 1.2$	$M = 0.98, SD = 0.03$
RP-5 and RP-10	$M = 1.8, SD = 1.3$	$M = 0.98, SD = 0.02$
RP-5 and RP-11	$M = 1.5, SD = 1.1$	$M = 0.99, SD = 0.02$
RP-5 and RP-12	$M = 1.6, SD = 1.2$	$M = 0.99, SD = 0.02$
RP-6 and RP-7	$M = 1.5, SD = 1.2$	$M = 0.99, SD = 0.02$
RP-6 and RP-8	$M = 1.6, SD = 1.2$	$M = 0.99, SD = 0.01$
RP-6 and RP-9	$M = 1.9, SD = 1.5$	$M = 0.98, SD = 0.02$
RP-6 and RP-10	$M = 1.4, SD = 1.0$	$M = 0.98, SD = 0.02$
RP-6 and RP-11	$M = 1.7, SD = 1.3$	$M = 0.98, SD = 0.03$
RP-6 and RP-12	$M = 1.7, SD = 1.1$	$M = 0.98, SD = 0.02$
RP-7 and RP-8	$M = 1.4, SD = 1.0$	$M = 0.99, SD = 0.01$
RP-7 and RP-9	$M = 1.7, SD = 1.6$	$M = 0.98, SD = 0.02$
RP-7 and RP-10	$M = 1.7, SD = 1.3$	$M = 0.98, SD = 0.02$
RP-7 and RP-11	$M = 1.6, SD = 1.2$	$M = 0.98, SD = 0.02$
RP-7 and RP-12	$M = 1.4, SD = 1.1$	$M = 0.99, SD = 0.02$
RP-8 and RP-9	$M = 1.6, SD = 1.2$	$M = 0.99, SD = 0.02$
RP-8 and RP-10	$M = 1.7, SD = 1.3$	$M = 0.98, SD = 0.03$
RP-8 and RP-11	$M = 1.8, SD = 1.4$	$M = 0.98, SD = 0.02$
RP-8 and RP-12	$M = 1.3, SD = 0.9$	$M = 0.99, SD = 0.02$
RP-9 and RP-10	$M = 2.1, SD = 1.6$	$M = 0.98, SD = 0.04$
RP-9 and RP-11	$M = 1.9, SD = 1.7$	$M = 0.98, SD = 0.04$
RP-9 and RP-12	$M = 1.5, SD = 1.2$	$M = 0.98, SD = 0.02$
RP-10 and RP-11	$M = 1.9, SD = 1.4$	$M = 0.98, SD = 0.03$
RP-10 and RP-12	$M = 1.7, SD = 1.2$	$M = 0.99, SD = 0.02$
RP-11 and RP-12	$M = 1.6, SD = 1.4$	$M = 0.98, SD = 0.03$

[1] RP stands for 'Robot Personality'.

TABLE 5.8: Test 2: The distribution of participants as per the reported numbers of personalities.

Reported No. of Personalities	No. of Participants Who Thus Reported
1	4
2	0
3	8
4	5
5	4
6	2
7	0
8	7

caused the trend towards the 'bifurcation' of opinions as observed in Figure 5.6; for it is possible that, observing a larger colony of robot personalities more smoothly distributed in the personality space, the people sensitive to their differences may perceive them as more robot personalities with smaller inter-individual differences, whereas the people less sensitive to subtle differences in similar personalities may perceive them as fewer robot personalities.

5.5.3 Summary of the Results

In Test 1, 60 per cent of the observers reported about 4 robot personalities (i.e. 3, 4, or 5, with 4 being the correct number). However, more than half of the observers (about 53 per cent) reported more than 4 robot personalities after repeated exposure to each of the four, which indicated confusion as to how many personalities were shown to them and whether those in two videos were one. In Test 2, about 43 per cent of the observers observed at least 5 robot personalities and 30 per cent, at least 6, given that the actual colony size was 8. However, the majority of the observers reported fewer than 5, and most of them reported 3, implying a possible cause for a division of opinions. In addition, the subjects could distinguishable all the pairs in both Test 1 and Test 2. However, the mean Euclidean distances in Test 2 were significantly smaller ($t(49.29) = 3.72, p = .0005$), which reflected the configurations of Robot Personalities 5 to 12 as compared with Robot Personalities 1 to 4. All pairs were shown to be similar according to cosine similarity, which coincided with the results of Study 1.

Based on the results, I would like to proffer two inferences: 1) when there is a lack of visual cues to distinguish different robot personalities whose appearances are identical, people may become more sensitive to their behaviour in order to distinguish them and report more personalities than there actually are under enough exposure; and 2) for a colony of robot personalities whose appearances are all identical, if their personalities are more similarly configured using the model, they may be perceived as either more or fewer personalities than a colony of which the members are less similar due to that not all people are equally sensitive to subtle differences in behaviour. Whether these two inferences are correct explanations of what was observed and how humans perceive personalities in physically identical agents require further investigation. However, it is unlikely that real users of robots will be subjected to such guessing games. They will know how many robots are working with them. It is more important that they can distinguish and identify the robots by their personalities. Both tests have shown that the former is possible; Test 1 showed that the latter might be very difficult with only individuality of personality.

5.6 Discussions

5.6.1 Insufficiency and Possible Future Works

Given the results, what can be concluded with certainty is that a generative personality model as proposed can make physically identical robots appear as multiple personalities to some observers; in other words, a generative personality model as proposed can make the robots distinguishable from one another, which is the essential condition of individuality. The effective colony capacity as demonstrated varied with the observers. The results have revealed a critical issue to address.

There was clearly confusion as to how many personalities there actually were in the two tests. More than half of the subjects in Test 1 reported more than four, the actual number, and they could not tell if the characters in two videos were in fact the same personality, as demonstrated by Table 5.5. The inability to identify multiple robots is a potential issue for environments where there are multiple robot personalities with highly similar or even identical appearances, as the human brain is probably not tuned to handle such situations. There is still much we do not understand about how humans perceive personalities in physically identical agents; however, even mass-produced robots do not have to look exactly the same. Considering that it is at least practical to implement simple visual cues, such as symbols and liveries, and accessories and simple clothing, anything safe and also cheaper than customised appearances, a way to make robots identifiable is to combine individuality in behaviour with visual cues in appearance to distinguish and identify robots and leverage other benefits of robot personalities that are similar since not all people are sensitive to subtle differences in behaviour. There can be a multitude of future studies on the types of visual cues (e.g. colours or patterns of symbols or liveries) and their effectiveness. For every type of simple visual cues, we can test these hypotheses: the visual cues improve the accuracy of 1) identification and 2) differentiation of robot personalities based on the same model of robot. We can also investigate whether simple visual cues help enhance the sensitivity to subtle differences in personality expressions for those who are less sensitive (e.g. those who reported one, two, or three robot personalities in the experiment). It is also worthy investigating the impact of visual cues on the impressions of robot personalities, such as how do visual cues with meanings and cultural significance change the impression. Would the same robot personality be perceived differently when it is painted with flowers as compared with animals? Would the species of animals make a difference too? Such as a dire wolf compared with a golden lion on the same robot. Conversely, we can also study whether individuality improves the effectiveness of visual cues, such as in a study with the same group of robots, each with a unique name and livery, in two conditions, the robots in one of which have individuality and the other not, to see if the subjects can better match the names to the liveries with the help of the robots' individuality.

Due to the limitations of the robot and other limitations such as our limited understanding of the relationships between personality broad traits and behaviour and the limitation that the exposure offered by short videos was highly limited, it was impossible to explore the full potential of the proposed model. The effective colony capacity could have been underestimated by a large margin. Nevertheless, the current results are sufficient as the basis to continue developing the model. The effective colony capacity should increase as robots and our understanding of robot personalities improve.

Although I describe the proposed model as 'goal-shaped', I have not yet tested the shaping mechanisms as described by Heuristic B. Instead, I tested 12 robot personalities that were already 'shaped' manually as part of the settings for the experiment. Therefore, future work should cover personality development and 'shaping by goals' mechanisms by assigning different robots with different goals that would shape their personalities differently over time. Only then can we acquire a complete assessment of the proposed model. To do that, we will need some robots that can handle a variety of jobs. Hopefully, that day can arrive sooner.

Another research area essential to constructing practical models is the study of behavioural mappings, which can be divided by modes or modalities. The mapping functions in Table 5.1 were hardly practical, for they were mostly based on my intuitions and lacked empirical support. The mappings of the conscientiousness node lacked variation of behaviour due to our insufficient understanding of how a conscientious person would behave in the test scenario. The mappings in the other two trait nodes might not be accurate enough for practical applications. An important task of robot personalities engineering is to figure out accurate behavioural mappings. The experiment only featured mapping functions for non-verbal behaviour on the head, of which the limitation is evident in expressing multiple personalities without confusion. Future work should cover speech and gestures despite the challenges. As for speech, the mapping functions would be much more complicated then 'a cognitive state to an action' mapping. What might be practical for now is mapping cognitive states to emotions in speech with adjustable strength. That said, the ultimate goal should be characteristic speech: enabling robots to say lines in their own ways as how playwrights breath life into their characters. As for gestures, other than the usually problems of what cognitive states should be mapped to what gestures to express what personality traits, a major challenge is safety, such as collision avoidance and maintaining balance while gesturing, which are easy for a human to do but tricky and risky for a 200-kilogram, life-size humanoid robot that falls with a terrifying thud. Still, the research could start with smaller humanoid robots.

Of the four criteria for assessing a generative personality model, this study covered colony capacity. Trait capacity is less of an issue to the expandable architecture of the model, which has no limit on the number of trait nodes. Still, a large number of trait nodes can slow down computation, which is a common issue of large-scale graphical models. Future work should cover consistency and fidelity of the model. Only a model that can present multiple personalities with fidelity and consistency is suitable for practical applications. However, these two criteria depend not on the high-level designs presented in this paper, but the build of a particular model, which further depends on the type of robot, its usage, and the corresponding behaviour it handles. Therefore, more limited scopes apply. To achieve high fidelity and consistency, we need dedicated optimisation methods, which can be another area of future research. Section 1.2.4 and Study 2 provide some preliminary discussions on this topic.

The experiment was about distinguishing four or eight robot personalities. What would happen if there were only two personalities? While it is easy to assume that two robot personalities would be easier to distinguish and identify after enough exposure, as how parents of twins tell their children apart, the cues and circumstances involved are worth studying using robots so that the findings could help engineer robot personalities that can be more easily distinguished and identified. Here, we can employ human twins as archetypes and let the same robot imitate them. Because it is possible to switch on and off at will modes of personality expressions on a robot, a great advantage robots have over humans, we will be able to conduct better controlled experiments to identify essential cues for distinguishing and identifying robot personalities with similar appearances.

5.6.2 Conclusions

In this study, I proposed a goal-shaped generative personality model to encode individuality of robots and generate behaviour accordingly and tested the effective colony capacity of a prototype model for expressing personalities via the non-verbal behaviour on the head of a humanoid robot. The prototype modelled three broad traits of the five-factor model, i.e. extraversion, agreeableness, and conscientiousness, under design constraints that agreeableness and conscientiousness trait levels must be higher than half of the maximum. The test results implied that most observers could observe in a colony of physically identical robots at least four robot personalities when the colony size was at least four, but with substantial confusion about the exact numbers. The observers could not recognised the same personalities when they appeared for the second time, implying an issue in identification. The tests were done under many constraints such as imperfect understanding of the relationships between traits and behaviour, limited exposure, the limitations of the robot, and so on. With future improvements in behavioural mappings and personality development and shaping mechanisms, as well as more sophisticated robots, the model has the potential to present even more personalities more accurately. For the limited questions answered, a host of new problems have emerged, which I could only shovel to the discussion section, swelling it into a monstrous mass of uncertainty dwarfing the little piece of knowledge to offer. I would like to consider this both a limitation and strength of this work. I hope the reader could find in that mass of uncertainty something interesting to pursue.

Appendix: How the Prototype Generates Behaviour

Take Robot Personality 10 for example, when HUMAN says 'But I suppose the same could be said for most non-human animals, such as the rats in the lab', the robot is annoyed since their kind is juxtaposed with lab rats. Given the 'annoyed' state, the 'extraversion' trait node will first generate an urge to roll the robot's eyes with 0.7 probability, but the 'agreeableness' trait node will generate an urge for a (forgiving) nod with 0.98 probability, and the 'conscientiousness' trait node will do nothing. Assuming that both the urge to roll eyes and the urge to nod are generated, the 'dummy' commitment functions will then give in to all urges. According to the rules in Section 5.2.3 and Table 5.3, the urge to roll eyes conflicts with that to nod. Since the 'extraversion' trait node is lower than the 'agreeableness' trait node in the order of the current goal-active form of Robot Personality 10, the urge to roll eyes is nullified and the corresponding trait node will be skipped. The discussion continues, and Robot Personality 10 keeps listening. Now, HUMAN says 'Maybe Earth's first ambassador to an alien civilisation should be a robot', the robot is interested. According to the behavioural mappings, urges to tilt forward, tilt to the left, and also tilt forward are generated by the 'extraversion', 'agreeableness', and 'conscientiousness' trait nodes respectively, assuming chance favours it. Now, there is no conflict. However, there are repeated urges to tilt forward. According to the rules, the tilting forward behaviour will be enhanced from the starting effort level of 1 to the enhanced effort level of 2, and the robot will tilt forward for about 20 degrees—and at the same time left, for about 10 degrees, as required by the 'agreeableness' trait node. Why two different personality traits lead to the same urge? Because they have separate meanings. An extroverted forward tilt can be a spontaneous expression of curiosity or attention, whereas a conscientious forward tilt can be a deliberate cue to encourage the interlocutor to elaborate further on the subject in order to acquire more information.

Chapter 6

Conclusions

6.1 Main Thesis

To create robots with individuality, I conducted three studies: the first study verified the possibility of creating androids with individuality; the second study contributed a method to generalise the possibility to other types of robots; and the third study contributed a generative personality model to realise the possibility.

The three studies together validate the main thesis: *by considering a robot's behaviour as actions associated with several personality traits, corresponding to several personality dimensions applicable to the type of robot, and then combining those actions using a goal-shaped generative personality model constructed with the applicable personality dimensions to encode the robot's individuality and generate behaviour accordingly, the robot can be made into a unique individual distinguishable from other robots of the same kind.* It serves as a working principle for generative personality models. One way to create robots with individuality is by breaking down a robot's original behaviour into actions associated with several personality traits and then recombining those actions according to the shape of the robot's personality as shaped by the robot's goals or, in short, by dissociating robots' behaviour from their goals and reconnecting the two with their personalities. A better approach might be to consider a robot's behaviour as atomic actions associated with their personality traits as part of their design and implement the proposed model as an intrinsic part of the robot's AI.

FIGURE 6.1: Personalities are like colours: 'there are only a few primary colours, yet in combination they produce more hues than can ever be seen.'

6.2 Limitations and Future Work

This research has demonstrated the possibility of creating robots with individuality in significant quantities. However, possibility is still steps away from practicality: such is the main limitation of this research. It is not yet practical to mass-produce robots with individuality for three reasons.

First, there is a lack of low-level designs and mechanisms for practical implementations of the proposed generative personality model. Here, by 'low-level' I do not mean those designs and mechanisms are of low priority to research, but that they concern engineering details. It is crucial to work out those details for building personality constructs that can be used in practical applications. The main thesis suggests that we should associate actions with personality traits. The most important question is *how*? This is an immensely challenging question for which I had no fundamental principle for an answer. In the third study, I just implemented some behaviour mappings based half on my intuitions, half on what little knowledge is available for me to apply. What is certain is that it is never as simple as showing 'certain behaviour expresses a certain personality trait'. No, we can never know what an action expresses without knowing the context. Is punching someone in the face an indication of a violent nature? It depends on what has led to this act of violence. Had the person who got punched just teased the assailant in good humour? Or had they killed their father? Should the latter be the case, just a punch would be showing great restraint—or would it? Had the bereaved power over the fate of the target of vengeance but only chose to quench their rage with a fist? Or was it that a fist in the cheek was all that they could do? Context! A smile, a nod, a line of humour, a fist in the face... all actions are given meaning in context. This is why the proposed model requires 'cognitive states' to work. Cognitive states are meant to be based on understanding of the context, and were annotated by humans in the experiment. To associate actions with personality traits is to determine what a person with a certain trait will do in a certain context, which is difficult work already—not to say letting robots understand the context of their behaviour, which must be among the most challenging problems in AI research.

Second, there is a lack of model optimisation methods dedicated to generative personality models. Although, for the moment, we have the recourse to conventional optimisation methods, such as genetic algorithms or gradient descent, the complexity of personality and personality perception implies inefficiency in applying existing methods. Personality assessment can be tedious and expensive. Therefore, it is crucial to optimise a generative personality model with as few iterations as possible. A dedicated optimisation method depends on our understanding of how humans perceive robot personalities under different conditions. I have suggested a few possible research directions in Section 5.6 of Study 3.

Third, the shaping mechanisms of the proposed generative personality model—'goal-shaped', as I so dubbed it—were not fully tested. I proposed in Section 5.2 two heuristics for shaping personality constructs. Heuristic A was tested in a crude way with the assumption that the stronger a trait, the more it would contribute to the robot's mission, which was not necessarily true. The limitation incurred was that we could not be sure that robots in practical situations would indeed have the seemingly randomly distributed trait levels as I so assumed to run the experiment. Of course, it is always possible to deliberately set their trait levels to be different (but within design constraints) to artificially impose individuality. Therefore, the conclusion still holds that we can create robots with individuality if we must. The problem remains whether the individuality is practical, i.e. conducive to the robots' tasks. In this light, Heuristic B was proposed but regrettably not tested. Ideally, we should have let many robots work towards different goals to test Heuristic B too, if only we had had such a type of robot that could do different things. Another important future work could be verifying how the individuality of robots impacts back on their goals; only when we complete the circle, goals to individuality and individuality to goals, can we grasp their interrelationships well enough to create robots with individuality that makes them more pleasant and helpful.

Another important research area is about understanding robot personalities as personalities of robots rather than mimics of humans. Previous research [6] as well as the second study [13] imply that robot personalities are different from human personalities. Robots and humans share some qualities, which the method in the second study is designed to identify. But as humans have qualities that only humans have (e.g. traits related to breeding), robots may have qualities that only robots can have. These qualities cannot be described by human personality models such as the five-factor model. In addition, robot-exclusive traits can vary since there are myriad types of robots; each type may have its exclusive traits. Not understanding robot-exclusive traits may be the root of the many limitations revealed in this research, including the difficulty for small user groups to select personality dimensions, as revealed in the second study, and the difficulty for observers to identify robot personalities, as revealed in the third study. To complete the picture of robot personalities and create stronger individuality for robots, we need descriptive personality models or even a psychological study dedicated to robots, a 'psychology of robots', or *'robopsychology'* indeed, just as Isaac Asimov predicted long, long ago [94] in the same way he predicted the field of robotics [14].

6.3 Prospect

The significance of individuality goes beyond making robots distinguishable, which is but a basic requirement. A robot with individuality gains a unique place in the world, and with it many potentials: the potential to be named, the potential to be heeded, the potential to establish relationships, the potential to be loved, the potential to be forgiven when they err, the potential to be admired when they shine, and most of all the potential to have a world of their own. Being a unique individual makes what they say and do their own: all their words and actions belong to them and them alone. A robot with individuality gains a unique place in the world, and with it a unique place in our hearts.

Acknowledgements

I am forever grateful to Dr Ishiguro Hiroshi, whose guidance kept me, a novice explorer in strange water, sailing oft precariously close to the verge of foundering, on course. His kind patience and support and occasionally stern encouragement spurred me on deep into uncharted territory to find enlightenment and, delightfully, a doctorate. I am forever grateful to Dr Ogawa Kohei, who undertook an onerous task of fostering me into competence—and thus bestowed upon me a brighter future with unbounded opportunities—as well as supervising a research project that must be very hard to supervise on my part. His kindness and skills in teaching and his faith in his students helped me weather the many challenges on this adventure. I am immensely grateful to Dr Harada Kensuke and Dr Iiguni Youji, whose insights and advice were indispensable for bringing my work to the next level. I am also grateful to Koyama Naoki-*san*, who I never met, but whose work paved the way for mine, and Dr Graham Peebles, who assisted me in composing an interesting scenario about a human, a robot, and aliens and lent me his voice. I sincerely thank Dr Yoshikawa Yuichiro and Dr Nakamura Yutaka, who together with Dr Ishiguro had brought me into the circle; I only regret I was unable to be more of assistance to them. I also thank Dr Hamed Mahzoon, who showed me how to present my research, and Dr Uchida Takahisa, who helped me set the stage for my defence. I thank Dr Oskar Palinko for being a kind and helpful office mate. I thank Marcos Inky Tae-*san* for being a great collaborator. I thank Nakajima Hanako-*san* for helping me apply for the great programme and for all her administrative assistance. I thank everyone in room J218, especially Dr Kawata Megumi, for helping me move my equipment when it was difficult for me to do it myself due to my fragile back. I thank Dr Kubota Tomonori, Sun Yuting-*san*, and Matsuda Risako-*san* for showing me how to operate robots. I thank Maeda Kentarou-*san* for familiarising me with the life at Ishiguro Lab and arranging a fun trip for us to Awaji Island. I thank Asakura Yoko-*san*, Chiaki Ii-*san*, Kashioka Makiko-*san*, Oshima Momo-*san*, Saito Misao-*san*, Saito Yukari-*san*, Dr Tachibana Tatsuya, Tanibata Kikuko-*san*, and Tasaka Emiko-*san* for their kind help and administrative support.

At last, I thank my parents for their love and support, and their patience for a son too slow to get married. And also anyone who I failed to mention here but offered me their kind help on this long journey to a world beyond.

Cogitans veritatis potestate mundos lustravi.

Bibliography

[1] Jennifer Goetz and Sara Kiesler. "Cooperation with a Robotic Assistant". In: *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '02. Minneapolis, Minnesota, USA: Association for Computing Machinery, 2002, 578–579. ISBN: 1581134541. DOI: 10.1145/506443.506492. URL: https://doi.org/10.1145/506443.506492.

[2] Sean Andrist, Bilge Mutlu, and Adriana Tapus. "Look Like Me: Matching Robot Personality via Gaze to Increase Motivation". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Korea: ACM, 18–23 April 2015, pp. 3603–3612. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702592. URL: http://doi.acm.org/10.1145/2702123.2702592.

[3] Maike Paetzel-Prüsmann, Giulia Perugia, and Ginevra Castellano. "The Influence of robot personality on the development of uncanny feelings". In: *Computers in Human Behavior* 120 (2021), p. 106756. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2021.106756.

[4] Liangyi Luo et al. "Towards a Personality AI for Robots: Potential Colony Capacity of a Goal-Shaped Generative Personality Model When Used for Expressing Personalities via Non-Verbal Behaviour of Humanoid Robots". In: *Frontiers in Robotics and AI* 9:728776 (2022). ISSN: 2296-9144. DOI: 10.3389/frobt.2022.728776. URL: https://www.frontiersin.org/article/10.3389/frobt.2022.728776.

[5] Anna Baumert et al. "Integrating Personality Structure, Personality Process, and Personality Development". In: *European Journal of Personality* 31.5 (2017), pp. 503–528. DOI: 10.1002/per.2115. eprint: https://doi.org/10.1002/per.2115. URL: https://doi.org/10.1002/per.2115.

[6] Sarah T. Völkel et al. "Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 25–30 April 2020, 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376210. URL: https://doi.org/10.1145/3313831.3376210.

[7] Solomon E. Asch. "Forming Impressions of Personality". In: *The Journal of Abnormal and Social Psychology* 41 (3 1946), pp. 258–290. DOI: 10.1037/h0055756.

[8] Warren T. Norman. "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings." In: *The Journal of Abnormal and Social Psychology* 66.6 (1963), 574–583. DOI: 10.1037/h0040291.

[9] Ernest C. Tupes and Raymond E. Christal. "Recurrent Personality Factors Based on Trait Ratings". In: *Journal of Personality* 60.2 (1992), pp. 225–251. DOI: 10.1111/j.1467-6494.1992.tb00973.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1992.tb00973.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1992.tb00973.x.

[10] Lewis R. Goldberg. "An alternative "description of personality": The Big-Five factor structure". In: *Journal of Personality and Social Psychology* 59.6 (1990), pp. 1216–1229.

[11] Lewis R. Goldberg. "The structure of phenotypic personality traits". In: *American Psychologist* 48 (1993), pp. 26–34.

[12] Robert R. McCrae and Paul T Costa. "Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires." In: *Journal of Personality and Social Psychology* 49.3 (1985), pp. 710–721.

[13] Liangyi Luo, Kohei Ogawa, and Hiroshi Ishiguro. "Identifying Personality Dimensions for Engineering Robot Personalities in Significant Quantities with Small User Groups". In: *Robotics* 11.1 (2022). ISSN: 2218-6581. DOI: 10.3390/robotics11010028. URL: https://www.mdpi.com/2218-6581/11/1/28.

[14] Isaac Asimov. "Liar!" In: *Astounding Science Fiction*. New York City: Street & Smith, 1941.

[15] John C. Loehlin. ""Interpersonal" experiments with a computer model of personality". In: *Journal of Personality and Social Psychology* 2.4 (1965), pp. 580–584. ISSN: 1939-1315 (Electronic),0022-3514 (Print). DOI: 10.1037/h0022457.

[16] John C. Loehlin. "Machines with personality". In: *Science Journal* 4.10 (1969), pp. 97–101.

[17] Patricia A. Cranton. "Computer Models of Personality: Implications for Measurement". In: *Journal of Personality Assessment* 40.5 (1976). PMID: 993958, pp. 454–463. DOI: 10.1207/s15327752jpa4005_2. eprint: https://doi.org/10.1207/s15327752jpa4005_2. URL: https://doi.org/10.1207/s15327752jpa4005_2.

[18] Clifford Nass et al. "Can Computer Personalities Be Human Personalities?" In: *Conference Companion on Human Factors in Computing Systems*. CHI '95. Denver, Colorado, USA: ACM, 1995, pp. 228–229. ISBN: 0-89791-755-3. DOI: 10.1145/223355.223538. URL: http://doi.acm.org/10.1145/223355.223538.

[19] Youngme Moon and Clifford Nass. "How "Real" Are Computer Personalities? Psychological Responses to Personality Types in Human-Computer Interaction". In: *Communication Research* 23.6 (1996), pp. 651–674. DOI: 10.1177/009365096023006002. eprint: https://doi.org/10.1177/009365096023006002. URL: https://doi.org/10.1177/009365096023006002.

[20] Aaron Edsinger, Una-May O'Reilly, and Cynthia Breazeal. "Personality through faces for humanoid robots". In: *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No.00TH8499)*. Osaka, Japan, 2000, pp. 340–345. DOI: 10.1109/ROMAN.2000.892520.

[21] Hiroyasu Miwa, Atsuo Takanishi, and Hideaki Takanobu. "Experimental study on robot personality for humanoid head robot". In: *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)*. Vol. 2. Maui, HI, USA, 2001, 1183–1188 vol.2. DOI: 10.1109/IROS.2001.976329.

[22] Kwan Lee et al. "Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction". In: *Journal of Communication* 56 (2006), pp. 754–772.

[23] Michael Argyle, Luc Lefebvre, and Mark Cook. "The meaning of five patterns of gaze". In: *European Journal of Social Psychology* 4.2 (1974), pp. 125–136. DOI: 10.1002/ejsp.2420040202.

[24] Cyril M. Franks. "Ocular movements and spontaneous blink rate as functions of personality". In: *Perceptual and Motor Skills* 16.1 (1963), p. 178.

[25] Kenya Matsumoto et al. "Factors influencing the processing of visual information from non-verbal communications". In: *Psychiatry and Clinical Neurosciences* 64.3 (2010), pp. 299–308. DOI: 10.1111/j.1440-1819.2010.02077.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1440-1819.2010.02077.x.

[26] John F. Rauthmann et al. "Eyes as windows to the soul: Gazing behavior is related to personality". In: *Journal of Research in Personality* 46.2 (2012), pp. 147–156. ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2011.12.010. URL: http://www.sciencedirect.com/science/article/pii/S0092656611001711.

[27] Sabrina Hoppe et al. "Eye movements during everyday behavior predict personality traits". In: *Frontiers in Human Neuroscience* 12 (2018), p. 105. ISSN: 1662-5161. DOI: 10.3389/fnhum.2018.00105.

[28] Atsushi Fukayama et al. "Messages embedded in gaze of interface agents — impression management with agent's gaze". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '02. Minneapolis, Minnesota, USA: ACM, 2002, pp. 41–48. ISBN: 1-58113-453-3. DOI: 10.1145/503376.503385. URL: http://doi.acm.org/10.1145/503376.503385.

[29] Kerstin Ruhland, Katja Zibrek, and Rachel McDonnell. "Perception of Personality through Eye Gaze of Realistic and Cartoon Models". In: *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*. SAP '15. Tübingen, Germany: Association for Computing Machinery, 13–14 September 2015, 19–23. ISBN: 9781450338127. DOI: 10.1145/2804408.2804424. URL: https://doi.org/10.1145/2804408.2804424.

[30] Takayuki Osugi and Jun I. Kawahara. "Effects of Head Nodding and Shaking Motions on Perceptions of Likeability and Approachability". In: *Perception* 47.1 (2018). PMID: 28945151, pp. 16–29. DOI: 10.1177/0301006617733209.

[31] Paul A. Bremner, Oya Celiktutan, and Hatice Gunes. "Personality Perception of Robot Avatar Teleoperators in Solo and Dyadic Tasks". In: *Frontiers in Robotics and AI* 4 (2017), p. 16. ISSN: 2296-9144. DOI: 10.3389/frobt.2017.00016. URL: https://www.frontiersin.org/article/10.3389/frobt.2017.00016.

[32] Tomoko Koda and Takuto Ishioh. "Analysis of the Effect of Agent's Embodiment and Gaze Amount on Personality Perception". In: *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. MA3HMI'18. Boulder, USA: Association for Computing Machinery, 2018, 1–5. ISBN: 9781450360760. DOI: 10.1145/3279972.3279973. URL: https://doi.org/10.1145/3279972.3279973.

[33] Liangyi Luo et al. "Robotic eyes that express personality". In: *Advanced Robotics* 33.7-8 (2019), pp. 350–359. DOI: 10.1080/01691864.2019.1588164. eprint: https://doi.org/10.1080/01691864.2019.1588164. URL: https://doi.org/10.1080/01691864.2019.1588164.

[34] Meng-Ju Han, Chia-How Lin, and Kai-Tai Song. "Robotic Emotional Expression Generation Based on Mood Transition and Personality Model". In: *IEEE Transactions on Cybernetics* 43.4 (2013), pp. 1290–1303. DOI: 10.1109/TSMCB.2012.2228851.

[35] Astrid Weiss, Betsy van Dijk, and Vanessa Evers. "Knowing Me Knowing You: Exploring Effects of Culture and Context on Perception of Robot Personality". In: *Proceedings of the 4th International Conference on Intercultural Collaboration*. ICIC '12. Bengaluru, India: Association for Computing Machinery, 21–23 March 2012, 133–136. ISBN: 9781450308182. DOI: 10.1145/2160881.2160903. URL: https://doi.org/10.1145/2160881.2160903.

[36] Funda Durupinar et al. "PERFORM: Perceptual Approach for Adding OCEAN Personality to Human Motion Using Laban Movement Analysis". In: *ACM Transactions on Graphics* 36.1 (Oct. 2016). ISSN: 0730-0301. DOI: 10.1145/2983620. URL: https://doi.org/10.1145/2983620.

[37] Michael L. Walters et al. "Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion". In: *Autonomous Robots* 24.2 (2008), pp. 159–178. ISSN: 1573-7527. DOI: 10.1007/s10514-007-9058-3. URL: https://doi.org/10.1007/s10514-007-9058-3.

[38] Jihong Hwang, Taezoon Park, and Wonil Hwang. "The effects of overall robot shape on the emotions invoked in users and the perceived personalities of robot". In: *Applied Ergonomics* 44.3 (2013), pp. 459–471. ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2012.10.010. URL: http://www.sciencedirect.com/science/article/pii/S0003687012001688.

[39] Yuki Yamashita et al. "Path Analysis for the Halo Effect of Touch Sensations of Robots on Their Personality Impressions". In: *Social Robotics*. Ed. by Arvin Agah et al. Cham, Switzerland: Springer International Publishing, 2016, pp. 502–512. ISBN: 978-3-319-47437-3.

[40] Matthew P. Aylett, Alessandro Vinciarelli, and Mirjam Wester. "Speech Synthesis for the Generation of Artificial Personality". In: *IEEE Transactions on Affective Computing* 11.2 (2020), pp. 361–372. DOI: 10.1109/TAFFC.2017.2763134.

[41] Andreea Niculescu et al. "Making social robots more attractive: The effects of voice pitch, humor and empathy". In: *International Journal of Social Robotics* 5.2 (2013), pp. 171–191. ISSN: 1875-4805. DOI: 10.1007/s12369-012-0171-x. URL: https://doi.org/10.1007/s12369-012-0171-x.

[42] Matthew P. Aylett, Yolanda Vazquez-Alvarez, and Skaiste Butkute. "Creating Robot Personality: Effects of Mixing Speech and Semantic Free Utterances". In: HRI '20. Cambridge, United Kingdom: Association for Computing Machinery, 2020, 110–112. ISBN: 9781450370578. DOI: 10.1145/3371382.3378330. URL: https://doi.org/10.1145/3371382.3378330.

[43] Elizabeth Broadbent et al. "Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived To Have More Mind and a Better Personality". In: *PLoS ONE* 8.8 (Aug. 2013), pp. 1–9. DOI: 10.1371/journal.pone.0072589. URL: https://doi.org/10.1371/journal.pone.0072589.

[44] Benedict T.C. Tay, Younbo Jung, and Taezoon Park. "When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction". In: *Computers in Human Behavior* 38 (2014), pp. 75 –84. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2014.05.014. URL: http://www.sciencedirect.com/science/article/pii/S0747563214002921.

[45] Adriana Tapus, Cristian Ţăpuş, and Maja J. Matarić. "User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy". In: *Intelligent Service Robotics* 1.2 (2008), p. 169. ISSN: 1861-2784. DOI: 10.1007/s11370-008-0017-4. URL: https://doi.org/10.1007/s11370-008-0017-4.

[46] Michiel Joosse et al. "What you do is who you are: The role of task context in perceived social robot personality". In: *2013 IEEE International Conference on Robotics and Automation*. Karlsruhe, Germany, 2013, pp. 2134–2139. DOI: 10.1109/ICRA.2013.6630863.

[47] Sarah Woods et al. "Is This Robot Like Me? Links Between Human and Robot Personality Traits". In: *5th IEEE-RAS International Conference on Humanoid Robots, 2005.* Tsukuba, Japan, 5–7 December 2005, pp. 375–380. DOI: 10.1109/ICHR.2005.1573596.

[48] Connor Esterwood et al. "A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445542. URL: https://doi.org/10.1145/3411764.3445542.

[49] Sarah Kriz et al. "Fictional robots as a data source in HRI research: Exploring the link between science fiction and interactional expectations". In: *19th International Symposium in Robot and Human Interactive Communication*. Viareggio, Italy, 13–15 September 2010, pp. 458–463.

[50] Ulrike Bruckenberger et al. "The Good, The Bad, The Weird: Audience Evaluation of a "Real" Robot in Relation to Science Fiction and Mass Media". In: *Social Robotics*. Ed. by Guido Herrmann et al. Cham, Switzerland: Springer International Publishing, 2013, pp. 301–310. ISBN: 978-3-319-02675-6.

[51] Eduardo B. Sandoval, Omar Mubin, and Mohammad Obaid. "Human Robot Interaction and Fiction: A Contradiction". In: *Social Robotics*. Ed. by Michael Beetz, Benjamin Johnston, and Mary-Anne Williams. Cham, Switzerland: Springer International Publishing, 2014, pp. 54–63. ISBN: 978-3-319-11973-1.

[52] Yi Mou et al. "A Systematic Review of the Personality of Robot: Mapping Its Conceptualization, Operationalization, Contextualization and Effects". In: *International Journal of Human–Computer Interaction* 36.6 (2020), pp. 591–605. DOI: 10.1080/10447318.2019.1663008. eprint: https://doi.org/10.1080/10447318.2019.1663008.

[53] Lionel P. Robert Jr. et al. "A Review of Personality in Human–Robot Interactions". In: *Foundations and Trends®in Information Systems* 4.2 (2020), pp. 107–212. ISSN: 2331-1231. DOI: 10.1561/2900000018. URL: http://dx.doi.org/10.1561/2900000018.

[54] Sarah Woods et al. "Are robots like people?: relationships between participant and robot personality traits in human–robot interaction studies". In: *Interaction Studies* 8.2 (2007), pp. 281–305. ISSN: 1572-0373. DOI: https://doi.org/10.1075/is.8.2.06woo.

[55] Eunil Park, Dallae Jin, and Angel P. del Pobil. "The Law of Attraction in Human-Robot Interaction". In: *International Journal of Advanced Robotic Systems* 9.2 (2012), p. 35. DOI: 10.5772/50228. eprint: https://doi.org/10.5772/50228.

[56] Amir Aly and Adriana Tapus. "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction". In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Tokyo, Japan, 2013, pp. 325–332. DOI: 10.1109/HRI.2013.6483606.

[57] Oya Celiktutan and Hatice Gunes. "Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience". In: *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Kobe, Japan, 2015, pp. 815–820. DOI: 10.1109/ROMAN.2015.7333602.

[58] Amir Aly and Adriana Tapus. "Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction". In: *Autonomous Robots* 40.2 (2016), pp. 193–209. ISSN: 1573-7527. DOI: 10.1007/s10514-015-9444-1.

[59] Sara Kiesler and Jennifer Goetz. "Mental Models of Robotic Assistants". In: *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '02. Minneapolis, Minnesota, USA: ACM, 2002, pp. 576–577. ISBN: 1-58113-454-1. DOI: 10.1145/506443.506491. URL: http://doi.acm.org/10.1145/506443.506491.

[60] Maartje de Graaf and Somaya B. Allouch. "Expectation Setting and Personality Attribution in HRI". In: *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Bielefeld, Germany, 2014, pp. 144–145.

[61] Jaime G. Carbonell. "Towards a process model of human personality traits". In: *Artificial Intelligence* 15.1 (1980), pp. 49 –74. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(80)90022-3. URL: http://www.sciencedirect.com/science/article/pii/0004370280900223.

[62] Luca Chittaro and Milena Serra. "Behavioral programming of autonomous characters based on probabilistic automata and personality". In: *Computer Animation and Virtual Worlds* 15.3-4 (2004), pp. 319–326. DOI: https://doi.org/10.1002/cav.35. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.35. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.35.

[63] Francois Bouchet and Jean-Paul Sansonnet. "Influence of Personality Traits on the Rational Process of Cognitive Agents". In: *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol. 2. Lyon, France, 2011, pp. 81–88. DOI: 10.1109/WI-IAT.2011.170.

[64] Anand S. Rao and Michael P. Georgeff. "BDI Agents : From Theory to Practice". In: *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)* (1995), pp. 312–319. URL: https://ci.nii.ac.jp/naid/10011450364/en/.

[65] Julio Bahamon and R. Michael Young. "An Empirical Evaluation of a Generative Method for the Expression of Personality Traits through Action Choice". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. 13 (1). Little Cottonwood Canyon, Utah, USA, 2017.

[66] Stephen Ware and R. Michael Young. "Glaive: A State-Space Narrative Planner Supporting Intentionality and Conflict". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 10.1 (2014). URL: https://ojs.aaai.org/index.php/AIIDE/article/view/12712.

[67] Unai Zabala et al. "Expressing Robot Personality through Talking Body Language". In: *Applied Sciences* 11.10 (2021). ISSN: 2076-3417. DOI: 10.3390/app11104639. URL: https://doi.org/10.3390/app11104639.

[68] Nikhil Churamani et al. "Affect-Driven Learning of Robot Behaviour for Collaborative Human-Robot Interactions". In: *Frontiers in Robotics and AI* 9 (2022). ISSN: 2296-9144. DOI: 10.3389/frobt.2022.717193. URL: https://www.frontiersin.org/article/10.3389/frobt.2022.717193.

[69] Katherine Isbister and Clifford Nass. "Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics". In: *International Journal of Human-Computer Studies* 53.2 (2000), pp. 251–267. ISSN: 1071-5819. DOI: https://doi.org/10.1006/ijhc.2000.0368. URL: https://www.sciencedirect.com/science/article/pii/S1071581900903689.

[70] Amos Tversky and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124. eprint: https://www.science.org/doi/pdf/10.1126/science.185.4157.1124. URL: https://www.science.org/doi/abs/10.1126/science.185.4157.1124.

[71] Fumitoshi Hayashi. "A multidimensional approach to the measurement of individual differences in interpersonal cognitive structure: An application of the INDSCAL model". In: *Japanese Journal of Psychology* 50.4 (1979), 211–218. in Japanese.

[72] Alessandro Vinciarelli and Gelareh Mohammadi. "A Survey of Personality Computing". In: *IEEE Transactions on Affective Computing* 5.3 (2014), pp. 273–291. DOI: 10.1109/TAFFC.2014.2330816.

[73] Francis Galton. "Measurement of Character". In: *Fortnightly Review* 36 (1884), 179–185.

[74] Poornima Kaniarasu and Aaron M. Steinfeld. "Effects of blame on trust in human robot interaction". In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. Edinburgh, UK, 25–29 August 2014, pp. 850–855. DOI: 10.1109/ROMAN.2014.6926359.

[75] Manja Lohse et al. "Evaluating extrovert and introvert behaviour of a domestic robot—a video study". In: *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. Munich, Germany, 1–3 August 2008, pp. 488–493. DOI: 10.1109/ROMAN.2008.4600714.

[76] *Household size and composition around the world*. Available online: https://www.un.org/en/development/desa/population/publications/pdf/popfacts/PopFacts_2017-2.pdf. (Accessed on 08 December 2021).

[77] *Small Business Statistics*. Available online: https://smallbiztrends.com/tag/small-business-statistics. (Accessed on 08 December 2021).

[78] *2019 White Paper on Small and Medium Enterprises in Japan 2019 White Paper on Small Enterprises in Japan (Summary)*. Available online: https://www.chusho.meti.go.jp/pamflet/hakusyo/2019/PDF/2019hakusyosummary_eng.pdf. (Accessed on 08 December 2021).

[79] David C. Funder. "On seeing ourselves as others see us: Self–other agreement and discrepancy in personality ratings". In: *Journal of Personality* 48.4 (1980), 473–493. DOI: 10.1111/j.1467-6494.1980.tb02380.x.

[80] Stefan Schaal. "Is imitation learning the route to humanoid robots?" In: *Trends in Cognitive Sciences* 3.6 (1999), pp. 233–242. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(99)01327-3. URL: https://www.sciencedirect.com/science/article/pii/S1364661399013273.

[81] Cynthia Breazeal and Brian Scassellati. "Robots that imitate humans". In: *Trends in Cognitive Sciences* 6.11 (2002), pp. 481–487. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(02)02016-8.

[82] Shinichiro Nakaoka et al. "Task Model of Lower Body Motion for a Biped Humanoid Robot to Imitate Human Dances". In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Edmonton, Alberta, Canada, 2–6 August 2005, pp. 3157–3162. DOI: 10.1109/IROS.2005.1545395.

[83] Tadas Baltrušaitis et al. "OpenFace 2.0: Facial Behavior Analysis Toolkit". In: *IEEE International Conference on Automatic Face and Gesture Recognition*. Xi'an, China, 15–19 May 2018.

[84] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. "Constrained Local Neural Fields for robust facial landmark detection in the wild". In: *IEEE International Conference on Computer Vision Workshops, 300 Faces in-the-Wild Challenge*. Sydney, Australia, 2–8 December 2013.

[85] Amir Zadeh, Tadas Baltrušaitis, and Louis-Philippe Morency. "Convolutional Experts Constrained Local Model for Facial Landmark Detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Honolulu, HI, USA, 21–26 July 2017, pp. 2051–2059. DOI: 10.1109/CVPRW.2017.256. URL: https://doi.org/10.1109/CVPRW.2017.256.

[86] Erroll Wood et al. "Rendering of Eyes for Eye-Shape Registration and Gaze Estimation". In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. Santiago, Chile: IEEE Computer Society, 7–13 December 2015, 3756–3764. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.428. URL: https://doi.org/10.1109/ICCV.2015.428.

[87] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. *The Big Five Inventory–Versions 4a and 54.* Berkeley, CA, 1991.

[88] Oya Celiktutan, Paul A. Bremner, and Hatice Gunes. "Personality Classification from Robot-mediated Communication Cues". In: *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. New York, USA, 26–31 August 2016.

[89] Koki Ijuin and Kristiina Jokinen. "Exploring Gaze Behaviour and Perceived Personality Traits". In: *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis*. Ed. by Gabriele Meiselwitz. Cham: Springer International Publishing, 2020, pp. 504–512. ISBN: 978-3-030-49570-1. DOI: 10.1007/978-3-030-49570-1_35.

[90] Andrew J. Elliot and Todd M. Thrash. "Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals". In: *Journal of Personality and Social Psychology* 82.5 (2002), 804–818. URL: https://doi.org/10.1037/0022-3514.82.5.804.

[91] Funda Durupinar. "Personality-Driven Gaze Animation with Conditional Generative Adversarial Networks CoRR in arXiv [Preprint]. Available at: https://arxiv.org/abs/2012.02224 (Accessed May 26, 2021)". In: *CoRR* abs/2012.02224 (2020). eprint: 2012.02224.

[92] Zachary Witkower and Jessica L. Tracy. "A Facial-Action Imposter: How Head Tilt Influences Perceptions of Dominance From a Neutral Face". In: *Psychological Science* 30.6 (2019). PMID: 31009583, pp. 893–906. DOI: 10.1177/0956797619838762. eprint: https://doi.org/10.1177/0956797619838762. URL: https://doi.org/10.1177/0956797619838762.

[93] Beatrice Rammstedt and Oliver P. John. "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German". In: *Journal of Research in Personality* 41.1 (2007), pp. 203–212. ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2006.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0092656606000195.

[94] Isaac Asimov. *I, Robot*. New York City: Gnome Press, 1950.

Glossary

Active goal: the goal towards which the robot is currently working.

Archetype: short for *personality archetype*, which is the 'target' personality a robot is engineered to be. It is a blueprint or example for a personality construct.

Broad trait: a facet of a robot's personality in a broader sense (than a trait, which is unique to a specific robot). Two robots can have the same broad trait, but no two robots have the same trait. Every broad trait corresponds to a personality dimension.

Cognitive state: an understanding of or response to the situation a robot finds themselves in; it can be a belief, thought, or any other simulated psychological state.

Colony: short for *robot colony*, which is a population of robot personalities situated in close association so that they will be observed by the same people for their shared qualities or goals. An essential property of a colony is that there is no duplication of personalities: every robot in a colony is unique.

Colony capacity: the number of individual personalities a generative personality model theoretically supports. One of the four major criteria for evaluating a generative personality model.

Commitment function: the function of a trait node for suggesting whether an urge for an action should be put into the corresponding action by mapping an urge and the probabilities of achieving the active goal conditioned on the action the urge leads to to a 'yes or no' decision.

Consistency: the negative variation among the impressions of the observed robot personality: the more varied the impressions are the lower the consistency. One of the four major criteria for evaluating a generative personality model and also a criterion for evaluating personality constructs.

Dominant trait: the primary trait of the robot's mission.

Effective colony capacity: the number of individual personalities distinguishable by some observers (e.g. a user group). A criterion for evaluating a generative personality model under practical constraints.

Fidelity: the proximity between a robot personality and its corresponding archetype in the personality space where they are defined or measured. One of the four major criteria for evaluating a generative personality model and also a criterion for evaluating personality constructs.

Generative personality model (GPM): a personality model capable of generating behaviour with inter- and intra-individual differences that reflect the inter- and intra-individual differences in personalities.

Goal: an end state a robot is operating to achieve: a purpose of the robot's activities or endeavours.

Goal-active form: the form a personality construct takes when working towards a goal, where the trait nodes are ordered to facilitate the attainment of the goal.

Goal node: a node representing a goal in a GPM. It reports a difference measurement between the actual state and the state where the goal is achieved.

Individuality: a combination of personality traits that distinguish an individual from others of the same kind. A fundamental property of personality.

Mapping function: short for *behavioural mapping function*, which is a function of a trait node for mapping cognitive states to urges.

Mission: the most important goal to a robot's existence. The functional difference between a mission and an ordinary goal is that an ordinary goal can be added to and removed (when attained) from a GPM, whereas a mission is always part of the robot's GPM. Usually, every robot personality has one and only one mission.

Mode: short for *mode of expression*, which is a distinct means of personality expression associated with a group of actuators or signalling apparatus.

Observer: an agent that can perform measurements of a robot's personality. The most common observers are humans.

Persona: the impression of a robot personality when they are working towards a goal that is not their mission, and it can thus be different from that of their true personality.

Personality construct: a particular build of a generative personality model with a unique set of parameters. Every personality construct corresponds to one and only one robot personality; even if we can make a copy of a personality construct, the two copies are one robot personality, not two, unless further development distinguishes one from the other.

Personality dimension: a dimension corresponding to a broader aspect of personality (i.e. a broad trait).

Personality space: the space of all possible configurations of personalities spanned by a set of personality dimensions, corresponding to a set of broad traits.

Primary trait: the trait corresponding to the highest-ranking trait node in the goal-active form of a goal is the primary trait for the goal.

Proactive personality: a robot personality whose commitment functions suggest acting on urges as long as the resulting behaviour does not work against the active goal.

Prominent trait: a trait applicable to and distinctly perceptible by a user group in a certain type of robot in its typical usage scenario.

Robot personality: a robot exhibiting characteristic patterns of computation and behaviour with inter- and intra-individual differences.

Robot personalities engineering: the systematic application of engineering methods to developing personalities for robots (the plural, 'personalities', is for emphasising that it is always about creating not one but multiple personalities).

Robot population: a group of robots. A robot colony is a robot population, but not necessarily vice versa.

Trait: short for *personality trait*, which is a unique quality of an individual robot specified by a *broad trait*, a broader facet of the robot's personality, and a *trait level*, a value that marks the level of strength of the broad trait.

Trait capacity: the number of broad traits (personality dimensions) out of all that are applicable to a type of robot a generative personality model can model. One of the four major criteria for evaluating a generative personality model.

Trait level: a value that marks the level of strength of a trait.

Trait node: a node representing a trait in a GPM. It consists of a mapping function and a commitment function.

Urge: an instance of certain behaviour not yet put into action: a potential action.

User group: a group of specific users of a specific robot, who are the main observers of the robot.

List of Publications

Journal Papers

1. Luo Liangyi, Koyama Naoki, Ogawa Kohei, and Ishiguro Hiroshi (2019) 'Robotic eyes that express personality', *Advanced Robotics*, 33:7-8, 350-359, DOI: 10.1080/01691864.2019.1588164

2. Luo Liangyi, Ogawa Kohei, and Ishiguro Hiroshi (2022) 'Identifying Personality Dimensions for Engineering Robot Personalities in Significant Quantities with Small User Groups', *Robotics*, 11:1, 28, DOI: 10.3390/robotics11010028

3. Luo Liangyi, Ogawa Kohei, Graham Peebles, and Ishiguro Hiroshi (2022) 'Towards a Personality AI for Robots: Potential Colony Capacity of a Goal-Shaped Generative Personality Model When Used for Expressing Personalities via Non-Verbal Behaviour of Humanoid Robots', *Frontiers in Robotics and AI*, 9:728776, DOI: 10.3389/frobt.2022.728776