Care Robots in Residential Homes for Elderly People

An ethical examination of deception, care, and consent

Karen Lancaster

PhD Thesis

University of Nottingham

2023

Acknowledgements

I would like to extend a huge thank you to my supervisors at the University of Nottingham, Dr Zachary Hoskins and Dr Neil Sinclair, for their diligent help, moral support, and constructive feedback.

I would also like to thank my parents for looking after Elliot while I attended conferences. To my friends and family, thank you for listening to me talk about robots. And Barney, I had many flashes of inspiration while out walking you; I miss you.

Most of all, I would like to thank my wonderful son Elliot: without you I would be incomplete. I am so proud that you are my son.

My studies have been funded by Midlands3Cities as a Collaborative Doctoral Award. Thanks are due to my industry partner Vecna, and David Clear in particular. All work and conclusions are my own.

Names used in thought experiments do not represent real persons.

Contents

	Page
Introduction	4
Chapter 1 Robot matrix: An examination of the definition of ‘robot’, types of robot, and ethical issues pertaining to robots	20
Chapter 2 Deception: A conceptual and normative analysis	59
Chapter 3 Robo-deception: Types of robo-deception and the danger (or lack of danger) of their occurring	103
Chapter 4 Fake compassion: A conceptual and normative analysis of emotionless carebots appearing to care	146
Chapter 5 Dignity: The varieties of dignity and their importance	182
Chapter 6 Consent: The nature, grounds, and importance of consent	211
Chapter 7 Consent and dignity: How consent-seeking promotes dignity	251
Conclusion	291
References	304
Appendix A Index of robots and related technologies	347

Introduction

"Counsellor Troi is going to need the comfort of a human touch,
not the cold hand of technology."

~ Dr Pulaski objects to Data (a robot) helping to deliver a baby, in *Star Trek: The Next Generation* ["The Child" S2, Ep1] (Bowman 1991).

Robots, AI, and related technologies are some of the most exciting innovations of our time. Even over the past few years, impressive advances have been made: we are seeing an explosion in the use and acceptance of smart speakers such as the Amazon Echo (Alexa)[1]; meanwhile, the abilities of AI software such as ChatGPT are phenomenal. People have long since dreamed of having robotic servants which can carry out the mundane and laborious chores we do not want to do, and the technological advances we are currently witnessing take us closer to this dream than ever before. The devices within our homes and workplaces are becoming increasingly

[1] More information on all the technologies – factual and fictional – mentioned in this thesis can be found in Appendix A.

interconnected (“the Internet of Things”) as security systems, thermostats, televisions, lights, and home appliances can all be controlled from a smartphone. At times it truly feels that we are on the cusp of an electrifying revolution.

Unfortunately, there are other aspects of life which do not fill us with the same levels of optimism about the future: the healthcare sector is an area where revolution seems increasingly necessary. By 2050, there are expected to be 2.1 billion people over age 60 worldwide (more than 1 in 5 people), including 426 million over age 80 (WHO 2022). Although life expectancy is steadily increasing, time spent living independently is not, meaning that there are mounting numbers of elderly people requiring care in residential homes. This increase in people requiring care, coupled with demographic shifts (a proportional decrease in the number of people of working age), is predicted to create a worldwide eldercare staffing shortfall, particularly in developed countries with low birth rates and high life expectancy. In the UK, there are already 100,000 caring vacancies which remain unfilled (Skills for Care 2021, NHS Support Federation 2022).[2] Similarly, in the US, there is expected to be a shortfall of care staff of around 150,000 by 2030, and 350,000 by 2040 (Miller 2017). The situation is unsustainable.

Since the industrial revolution, humans have been replacing themselves with machines which can perform tasks more quickly, more accurately and for longer durations than human workers possibly could. Robots continue to

[2] The covid-19 pandemic initially elicited a fall in unfilled vacancies (perhaps as people were keen to play their part in fighting the virus) but since October 2021, vacancy rates have returned to pre-pandemic levels (Skills for Care 2021).

replace today's factory workers, and clerical jobs are now under threat too: predictions regarding how many of us will be replaced by robots or AI by 2030 range from 20 million to 800 million worldwide (Connley 2017, Cellan-Jones 2019).

Robots designed for caring purposes – carebots – present a possible solution to the staffing shortfall in residential homes: they perform some of the work which has been hitherto undertaken by human nurses or informal carers. In Japan – a country where more than a quarter of citizens are over 65 – carebots are being increasingly utilised (Siripala 2018, Matuszek 2017). Carebots not only offer a potential solution to the staffing shortfall; they also present huge financial savings: robotic devices and AI applications are expected to save the US healthcare economy $150 billion each year (Accenture 2019).

Robots' abilities are ever-increasing, and we are becoming more accustomed to the idea of having robots in our homes, on our roads, and in our workplaces. However, although carebots are an exciting and unique development in healthcare provision, they are not without controversy. Some philosophers (Sharkey and Sharkey 2012, 2020, Sharkey 2014, Turkle 2017, Turkle et al. 2006, Sparrow and Sparrow 2006, Sparrow 2002) express serious concerns about carebots being used in residential homes. This thesis addresses some pertinent concerns, and goes some way towards justifying the use of carebots in residential homes in the near future, as well as suggesting some ways in which carebots ought to behave.

One might be forgiven for thinking this thesis is unnecessary – that programming nursing codes of conduct into carebots would be sufficient guidance for them. However, nurses are not simply walking codes of conduct:

they have years of life experience which they bring to the job – experience which carebots simply do not have. Nurses will (ideally) have learned acceptable ways of behaving towards other people, and they can draw upon their life experience to know – probably without being told – that some actions are more appropriate than others. A carebot has no such background knowledge, and would need to be programmed with information and instructions on how to behave vis-à-vis patient consent and dignity – topics which this thesis addresses. It is important to note, however, that this is a philosophical thesis; I do not suggest technical specifications for robots, nor programming guidance for roboticists.

1 Questions this thesis addresses

It is often useful to lay out in the introduction what one means by ambiguous or unfamiliar terms such as ‘carebot’. Defining a carebot as a robot which cares merely bisects the problem: now we might ask “What is a robot?” and “What is caring?” These questions are both explored in more detail within this thesis.

It might initially seem that ‘robot’ is a term which is commonly understood, but this is far from the case: is Alexa[3] a robot? Is a remote control C-3PO toy a robot? Is a cell phone? These questions are not simple, and some conceptual exploration is necessary to fully understand what is meant by the term ‘robot’, and to understand the different types of robot which exist. In Chapter 1, I provide a working definition of what a robot is: it is a machine which senses,

[3] Although the terms ‘Alexa’ and ‘Siri’ technically refer to the AI software rather than the devices *themselves*, I use ‘Alexa’ and ‘Siri’ to refer to the Amazon Echo and the Apple HomePod respectively, since this is how they are generally known.

thinks, and acts. I show why folk conceptions of 'robot' are not always useful; I discuss why driverless cars *are* robots, whereas remote control toys – whatever they look like – are *not* robots. More importantly than merely defining 'robot', I map the conceptual space and provide a matrix through which we can distinguish different types of robot according to their external appearance, and their intelligence levels. Onto the matrix we can plot robots as dissimilar as sexbots, industrial robots, and HAL-9000.[4] With the matrix established, I outline some of the different ethical questions which pertain to different types of robot. For example, concerns about robots' rights and responsibilities are relevant to highly advanced sentient robots, but not to industrial production-line robots; concerns about exacerbating gender inequalities arise from our interactions with sexbots, but not from the production of autonomous weapons.

One concern which permeates all areas of the matrix, and has been discussed within the robot ethics (or 'roboethics') literature, is deception. This is the focus of Chapters 2, 3, and 4. Chapter 2 takes a step back from robotic discussions in particular, to provide conceptual and normative analyses of both self-deception and other-deception (deceiving another person). What initially seems like a relatively clear phenomenon becomes less clear with more detailed analysis. For example, is it deceptive if I try to trick you into having a false belief, but you end up having a true belief? Can someone deceive unintentionally? These questions (and others) are examined, and I reach some necessary and sufficient conditions for other-deception. My normative analysis details why other-deception is generally viewed as wrong, yet there

[4] Although this thesis is grounded in reality, I often use fictional robots as examples.

are forms of prosocial deception – such as falsely saying I like your hairstyle – which facilitate pleasant social interaction and are morally unproblematic.

Chapter 3 focuses on robo-deception[5], and examines whether it is possible for robots to meet the necessary and sufficient conditions for deception laid out in Chapter 2. I suggest that although it is possible for highly advanced (fictional or futuristic) robots to deceive people, it is simply not possible for today's robots to do so, because they cannot *intend* to deceive. However, roboticists may deceive users via the robots they create. I argue that many writers expressing fears about robo-deception are what I call 'robo-deception alarmists' – that is to say, their fears are unwarranted or disproportionate to the (putative) threat. I distinguish four possible types of robo-deception: anthropomorphic deception, zoomorphic deception, disanthropomorphic deception, and basic other-deception, and I specify which type of deception pertains to which type(s) of robot.

Chapter 4 examines one particular type of (putative) robo-deception: when robots appear to care. Writers have questioned whether robots can indeed care (Sparrow and Sparrow 2006, Sharkey and Sharkey 2020, Turkle 2017, chap. 6), but we can only answer this question when we are clear about the meaning of 'care'. I outline two different meanings of the term: practical care (completing a set of necessary tasks), and emotional care (having a benevolent or compassionate affective state). Carebots can practically care, but cannot (yet) emotionally care – however, they can give the appearance of emotional care, by displaying what I call 'fake compassion'. Since carebots'

[5] I use this term to refer to instances where robots *themselves* deceive users, and to instances where roboticists use robots to deceive users.

outward behaviour is not caused by any affective state, critics might deem this to be deceptive. I argue that fake compassion is seldom deceptive, but even when deception does occur, it is a prosocial form of deception which promotes subjective wellbeing and better health outcomes for patients, and is morally unproblematic.

Carebots offer a pragmatic solution to the shortage of eldercare nurses, and it is reasonable to consider how carebots should behave. For practical care tasks such as folding bed sheets and dispensing medication, robots' programming is straightforward and philosophically uninteresting. However, determining how carebots should behave towards patients requires deeper exploration.

Ensuring that carebots respect and promote patients' dignity seems like a laudable goal; however, dignity is a slippery concept which is often poorly conceptualised and inconsistently used. In Chapter 5, I explore two conceptions of dignity: universal dignity (which we possess in virtue of being human), and variable dignity (which can increase or decrease depending on behaviour, treatment, and self-image). This lays the foundation for Chapters 6 and 7.

One crucial way of promoting someone's dignity is by taking their wishes into account: patient consent to medical procedures is well-covered in the philosophical literature – as is sexual consent. However, consent to non-medical care which nurses provide for elderly residential home patients is conspicuous by its absence in the philosophical consent literature, where sexual and medical consent receive almost all the attention. In Chapter 6, I put forward a schematic which categorises routine care activities which

carebots might some day undertake (such as helping patients with feeding, bathing, dressing, toileting, and walking) according to their level of invasiveness; consent is crucial for these activities because it promotes patient autonomy, bodily integrity, and dignity, as well as trust in the carebot (or nurse).

Some might contend that an account of routine consent is unnecessary for carebots, because they could simply utilise existing accounts of (sexual or medical) consent and apply these in routine care situations. In Chapter 7, I outline why this is not possible. Routine consent differs from medical consent in terms of information-giving, and it differs from sexual consent in terms of its transferability – and it differs from both medical *and* sexual consent in terms of its frequency. I argue that repeated non-consensual routine care can cause cumulative reductions in patients' dignity and wellbeing; thus, carebots (and human nurses) must obtain consent before providing routine care.

The concluding chapter draws together the themes and conclusions reached in each of the chapters, and I consider whether this thesis is over-predicting the usefulness of carebots, or whether we really *are* progressing into times when our lives are irrevocably intertwined with robots. I discuss what the covid-19 pandemic has shown us about the emotional necessity – but also the dangers – of human contact, and I consider carebots' role should another pandemic occur.

The robotic revolution is still in its infancy, and many questions still remain; some of these are outlined in my concluding chapter. Although further study is still required, I believe this thesis goes some way towards demonstrating

that carebots will be a useful addition to residential homes, and that their use is ethically defensible.

2 Carebots currently in existence

One might be forgiven for thinking that this thesis is merely a thought experiment based on science fiction – a discussion of robots which do not currently exist. However, this is not the case. Carebots with remarkable abilities currently exist (and robots with even more remarkable abilities are no doubt in development). Presently, I describe some real-life carebots and their abilities, as well as some related technologies whose abilities may be integrated into carebots in the near future.

Throughout this thesis I refer to ‘carebots’ generally, but carebots are not a homogeneous group of technologies. Nonetheless, they share some common traits – namely that they are robots which are designed to perform some caring functions (such as helping with medication, dressing, feeding, socialising, setting reminders, or monitoring vital signs). I now describe a few of today’s carebots, so that it is clearer which sorts of robots are under discussion in the thesis.

2.1 Gecko CareBot

The Gecko CareBot is a mobile service robot produced specifically for helping elderly people within their own homes – although it could also be used within care institutions, or to look after a child or disabled person (Gecko Systems 2019). Its electronics are concealed within a metallic frame, with a grey plastic outer shell. It is around 120 cm tall, with good stability, and moves about on wheels. It has a rectangular screen in its top section, which loosely resembles

a head; it has no limbs, and aside from its upright shape, is otherwise unhumanlike in appearance, though it has two-way verbal communication.

The Gecko CareBot has the following functions:

- moves autonomously about the environment, avoiding obstacles, and assists the patient in their daily routine
- telepresence[6]: a family member can remotely control it and move it about the patient's home
- sets reminders
- alerts patient when a caller is at the front door
- holds brief conversations, tells jokes, family anecdotes, recites Bible verses, plays music; has a customisable 'personality' and voice
- detects emergencies such as intruders, fires, patient falls, patient unresponsiveness, unexpected patient absence, and calls for help – and alerts family members / emergency services if needed
- can interface with medical equipment to monitor patients' vital signs, blood pressure, pulse, oxygenation levels, and blood sugar levels – and alerts family members / emergency services if needed

(Gecko Systems 2008, 2019, Cision 2012, Kerr et al. 2018)

2.2 Care-o-bot

The Fraunhofer-Gesellschaft Care-o-bot 4 (hereafter, Care-o-bot) is intended to support and care for people of any age in their home environment, in

[6] Roughly, telepresence consists of two-way video conferencing, where the absent party is also able to remotely control at least some of the robot's actions (Telepresence Robots 2019, Robot Center Ltd 2021).

customer service roles, delivering items in offices, or as a mobile and verbally-interactive guide (Fraunhofer-Gesellschaft 2018a). It has a white plastic outer shell, an upright body shape, two (removeable) arm-like limbs and a rounded 'head', giving it a loosely humanlike appearance; it is 158 cm tall, weighs 140 kg, and moves about on wheels. Its head is a touchscreen which can display text or a schematic face.

The Care-o-bot has the following functions:

- Moveable arms with grasping hands
- Able to navigate busy or messy areas
- Good overall physical dexterity (e.g. arms have spherical rather than simple hinge joints) with 29 degrees of freedom[7]
- Follows verbal commands to fetch and carry items
- Can support patients as they walk
- Can set reminders, connect to the internet to give responses to factual queries, and can display written information on its screen
- Facial recognition technology, captures moods, and can adapt appropriately
- Has an adaptable 'personality'

(Fraunhofer-Gesellschaft 2018a, 2018b, 2020)

[7] Degrees of freedom are a measure of manoeuvrability; the more degrees of freedom, the more manoeuvrable a robot is.

2.3 Pearl

Pearl was created by researchers at the University of Pittsburgh, University of Michigan, and Carnegie Mellon University in 2000,[8] primarily aimed at assisting elderly people suffering cognitive decline or chronic conditions to go about their normal everyday lives (Robotics Today 2021, Pollack et al. 2002). It is metallic, around 110 cm tall, with an upright body shape and a loosely humanlike face, with red lips, and movable eyes and eyebrows; Pearl's torso has visible electronics and a rectangular touchscreen.

Pearl has the following functions:

- Displays text or information on its screen
- Can support patients as they walk (adapting to patient's speed and rhythm)
- Can set intelligent reminders based on what it observes, send information to family or doctors, and facilitate video calls
- Sensors allow it to move about and detect patient falls
- Facial recognition
- Can engage in rudimentary chat

(Pollack et al. 2002)

2.4 Other support robots

The past few years has seen an explosion in the development and widespread acceptance of smart speakers – such as Alexa and Siri – which can:

[8] Although a little old now, Pearl's functions are impressive enough to make it worth mentioning here.

- understand natural language and respond verbally
- set reminders
- access the internet to find information
- play games
- some can also show videos, images, and written instructions on a screen, and enable video calls

It is reasonable now to expect a carebot to possess these functions as a bare minimum, but there are additional things we might want from a carebot. Here I briefly list some features which could be useful in carebots, and I note some technologies which currently possess these features:

- Companionship through (verbal) chat, humour, and adaptive or customisable 'personalities' (Buddy; Olly)[9]
- Learning people's faces, recognising them, and responding accordingly (Riba; Jibo; Pepper; Asimo)
- Sensing whether (and how) the device is touched, and responding accordingly (Paro; Jibo)
- Detecting hazards and emergencies within the home, and alerting emergency services or loved ones (iPal; Buddy)
- Learning routines and anticipating patients' needs (Olly; Jibo)
- Behaving like a pet (Paro; Aibo; Companion Pets)
- Communicating in multiple languages and translating (Pepper)
- Interfacing with medical equipment to monitor health (Jibo)

[9] See Appendix A for more information on these technologies.

- Providing physical assistance to patients, such as dressing them (PR2); lifting them (Riba); feeding them (MySpoon); or augmenting their strength when walking (Stride Management Assist; Hybrid Assistive Limb)

When I speak of present-day carebots, I mean to refer to devices such as Pearl, the Gecko CareBot, and the Care-o-bot, and others like them – particularly those which are verbally interactive and multi-functional. However, this thesis is also forward-looking, and a number of my discussions revolve around the sorts of carebots which we can reasonably expect to exist within the next couple of decades; I take it that future carebots may have the sorts of abilities currently present in the Gecko CareBot, Care-o-bot, and Pearl, plus some of the above listed abilities, such as social interaction, and helping the patient with feeding, toileting, dressing, and moving about (this is discussed further in the thesis).

In the near future, due to technological convergence, we are likely to see carebots which exhibit an increasing number of features from the list given above – and more besides. Technological convergence is a process whereby previously separate technologies become integrated into a single device. Smartphones are the paradigmatic example of technological convergence: the smartphone is a single device which has subsumed the functions of a camera, alarm clock, sat nav, calculator, computer, video games device, translator, mini television, and telephone. We are also seeing more interconnectivity between devices; future carebots may be able to connect to and interact with smartphones and home appliances to provide better care than ever before.

3 Terminology

I presently explain some of the terms used in this thesis. I use the term ‘nurse’ to refer only to professional human nurses and nursing assistants, plus other people employed in caring roles for elderly people, not including doctors. I use ‘healthcare professional’ to mean any professional employed in the health sector, including nurses, doctors, surgeons, and others.

The term ‘patients’ refers to elderly people (aged 65+) living in residential homes (non-familial eldercare facilities and nursing homes). They are taken to be neurotypical adults: by this I mean that they do not have learning or cognitive disabilities, severe mental health problems such as schizophrenia, or dementia. It is important to stress this from the outset, because one might reach different conclusions if one were to consider patients with dementia[10] or severe cognitive decline. For example, when I discuss humanlike robots and anthropomorphic deception, I suggest that a patient would probably not believe the robot is a human with emotions; this may not be true of patients with dementia. There are also different normative considerations pertaining to carebots for patients with dementia, which would need to be explored elsewhere. For example, issues of consent could be different, since diminished rationality and severe memory loss can occur among people with dementia. Furthermore, it may be confusing or frightening for patients with dementia to interact with robots rather than humans. The use of carebots for patients with dementia is an interesting topic which should be philosophically

[10] The risk of dementia increases with age: around 2% of people aged 65-69 have dementia, and around 17% of people aged 85-89 have dementia (Alzheimer’s Research UK 2022). People with dementia may require additional care.

explored; however, it is beyond the scope of this thesis, which pertains solely to neurotypical elderly patients.

4 Conclusion

We are living in unprecedented times: astonishing progress is being made in AI and robotics: the technology is not only becoming more advanced, but also cheaper to produce, lighter weight, and more accessible (Stanford University 2022). These advances will be useful in many different areas of our lives, and the eldercare sector is one area where robots will be invaluable. Reckless or hasty use of carebots may be problematic, however: the 'Move fast and break things' mindset which catapulted Facebook to success – and controversy (Ghosh 2018, Taneja 2019) – is not a sound model for the ethical deployment of carebots into residential homes. Their use needs to be thoroughly philosophically explored and empirically trialled in order to ensure that some of the most vulnerable members of our society continue to receive dignified, compassionate treatment throughout their later years. Although the subject is vast and the ethical issues are many, I believe that this thesis goes some way towards showing that although carebots are not a panacea, they will have an ethically defensible place in our future.

Chapter 1

Robot matrix: An examination of the definition of 'robot', types of robot, and ethical issues pertaining to robots

In the introductory chapter, I laid out what I mean by particular terms, such as 'nurses', 'patients', and 'residential homes'; I did not, however, explain what I meant by the term 'robot'. This was not an oversight: rather, it requires more thorough attention. This chapter provides a definition of 'robot' and a matrix to distinguish between different types of robot; it also discusses how this helps elucidate the ethical issues at stake for different types of robot.

Despite an increasing roboethics literature, it is surprising that few writers adequately define what they mean to refer to by 'robot'. This would be acceptable it were unanimously understood what a robot is; however, folk conceptions of the term are ill-defined, and the term is used in disparate ways within computer science, engineering, social science, and the humanities. Evidently, a mapping of the conceptual space is required so we can not only determine which contenders are robots, but also distinguish between different types of robot, whilst apprehending their similarities and differences. This

would then allow us, as roboethicists, to more easily visualise the ethical issues at stake for different types of robot, and to appreciate the interconnections between various roboethical concerns.

One might wonder why we should bother to define 'robot' at all; we do not dedicate philosophy theses or articles to defining or distinguishing between other objects such as mountains, chairs, or clocks. The reason is that classifying robots is of ethical interest and significance in a way that classifying mountains, chairs, and clocks is not. Specifying what robots are could have broad-ranging ethical and pragmatic implications for the future.

Not all roboethical concerns pertain equally to all robots: concerns about driverless cars, such as their counterintuitive or morally troubling decision-making (Lorente 2020, Hansson, Belin, and Lundgren 2021), are markedly different from concerns regarding sexbots, such as that they could objectify women and reduce empathy (Campaign Against Sex/Porn Robots 2022, Richardson 2019, 2016, Lancaster 2021). These concerns are different again from concerns about social robots, such as their deceptiveness (Leong and Selinger 2019, Danaher 2020).

The main objective of this chapter is therefore to demonstrate how different ethical issues pertain to different types of robot. First, I define 'robot': this helps map the conceptual space and determine the technologies with which we are working. I examine folk conceptions of 'robot', and the sense-think-act paradigm from the robotics industry. In §2, I provide a matrix which enables us to distinguish between different types of robot: it consists of two axes onto which robots can be plotted – one axis tracks visual resemblance to a human, and the other axis tracks the robot's intelligence. I show how this overcomes

some of the ambiguity and opacity in the literature, and provides us with a classification system suitable for future use. Finally, in §3, I discuss which types of robot (in which areas of the robot matrix) are most pertinent to which ethical issues.

This enables us to get a holistic view of the roboethics literature and the robots on which it focuses, as well as seeing connected ethical concerns relating to robots with different functions (for example, driverless cars and autonomous weapons). The robot matrix shows how ethical concerns may vary even when they relate to robots with similar functions – for example, carebots with advanced intelligence raise different concerns from carebots with rudimentary intelligence; we may be concerned about the rights of the former, but not the latter. Although it can be useful to limit one's work to particular types of robot, one risks overlooking a broader understanding of roboethics as a whole, and one could miss connections between ethical issues and types of robot. This chapter aims to rectify this by 'zooming out' and mapping various roboethical concerns and types of robot.

1 Defining 'robot'

The robotic revolution is gathering speed. Popular media frequently feature robot-related articles: themes include industrial robots taking our jobs (Cellan-Jones 2019, Ashbrook 2016, Carson 2019, Schulz 2013); 'cobots' becoming our new work colleagues (Business Life 2018, I-Scoop 2016); medical robots out-performing human doctors (Moon 2017, Summers 2019, Walsh 2020); driverless cars on our roads (Autoweek 2018, Cusack 2021, Hawkins 2022, Titcomb and Sabur 2018); military robots being used in warfare (Euronews 2022, Hambling 2022, Johnson 2022, Thompson 2022); sexbots becoming

our lovers (Wade-Palmer 2021, Tamblyn 2014, Shen 2020, Ghosh 2020); and carebots serving as doctors or nurses (Hotzak 2015, Matuszek 2017, Siripala 2018, Cairns 2021, Guevarra 2021). Robots will be increasingly present on our roads, as well as in our workplaces, homes, schools, and care institutions – and elsewhere – within the next few decades. Consequently, we are seeing a mushrooming philosophical literature dealing with the potential benefits and ethical problems posed by robots.

Despite increased scholarship, many philosophical papers focusing on roboethics fail to adequately define what they mean to refer to by the term 'robot'. Some writers specify the types of robots to which their work pertains – for example, robot nannies (Sharkey and Sharkey 2010) or robot carers (Sorell and Draper 2014) – without spelling out what they consider a robot to be. Some writers give 'definitions' of robots which are so broad that they capture just about all kinds of technology, such as "machines, software, programmes" (Rainey 2016: 225). Some writers (Borenstein and Pearson 2010, Vallor 2011, Wallach 2010, Coeckelbergh 2014) do not indicate what they mean by 'robot' nor what types of robot they are discussing; they surely do not mean to discuss to *all* types of robot, since one argument would not apply equally to sex robots, driverless cars, military robots, and social robots. This terminological ambiguity means that the roboethics literature can be difficult to untangle: how can one engage with an ethical argument when our understanding of precisely *what* is under discussion remains opaque?[11]

[11] Some writers (such as Latikka et al (2019)) give clear definitions of what they mean to refer to by specific terms such as 'service robot' but such definitions only pick out a subset of robots.

If philosophers use the term as if it is universally agreed upon, and there is no simple way to distinguish between robots of different kinds, then as the roboethics literature continues to grow, it is likely to become increasingly muddled. Talking at cross purposes is a phenomenon which is already being borne out in the literature. For example, Sparrow and Sparrow (2006) argue that a robot cannot care for a patient; Meacham and Studley (2017) on the other hand, maintain that a robot *can* care. It may seem that these philosophers oppose one another: Meacham and Studley (2017) even claim they are arguing *against* Sparrow and Sparrow. However, on closer inspection, one sees that the robots under discussion in the two papers are a world away from one another. Sparrow and Sparrow's arguments demonstrate they are concerned with real-life robots (from 2006) with limited intelligence and social skills. Meacham and Studley's futuristic and hypothetical robots, by contrast, are intelligent, adaptive, highly social, and offer an excellent standard of emotional care to patients. With such different conceptions of robots, it is little wonder that the two articles reached different conclusions about whether a robot can care.

The term 'robot' is plainly in need of philosophical clarity. Gunkel (2018: 13–26) offers a discussion of our use of the term, and notes the difficulty in defining it, as well as people's varied use of it. What he does not offer, however – and what I offer herein – is a way to distinguish between different *types* of robot to help elucidate roboethical arguments. This not only helps me be specific about the types of robot *I* discuss, but could enable future roboethicists to ascertain whether they are discussing the same types of robot as their ostensible opponents.

I thus begin by giving an expansive definition of robots as a whole; once we have a broad definition, more nuanced *types* of robot can be situated within the conceptual space, thus providing a comprehensive framework.

When it comes to the definition of a robot, one commonly agreed-upon thing is that there is not a commonly agreed-upon definition of a robot! Indeed, some scholars (Bertolini 2013: 216, Nourbakhsh 2013: xiv) suggest that the difficulty of defining 'robot' means that we should perhaps refrain from even *attempting* to define it. However, I disagree: we do not have universally accepted definitions of other concepts such as knowledge, the mind, and wrongness, despite millennia of scrutiny. Yet these subjects are philosophically interesting and worthy of study in spite of (or perhaps because of) the disagreement, and discussing definitions of 'robot' is still useful even if no perfect definition exists.[12]

Definitions can be stipulative (what is meant by a term in a given context), or linguistic (what people in general mean by a term). I use a linguistic definition which is frequently utilised in industry. Although my definition of a robot may not enable us to clearly categorise all entities as either 'robot' and 'non-robot', it should give us a yardstick against which we can check whether a particular entity is clearly a robot, clearly not a robot, or a borderline case. Borderline cases can be included in the robot matrix, particularly if they are of ethical interest.

[12] Universal agreement with the definition of 'robot' I suggest is not a prerequisite for using my matrix; the matrix remains a useful way of categorising robots even if a slightly different definition is preferred by others.

1.1 Folk conceptions of robots

Let us begin by examining folk conceptions of robots – viz. ways in which most people use the term; these are useful because a definition which bears little resemblance to general usage is implausible. However, folk conceptions are sometimes poorly thought through, and even self-contradictory, so it would be poor scholarship to take them as gospel.

A folk conception of a robot is something metallic with a loosely humanoid[13] shape; for example, C-3PO and Asimo have sleek, humanlike bodies and bipedal locomotion; one could easily believe that Asimo, like C-3PO, is a human actor in a suit. Number 5 and Pearl are slightly less humanlike technologies which nonetheless fulfil folk conceptions of robots: they are metallic, with upright body shapes, visible circuitry, and basic 'faces', but do not resemble humans in suits.

Not everything that looks like a metallic human is a robot, however. In London's *Madame Tussauds*, one can find a waxwork figure of C-3PO (Merlin Entertainment Group 2020), and skilled pastry chefs could create a cake which looks just like C-3PO. Such creations are not robots: they are waxworks and cakes, just as a waxwork or cake resembling a dog is not in fact a dog.

Folk conceptions of 'robot' would also pick out remote control toy robots (which look like a metallic man or animal) and inaccurately call them robots. However, such human-controlled toys are not in fact robots. We must therefore move

[13] Something which is shaped similar to a human. This can come in degrees, as discussed below.

beyond folk conceptions of 'robot' if we wish to avoid identifying robots merely by their appearance.

1.2 Sense-think-act paradigm

Unfortunately, even roboticists and engineers do not fully agree on the definition of a robot (Jordan 2016: 3–4, Gunkel 2018: 13–26). Joseph Engelberger, the 'father of robotics', admitted even *he* could not define a robot (Guizzo 2020). However, the sense-think-act paradigm (sometimes sense-plan-act) was accepted by the International Service Robot Association (ISRA) as a working definition of a robot during the 1990s, and continues to be commonly cited today (see Pransky 1996, Siegel 2003, Nash and Open Robotics 2022).[14] The paradigm suggests that a robot is a machine which senses its environment, thinks about what it senses and plans a course action, then takes action. The conditions are individually necessary and jointly sufficient; thus, all and only those machines which can sense, think, and act are robots.

As philosophers, we cannot help but want to drill down further into what is meant by 'machine', 'sensing', 'acting', and probably most of all, 'thinking', given that all these terms are open to interpretation. One could write scores of theses analysing each of these terms, and there is insufficient space here to explore each one fully, though I shall briefly explain each of these features (thinking and intelligence are discussed further in §2.2):

[14] One reason for providing a linguistic definition rather than my own, stipulative, definition is that the sense-think-act paradigm is widely (though not universally) used, and is fit for purpose.

- **Machine**: An inorganic, artificial construction (typically made from metal and/or plastic) with different parts which work together
- **Senses**: Acquires information from the environment; senses may be similar or dissimilar to human senses
- **Thinks**: Has programming which enables it to process information and autonomously plan a (simple or highly complex) course of action
- **Acts**: Takes purposeful[15] action; this could involve a speech output, a movement, or something else

(list adapted from Winfield 2012: 8)

I accept that these are not perfectly complete analyses of the terms involved, but they provide a little more clarity than existed previously. The elements of the sense-think-act paradigm are causally interconnected: the sensing causes the thinking, which causes the action (which may lead to yet more sensing); all are made possible due to the robot's design. The sense-think-act paradigm is generally accepted by roboticists (Zaraki et al. 2017, Chapman, Gray, and Headleand 2015, Toshiba 2020, Bosch 2022, Aerospace Robotics 2014, Arkin 1998: 130, Bekey 2015: 2),[16] and is used in some philosophical literature (Winfield 2011, 2012, van Wynsberghe 2016: 40, Danaher 2020: 118, Nyholm 2020: 9).

[15] Defining 'purposeful' action is complex, but I mean to refer to an action which has been decided upon for some reason other than mere randomness – perhaps to accomplish some goal. For example, a robot saying "Hi there!" when it detects a new person in the room is purposeful action; a robot saying "Hi there!" at random intervals is not purposeful action.

[16] There are, however, examples of roboticists themselves using the term 'robot' to apply to technologies which do not think or sense, and are remotely controlled by a human (for example, Guizzo 2010).

An alternative, narrower definition used within some of the roboethics literature is ‘embodied artificial agent’ (Chrisley 2003, Wykowska, Chaminade, and Cheng 2016, Cappuccio, Peeters, and McDonald 2019: 2, Steels and Brooks 2018, Danaher 2020: 118, Fernández-Rodicio et al. 2022). Although such a definition does refer to robots, it only serves to pick out a subset of robots – ones which are intelligent enough to be called an agent. Defining agency is no easy task; it can be understood in various ways (see Schlosser 2019). Agency may involve some form of ‘mental state’ (another multifaceted term) or first-personal experience; agency seems to be something more than merely planning and carrying out an action. If I am right about agency, then less intelligent robots (such as industrial robots) are not artificial agents, and so the sense-think-act paradigm provides the broader definition of ‘robot’ with which I will be working. (If I am wrong, and agency is no different from thinking and acting, then the two definitions are the same and thus equally broad).

One might wonder how robots and AI are related. Some writers suggest that a robot – but not AI – is necessarily embodied and can interact with its environment (Sparrow and Sparrow 2006: 145). However, this distinction is problematic because AI systems necessarily exist within some sort of hardware – a machine – which is a form of embodiment, even if the software can be transferred to another device. AI chatbots (such as Replika and ChatGPT) sense speech or text inputs, think about (plan) a suitable response, then interact with the ‘environment’ by giving a speech or text output to the human user; this meets all the sense-think-act criteria. Therefore, this thesis considers all AI entities (which sense, think, and act) to be robots, echoing some roboethical arguments focusing on AI technologies (Danaher 2019a: 21,

2020: 118, Rainey 2016: 225, Bryson and Winfield 2017: 117, Leong and Selinger 2019: 300).

Many machines which we have in our homes and offices – such as stereos, fridges, and power drills – although increasingly sophisticated, are not robots, as they do not (currently[17]) fulfil the thinking criterion and take purposeful action, and they have few or no senses. Accordingly, we do not see ethical arguments within the roboethics literature focusing on stereos, fridges, and power drills. We should also note that most toy robots are not actually robots (they do not sense, think, and act), just as teddy bears and dolls are not *really* bears and babies: they are infantile approximations of something.

It is important to note that there are not always clear dividing lines between robots and non-robots, because some features of a robot – such as sensing and thinking – can come in degrees, meaning there may be some borderline cases; this is unproblematic, and the broad definition is still useful.

2 Robot matrix

I have defined a robot as a machine which senses, thinks, and acts: this allows us to determine which technologies count as robots. We are now about to explore varieties of robots within the broad definition. In this section, I explain my robot matrix, which gives us a way to distinguish between types of robot according to their intelligence and their appearance.

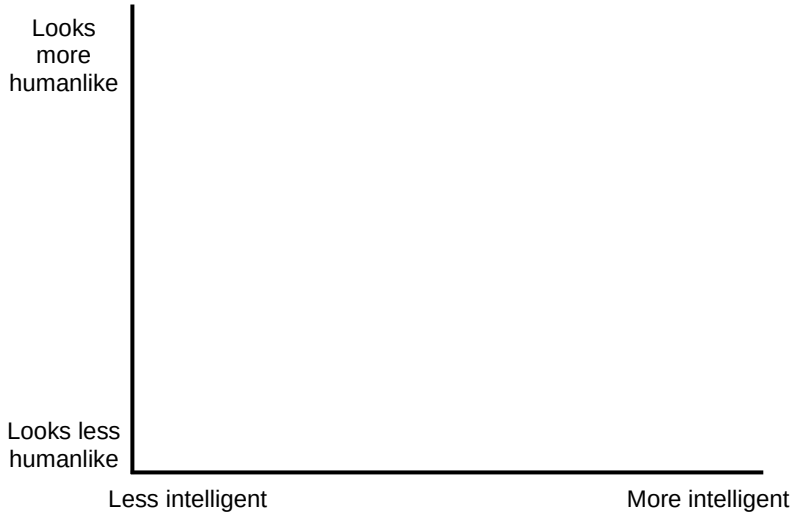
[17] This may change as technologies become smarter and more interconnected. For example, a smart kitchen could interact with the user's calendar and social media to find the food preferences of dinner guests, then design recipes and order appropriate food, and prepare the food for consumption.

Although not all roboethicists define what they mean by ‘robot’, those who do clarify the sorts of robots they are discussing commonly refer to the same two variables: the robot’s appearance, and its intelligence. These two variables form the axes of my matrix, onto which we can plot different types of robot. The robot matrix helps to capture existing robots, future robots, and fictional robots, all of which may be of interest to roboethicists. Later, in §3, I show how the robot matrix can be useful in analysing the roboethics literature holistically, clarifying roboethical issues and how they relate to robots of particular types.

Although many robots of science fiction look very humanlike and are highly intelligent (Data; Pris; Mia), it would be a mistake to assume that there is a necessary connection between a robot’s appearance and its intelligence. Robots which look very humanlike may have very limited intelligence (for example, a sexbot), or they may have intelligence levels which far surpass humans’ (for example, Data); on the other hand, robots may look nothing at all like a human whilst having very limited intelligence (for example, an industrial robot), or look unhumanlike but be highly intelligent (for example, HAL-9000) (Nyholm 2020: 9–10).

I discuss the terms I use (intelligence, humanlike appearance) in more detail below, and explain why I use them, but for now let me briefly explain the robot matrix itself. What I propose is as follows: a vertical axis which tracks the robot’s appearance: robots which look less humanlike appear towards the bottom of the axis, and those which look more humanlike are located towards the top. The horizontal axis tracks the robot’s intelligence: less intelligent robots appear towards the left of the axis, and those with greater intelligence

appear towards the right. In this way, we can visually compare various robots according to both their intelligence and their appearance.



It can sometimes be useful to group robots by their function (Royakkers and van Est 2015a: xii) – for example, concerns about indiscriminate killing by military robots are relevant regardless of the robots' intelligence levels or appearance. However, grouping by function alone can sometimes lead to misunderstanding: as I mentioned in §1, sometimes philosophers (such as Sparrow and Sparrow (2006) and Meacham and Studley (2017)) discuss robots with the same function (caring), which are radically different from one another, and the result is that arguments miss the mark. This is because carebots (and some other robots) come in a range of shapes, with substantial variation in their intelligence. My robot matrix can help to untangle the literature and facilitate a holistic understanding of roboethical issues.

Now I shall discuss why intelligence and humanlike appearance are pertinent variables in roboethical debates, and how we are to understand the terms. My use of these terms draws on existing literature, but they are stipulative

definitions: I explain how these terms should be understood when utilising my robot matrix. After that, I demonstrate the usefulness of the matrix in disentangling the complex roboethics literature, and the sorts of issues that are most applicable to different types of robot.

2.1 Appearance

This sub-section answers the questions: Why is 'appearance' one of the axes on my robot matrix (why not some other feature)? Why is a robot's humanlikeness an important or useful way of categorising robots?

A robot's visual appearance[18] is usually the first thing we notice, and our familiarity with humans' physical appearance means that humanlikeness is a yardstick against which robots can be roughly measured. Numerous roboethicists comment on the appearance of the robots they discuss, or divide technologies based on whether or not they look humanlike. For example, Cappuccio et al distinguish humanoid robots from "non-anthropomorphic machines, like vehicles, weapons, or industrial robots" (2019: 12). Danaher variously remarks that today's robots can take on several forms, sometimes looking recognisably humanoid, and sometimes looking decidedly unhumanlike – such as a tabletop or handheld device (Danaher 2019a: 21, 2019b: 2, 2020: 118). Sorell and Draper neatly distinguish between humanoid and non-humanoid robots: defining a humanoid robot as having an upright body with arms, and a head with facial features (Sorell and Draper 2014: 184). Nyholm discusses locating robots along a spectrum according to their

[18] 'Appearance' only refers to external appearance. Clearly, a robot's inner workings will be different from a human's.

humanlikeness: he writes that if robots such as Pepper are in the middle of a spectrum, then many real-life robots such as warehouse robots and driverless cars are at one end of that spectrum, and highly humanlike robots such as Sophia are at the other end (Nyholm 2020: 8).

These writers and others like them mention the humanlike or unhumanlike appearance of the robots they discuss because it is of some philosophical importance. We may, for example, be more morally troubled by violence towards a robot which looks highly humanlike (or animal-like) as opposed to one which looks like a box (Whitby 2008).[19] Concerns may relate to the robot itself – whether it can be raped, harmed, enslaved, and suchlike (Cappuccio, Peeters, and McDonald 2019, Danaher 2019b, 2019c, Eskens 2017, Chomanski 2020); other concerns may relate to how treatment of humanlike robots can negatively affect human individuals or society in general (Richardson 2016, 2019, Lancaster 2021, Sharkey 2014, Turkle 2017, Harvey 2015). Whether a robot looks humanlike is of key importance to roboethicists in a way that many other factors are not.

There are alternative ways in which we *could* categorise robots on a matrix – such as by their colours or materials (plastic, metal, silicone). Although robots come in a variety of colours and materials, roboethicists almost never comment on them, nor do they categorise robots according to these features, because the differences between robots' colours and materials are simply not

[19] This issue was explored in an episode of *Star Trek: The Next Generation* ["The Measure of a Man" S2, Ep9] (Scheerer 1989). In the episode, roboticist Commander Maddox wants to disassemble Data, however Data and other characters object. Maddox scoffs: "You are endowing Data with human characteristics because it looks human – but it is not. If it were a box on wheels, I would not be facing this opposition!"

of consequence to roboethicists.[20] The sorts of ethical issues which are applicable to red plastic driverless cars are also relevant to metallic blue driverless cars: it is simply unimportant and unhelpful to categorise robots according to colour or material (and indeed many other features too). But roboethicists frequently *do* categorise robots according to their humanlikeness and/or comment on the extent to which a robot appears humanlike, because this factor is highly relevant to roboethical debates.

What do I mean by humanlikeness? Some features which would make a robot more visually humanlike would be a vertical body, a head and face, arms, grasping hands, legs with feet, and being roughly human height (4-6 feet tall). A robot with only a couple of these features, such as PaPeRo (armless, legless, 50cm tall, upright, basic face) is less humanlike than a robot with more of these features, such as C-3PO (upright, adult height, moveable arms, legs, and hands, and a basic face).

Bołtuć gives some criteria for ascertaining whether a robot can pass for a human (what he calls the 'Church-Turing standard'). The visual criteria are (1) looking indistinguishable from a human at a set distance or in a photograph, and (2) the robot's motor movements and facial expressions are indistinguishable from a human (Bołtuć 2017: 216).[21] There are many present-day robots such as Pepper, QTRobot, iPal, and PaPeRo, which are loosely humanoid in appearance. Some robots – such as sexbots – fully meet

[20] When robots are *very* humanlike in their looks, their colour may be more pertinent – e.g. the 'skin' colour of robot slaves is important because of the troubling messages which such robots may convey about human enslavement.

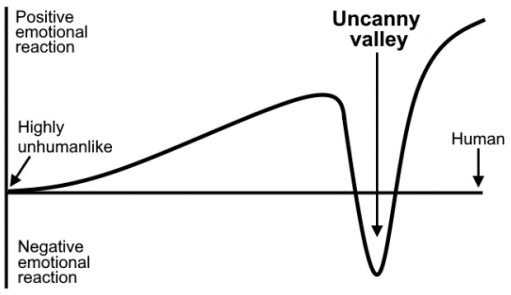
[21] Bołtuć also suggests some other features which are not related to visual appearance.

criterion 1, but there are currently no robots which meet criterion 2; they are thus within the uncanny valley.[22] In fiction, the replicants in *Blade Runner* (Scott 1982), and some of the synths in *Humans* (Arnold et al. 2015) meet both of Bołtuć's criteria, and can indistinguishably blend in with humans: these are the pinnacle of the humanlikeness scale.

Of course, many robots do not look humanlike – some are animal-like (such as Paro, Aibo, and Companion Pets), and some (driverless cars, military robots, smart speakers, and industrial robots) do not resemble any living creature. Roboethical arguments regarding smart speakers, for example, do not focus on appearance, because whether the speaker is a cube, a sphere, or a cylinder is irrelevant. However, when a robot appears humanlike – even only loosely humanlike – writers are quick to point this out, because the robot's appearance – its humanlikeness – affects many of the roboethical issues at stake. This is why visual appearance (*qua* humanlikeness) is one of the axes of my robot matrix.

When plotting robots on the vertical axis, they are ordered according to how humanlike they appear (see next page). I do not claim that robots which resemble a human are *better* than those which do not, I merely suggest that we can organise robots in terms of how humanlike they appear. At the top of

[22] People find robots which appear fairly humanlike – yet not *entirely* humanlike – creepy and unsettling (e.g. Sophia and the Geminoid). Loosely humanlike robots (e.g. Pepper) are perceived more positively (Mori, MacDorman, and Kageki 2012).



Most humanlike

Pris (looks totally human)

C-3PO (like a gold suit of armour)

Number 5 (quite mechanical, loosely humanlike)

Spot (yellow with 4 legs, no head)

Alexa (small black cylinder)

Least humanlike

the axis, we have robots which are indistinguishable from humans (Pris); lower down we have robots with a loosely humanoid appearance but which would not be mistaken for a human (C-3PO, Number 5). Lower still we have animal-like robots, which have limbs and a 'body', which humans also possess (Spot), and at the bottom are robots which are very visually unhumanlike (Alexa).

I acknowledge that towards the bottom of the axis, it becomes tricky to sort robots according to their (un)humanlikeness (for example, a driverless car and a smart speaker). This is not an insurmountable problem, however, since my robot matrix does not demand *exact measurements* of humanlikeness, but rather, it provides a 'family resemblance'. If someone believes a driverless car looks slightly more humanlike than a smart speaker (say, because its headlights resemble eyes and its wheels resemble limbs), but someone else believes a particular smart speaker looks more humanlike (say, because it has an upright shape and its buttons resemble eyes), then this is not hugely problematic. We can all agree that both smart speakers and driverless cars look very unhumanlike, and belong very close to the bottom of the axis.

There are some other difficult cases which do not easily map onto my robot matrix, such as AI software. Suppose there is a chatbot (such as Replika) which appears as a humanlike avatar on a smartphone screen: should this chatbot be placed high on the vertical axis, because the avatar appears humanlike, or should it be placed low on the vertical axis, because the

hardware on which it appears (the smartphone) is a very unhumanlike black rectangle? This is not an easy question to answer: for most robots, it is the hardware which we judge to be visually humanlike or unhumanlike – however this is not always the case. Some robots, such as the Care-o-bot and QTrobot, have heads with a screen at the front, which displays the face. In such cases, I believe we would say that the Care-o-bot and QTrobot "have faces", because we perceive what is displayed on the screen to *be* the robot's face. Similarly, when we converse with software on a screen, we perceive the humanlike avatar – not the background, nor the smartphone itself – to *be* the robot, just as if we were having a video call with another human. For example, in *Red Dwarf*, Holly is seen as *being* the (very humanlike) face on the screen – but because he exists purely on a screen, he does not seem as humanlike as a robot with a humanoid body does. Thus, if the avatar is humanlike, I suggest that the robot should be placed high on the vertical axis, but not as high as a robot with a body would be.

An even trickier case would be a chatbot which converses solely through text (or speech), and the user never sees an avatar at all. In cases where no avatar is seen, we only judge the (un)humanlike appearance of the hardware we see. Thus, ChatGPT and Babylon Health's AI doctor should be placed low on the humanlikeness axis, since the hardware on which they appear (a smartphone or computer) is very unhumanlike.[23]

[23] The nature of software means that it can be transferred onto other devices; it seems theoretically possible to run the ChatGPT software on a humanoid robot, in which case, one should place it according to how humanoid the robot looks.

2.2 Intelligence

In the sub-section above, I focused on the 'appearance' axis of the robot matrix; we now move on from that, to focus on the 'intelligence' axis. On this horizontal axis, robots with lower intelligence appear on the left of the axis, whereas robots with greater intelligence appear on the right. This sub-section answers the questions: why is intelligence one of the axes on my robot matrix (why not some other feature)? Why is it important or useful to categorise robots according to their intelligence?

What a robot can *do* is something which is of interest to roboethicists and lay people alike. There is not an extensive philosophical literature on go-karts, mannequins, radios, and cuddly toys. Although these things bear some similarities to driverless cars, sexbots, smart speakers, and robotic pets (respectively), the first set of items is philosophically uninteresting because they cannot do anything which elicits ethical debate. By contrast, the robots in the latter set are more ethically interesting because of what their intelligence enables them to do.

Intelligence is not a simple concept to define. Artificial intelligence (AI) is often described as "getting machines to do things that would be considered intelligent if done by people" (Turkle 2017: 63), but this definition does not fully capture the intelligence of machines which do not meet human standards. For example, a robot which can pick up an apple from a table would require AI (it would need to understand what it sees, plan what to do, and control its limb). However, if a human were simply to pick up an apple, we would not think: "That's intelligent!" Many animals, and people with profound cognitive impairments can do this simple task, so apple-grabbing does not really seem

intelligent when done by humans, but *does* require AI if done by a machine. There are levels of intelligence, and although picking up an apple requires a *little* intelligence, intelligence can extend far beyond that simple capability.

Above, when I discussed humanlike appearance, I noted that the pinnacle of humanlikeness is a robot which is indistinguishable from a human. Intelligence knows no such bounds.[24] Robots may have abilities which humans have, but to a far greater degree (for example, performing complex mathematics in a nanosecond), or they may have intelligence enabling them to do things unlike anything humans can do. Robots' intelligence could include:

- **Human intelligence**: conversing in natural language in real-time; sentience;[25] having a theory of mind;[26] understanding sensory information; making autonomous decisions based on deliberation; navigating their environment; learning through trial and error, and other sources of information; having and understanding its emotions
- **Superhuman intelligence**: acquiring and understanding information at astonishing speeds; understanding dozens of languages; deciphering codes; incredible mathematical ability
- **Animal-like intelligence**: understanding sensory information such as infra-red and ultraviolet, heat signatures, or electrical fields

[24] There is no limit to how far the horizontal axis can extend. If we can conceive of a robot which far exceeds even C-3PO's abilities, the horizontal axis can extend further to accommodate such robots.

[25] Sentience can be understood in different ways. I use it to mean 'phenomenally consciousness' – in other words, that there is something it is *like* to be that entity.

[26] By this I mean the ability to ascribe mental states to others, and to understand the sorts of mental states others might have.

- **Technological intelligence**: communicating with other technologies; connecting to the internet and understanding information gathered

Roboethicists who outline the sorts of robots to which their work pertains often distinguish between robots based on their intelligence. Some philosophers distinguish between sentient and non-sentient robots (Chomanski 2019: 1009), whilst others distinguish between robots with social abilities and those without (Latikka, Turja, and Oksanen 2019: 157). Others separate robots according to their levels of autonomy (Sullins 2011: 234). Evidently, a robot's intelligence is relevant to roboethical discussions – for example, discussions of friendships with robots tend to focus on intelligent robots which can chat, understand emotions, and make jokes, but not on less intelligent robots which simply move objects around in a warehouse (Danaher 2019a, Rainey 2016, Newton 2008, Mulvey 2018).

We judge a robot's intelligence by observing what it can do: its behaviour demonstrates its intelligence.[27] It would thus be almost impossible to categorise robots according to their intelligence without reference to their behaviour. To clarify this: imagine two robots, R1 and R2: Both are equally intelligent and indeed sentient. R1 is able to understand its human users, and communicate with them, whereas R2 is able to understand its human users, but cannot communicate with them because it does not have a speech synthesiser, nor a screen on which to put messages. Both robots are equally intelligent, but R1 *demonstrates* its intelligence, and thus gives the impression of being more intelligent than R2. When an engineer makes a claim about the

[27] Though we should note that mere physical abilities such as moving its wheels quickly, video recording, and sensing its environment are not features of intelligence.

robot's intelligence (such as claiming that it is able to understand human speech), the burden of proof lies with the person making the claim, and thus it makes sense for us to judge the robot by its behaviour, rather than by calculating its computing power.[28] Thus, the intelligence axis tracks the robot's observable intelligent abilities.[29]

Some writers distinguish between weak (or narrow) AI, and strong (or general) AI. Weak AI refers to the ability to perform a specific task – such as playing chess – in an intelligent way. Strong AI refers to the ability to perform a wide range of tasks (such as playing board games, writing poetry, and understanding human behaviour) (Nyholm 2020: 9). I refer to robots with stronger (more general) AI as having higher intelligence, and those with weaker (narrower) AI as having lower intelligence. Some robots can only make very basic decisions; others can converse in multiple languages, and make complex decisions – sometimes life-and-death decisions. Some fictional robots (Data; C-3PO) have capacities for learning which far surpass human intelligence, and are sentient.

My matrix requires that we order robots along the horizontal axis according to their intelligence, whereby robots with the lowest intelligence are located on

[28] We can obtain a raw figure of computing power – often measured in calculations per second or floating-point operations per second (flops), but this does not necessarily elucidate a robot's intelligent abilities, such as sentience or social skills.

[29] Note that observable abilities are not the same as observed abilities. If all I ever do with my smart speaker is get it to play my favourite song, then I have not observed the full range of its intelligence – translation, mathematical ability, playing games, accessing encyclopedia articles, etc. Its intelligence level is what *can* be observed.

the left, and robots with higher intelligence are on the right, to get something like this:[30]

Lower intelligence → Higher intelligence

Sexbot	**Care-o-bot**	**HAL-9000**	**Data**	**C-3PO**
Recognises some phrases. Says numerous phrases. (Realdoll 2019)	Navigates through its environment. Understands some commands. Basic social functions. (Fraunhofer-Gesellschaft 2020)	Sentient. Social functions.[31] Speaks and understands natural language. Controls spaceship. Facial recognition. Other abilities. (Fandom 2022a)	Sentient. Fluent in many languages. Learns at incredible speed. Understands human emotions.[32] Understands sensory information as humans do. Many other advanced abilities. (CBS Studios 2022)	Sentient. Fluent in 6 million languages. Emotional. Understands sensory information as humans do. Many other advanced abilities. (Fandom 2022b)

Note that I am not claiming that robots with higher intelligence are *better* than robots with lower intelligence; I merely claim that it is possible to (roughly) order robots according to their intelligence along an axis, once one knows the robots' intelligence. Discovering a robot's intelligence is not restricted to first-hand observation – for example, we may need to read about the robot. In the case of fictional robots, we simply have to accept what we are told about the

[30] I have summarised some of the most important abilities of these robots, since listing all of them would be too cumbersome.

[31] In *2001: A Space Odyssey,* HAL-9000 is the antagonist, and kills some of the spaceship crew; at times he is decidedly *un*sociable. However, this is not because he lacks the *ability* to be sociable; rather, he has social functions, but *decides* to be unsociable.

[32] Occasionally, Data uses an emotion chip to experience emotions; however, for the vast majority of episodes, he does not experience emotions, but understands them (he is tactful, polite, kind, etc)

robot's intelligence-based abilities – for example, no one has observed C-3PO speak 6 million languages, but we are told he can do so (Fandom 2022b), so let us say that he can.

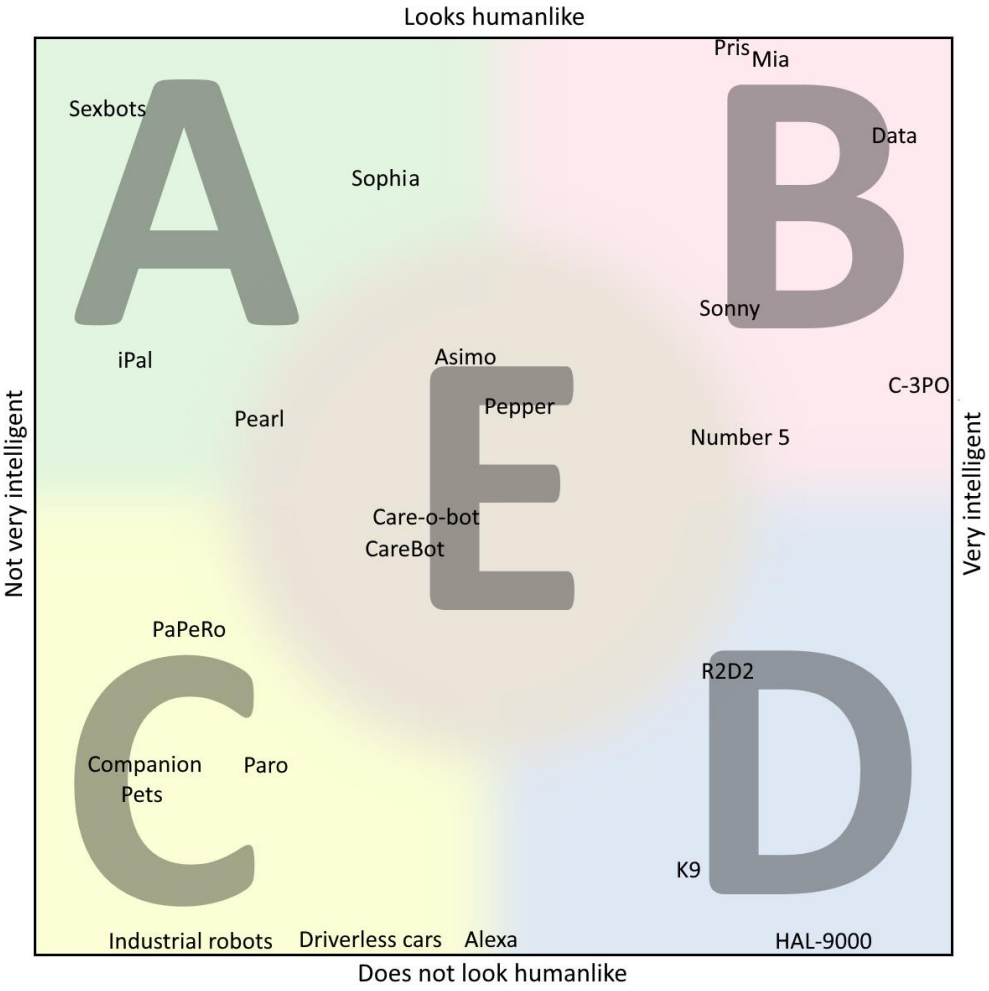
Assessing a robot's intelligence may be difficult, and there are many factors at play. One problem is that there are different ways to measure intelligence, and some aspects perhaps cannot be measured at all. Even human intelligence is not fully understood, and some writers suggest that intelligence is not a single ability, but rather that it consists of different aspects, such as mathematical ability, linguistic ability, musical ability, and interpersonal understanding (Gardner 2011), and these are not always at equal levels to one another. For example, some autistic savants can perfectly replicate any piece of music they have heard, but are unable to verbally communicate beyond the ability of a two-year-old. It is difficult to determine whether such a person has a higher level of intelligence than the average person, because clearly they are more advanced in some ways, but less advanced in others. The same may be true of some robots: suppose Robot A is not sentient, but has a broad general knowledge and is capable of learning, whereas Robot B is sentient and emotional, but only has the knowledge of a two-year-old, and struggles to learn anything new. Although Robot A has a greater *knowledge*, we may think that some aspects of intelligence – such as sentience and emotions – are so important that they trump other aspects of intelligence, thus Robot B should be placed further to the right of the intelligence axis than Robot A. Although there is not a way of measuring intelligence which is universally agreed upon, there is widespread agreement that different dimensions of intelligence can be distinguished from one another, and these can be

measured in different ways (though there is insufficient space to examine all of these herein).

With these difficulties in mind, ascertaining a robot's *precise* intelligence level can be difficult or impossible, but thankfully, the robot matrix does not demand a *precision* calculation of a robot's intelligence level. It is difficult to accurately measure animals' intelligence levels, but we are nonetheless able to understand 'family resemblances' and create an approximate ordering of intelligence: we know that humans, gorillas, dogs, bees, and slugs are ordered from most intelligent to least intelligent. Although it may be difficult to *precisely* determine the intelligence levels of the grey wolf, spotted hyena, coyote, and cheetah, we can nevertheless determine they have *similar* levels of intelligence – and that they are all less intelligent than gorillas, but more intelligent than bees. The same is true for robots: we can comprehend the approximate levels of intelligence they possess, which is enough for us to see 'family resemblances' on the robot matrix.

2.3 Robot matrix

Now that we understand both axes and how robots are ordered on them, we can put the two axes together. If one plots a range of robots onto the matrix in terms of their humanlikeness and intelligence, it would look something like this:



I have roughly divided the robot matrix into five different areas (robot Types A-E). There are not clear lines between one type of robot and another, and some robots (Pearl; Alexa; Number 5) are in an indeterminate location. Of course, one could divide the matrix into a greater number of areas, with a finer level of granularity; however, I believe that splitting the robot matrix into too many areas could prevent us from seeing the 'family resemblances' between robots which are similar, and more importantly (as I discuss shortly) have similar ethical issues pertaining to them. The robots, therefore, come in five Types, which I refer to throughout the thesis:

- Type A robots: Humanlike appearance, with low intelligence – e.g. sexbots
- Type B robots: Humanlike appearance, with high intelligence – e.g. Data
- Type C robots: Unhumanlike appearance, with low intelligence – e.g. industrial robots
- Type D robots: Unhumanlike appearance, with high intelligence – e.g. HAL-9000
- Type E robots: Loosely humanlike appearance, with mid-level intelligence – e.g. Pepper

It makes sense to have the central area for Type E robots, which exist mid-way along both axes, not only because they are paradigmatic robots which readily spring to mind when one envisions a robot, but also because some of our most advanced robots (at present) are Type E robots. Consequently, many roboethical discussions – including this thesis – focus on Type E robots, so it is worth distinguishing them from the other four types of robot.

If a robot which is identical to a human in its appearance and intelligence were plotted on the matrix, it would be at the very top of the vertical axis (because it looks completely humanlike) and perhaps three quarters of the way along the horizontal axis (it has quite a high level of intelligence, but not as high as Data and C-3PO). So, it would appear somewhere close to Pris. This is because no robot can appear more humanlike than a human,[33] but it could

[33] There are some conditions – congenital deformities, illnesses, and injuries – which cause someone to look 'unhumanlike'. Sadly, some robots may look more 'humanlike'

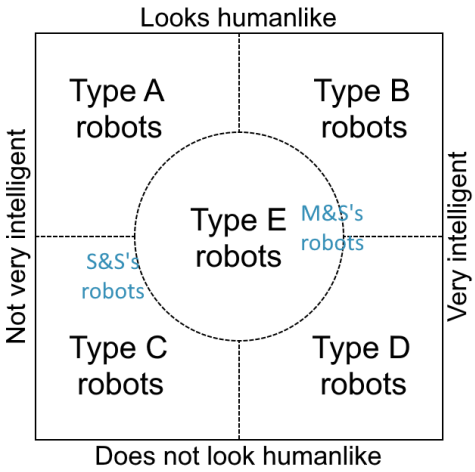
have higher intelligence than a human. All the robots on the right-hand side of the matrix are fictional robots: this is because we are yet to create a robot which has higher (general) intelligence than a human – although some robots out-perform humans in specific (narrow) ways, such as mathematics. Even though robots like Data and C-3PO do not (yet) exist in reality, it is still useful to discuss the ethical conundrums which they pose, and to include them within the robot matrix.

Robots' locations on the robot matrix represent their appearance and intelligence *at the time of writing,* and these may change as robots are developed further. For example, if driverless cars are given increased intelligence – say, conversational skills to chat with passengers, a broad general knowledge, and emotions, then they would be placed further to the right of the matrix, and we would call them Type D robots. Similarly, if the Care-o-bot was given hair and a flesh-coloured silicone face with moving lips, then the new Care-o-bot would be placed closer to the top of the matrix, and be a Type A robot. Developers may also decide to reduce a robot's intelligence, or to make it appear less humanlike, in which case, the robot will move left or downwards on the matrix, respectively.

Recall that my motivation for creating the robot matrix is to enable us to see how different roboethical discussions pertain to types of robot – but the matrix also helps clarify whether seemingly related roboethical discussions are indeed referring to the same type of robot. For example, I explained earlier that Meacham and Studley (2017) set up their argument in favour of carebots

than such people. My claims about looking 'humanlike' refer to the majority of human appearances.

as a response to Sparrow and Sparrow's (2006) criticisms of carebots. Although they both discuss carebots, scrutiny of their work reveals that Meacham and Studley's robots are not similar to Sparrow and Sparrow's robots. Using the robot matrix reveals that Meacham and Studley's (fictional or futuristic) robots which are highly social and look loosely humanlike are Type B/E robots, whereas Sparrow and Sparrow's robots from 2006, which do not look humanlike, and possess only basic intelligence, are Type C robots.[34]



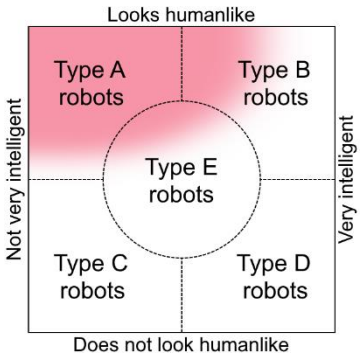
3 Ethical issues and the robot matrix

We have just seen that the robot matrix is a useful way of categorising robots to untangle roboethical discussions and ascertain whether different writers are in fact targeting the same robots. A further – and more substantial – benefit is that it enables us to establish links between ethical issues and different types of robot. Now that we have an understanding of which robots appear in which areas of the matrix, let us move on to discussing some roboethical issues, and how the robot matrix helps us see to which types of robot(s) the issues pertain.

An examination of the roboethics literature reveals numerous areas of interest. What follows is not an exhaustive list of ethical concerns, but it includes some of the most prominent roboethical discussions in the literature, and the types of robot most relevant to these concerns.

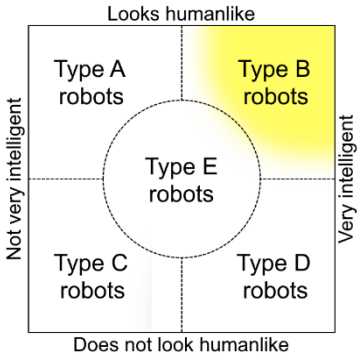
[34] The diagram shown here is a simplification of the robot matrix, and one should note that there are not *really* clear dividing lines between different types of robot.

- **Objectification of (human) women**: There are several concerns regarding sexbots which look quite humanlike. These include the possibility of sexbot users becoming less sensitive towards (human) women, sexbots exacerbating the objectification of (human) women, and that robot sex is vacuous and non-reciprocal (Lancaster 2021, Richardson 2016, 2019, Danaher 2019d, 2017a, 2017b, Harvey 2015, McArthur 2017).



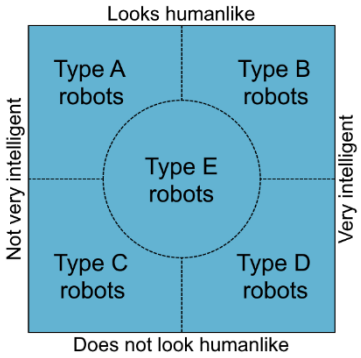
These sorts of concerns relate primarily to Type A robots and some Type B robots (shown in pink) which are capable of sexual activity. Concerns relate to the fact these robots physically resemble women, but have only very limited intelligence and can be treated however the user wishes.

Concerns about the welfare of sexbots *themselves* (Petersen 2017, Cappuccio, Peeters, and McDonald 2019) pertain more to Type B robots which are capable of consenting or suffering (shown in yellow).

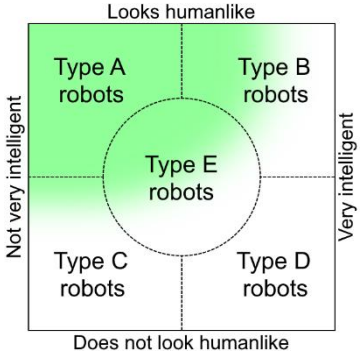


- **Loss of jobs**: Since the industrial revolution we have worried about machines displacing human workers, but robots may threaten more jobs today than ever before (Royakkers and van Est 2015b, Borenstein 2011, Smids, Nyholm, and Berkers 2020, Bellucci 2019), although robots could fill staffing gaps where there are insufficient numbers of human workers – such as nurses. Fears about loss of human jobs

relate to robots in all areas of the robot matrix (shown in blue), however the *types* of jobs under threat varies. Type A robots include sexbots, which may threaten the jobs of sex workers (Danaher 2014). Type C robots include industrial robots, weapons, and driverless cars; these may pose a threat to the jobs of factory workers, military personnel, and drivers respectively. Type D robots are highly intelligent but without humanoid form, and such robots replace clerical workers (lawyers, teachers, web designers, artists, administrative workers, and suchlike) (Sparkes 2023). Type E robots may threaten care work or service work (nurses, waiters, shop assistants, tour guides) (Meacham and Studley 2017, Sharkey and Sharkey 2012). Type B robots resemble humans physically but may have greater intelligence than humans – such robots may threaten most if not *all* types of human job!

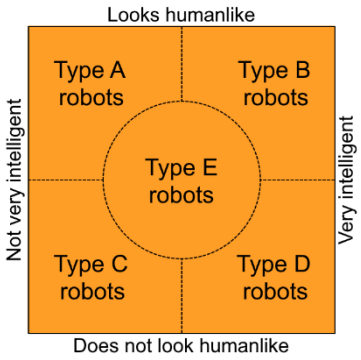


- **Anthropomorphic deception**: We have a tendency to anthropomorphise – particularly when things *already* look humanlike – which creates the possibility of being deceived (or deceiving ourselves) into seeing human-robot relationships as more reciprocal than they actually are (Leong and Selinger 2019, Eyssel, Kuchenbrandt, and Bobinger 2011, Fink 2012, Duffy 2003). These sorts of concerns relate primarily to Type A and Type E robots (shown in green). The concern dissipates with increasing intelligence, because advanced Type B robots may have the capacity to provide a fairly

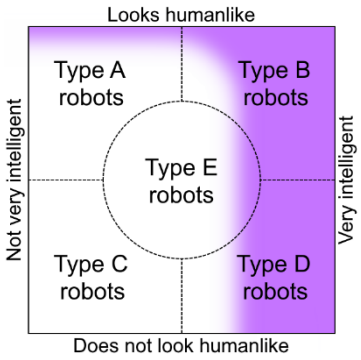


reciprocal relationship, so one is not anthropomorphising or fantasising if one perceives such robots as having abilities similar to a human. Anthropomorphic deception is explored in greater detail in Chapter 3.

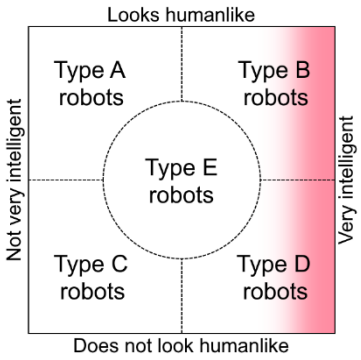
- **Trust and deception**: Trusting involves an element of vulnerability: if we trust people or robots which are not trustworthy, problems could arise. One danger with trusting robots is that we might be deceived or even betrayed by them (Grodzinsky, Miller, and Wolf 2015, Ferrario, Loi, and Viganò 2020, Danaher 2020, Shim and Arkin 2016, Isaac and Bridewell 2017, Matthias 2015). Ethical concerns about trusting robots permeate all areas of the robot matrix (shown in orange); however, the *type* of trust differs depending on the type of robot involved: one may trust a (Type A) sexbot to provide a satisfying experience, or we may trust driverless cars (Type C) to get us safely to our destination; such trust is domain-limited. Type B and Type D robots may be intelligent agents which we trust in a richer, deeper way, like the way we trust our parents – this would make any deception or untrustworthiness feel all the more hurtful (Danaher 2020). Robotic deception is explored in greater detail in Chapters 3 and 4.



- **Rights**: Discussions regarding whether robots should or do have rights (Steels and Brooks 2018, Coeckelbergh 2010) – including arguments regarding slavery (Chomanski 2019, 2020, Petersen 2011) – relate primarily to robots of Types B, D, and E which have advanced intelligence and possibly some form of self-awareness. These discussions hinge at least partly on whether robots can be moral patients (Danaher 2019b, 2019c, Cappuccio, Peeters, and McDonald 2019, Gerdes 2016). Not all rights-based discussions focus on highly intelligent robots, however. For example, concerns have been brought up in the literature regarding robots' sexual autonomy – whether sexbots (Type A robots) should be given the right and ability to decline sex, and whether having sex with a robot which has not given its consent amounts to rape (Steels and Brooks 2018, Eskens 2017, Frank and Nyholm 2017a, Coeckelbergh 2010, Sparrow 2017, Petersen 2017). Concerns about sexbot rape increase the further along the intelligence axis the robots are, since the likelihood of their being a moral patient increases. The types of robot about which rights-based arguments have been articulated are all shown in purple.

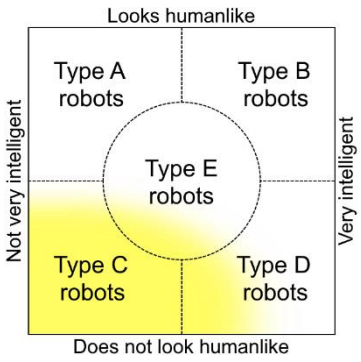


- **Emotions**: Discussions about apparent robot emotions and how we should respond to them (Fernández-Rodicio et al. 2022, Coeckelbergh 2014, Danaher 2019c) relate most clearly to highly advanced robots at the far right-hand side of the matrix (Types B and D – shown in pink). Currently, only fictional robots possess emotions, but this may change as technology develops. We may be more immediately concerned with the emotions of Type B robots (because they look human – see anthropomorphism, above), but concerns about Type D robots' emotions should not be overlooked. It might be possible for robots of lesser intelligence (e.g. Type A robots) to convincingly simulate emotions without actually *feeling* the emotions, and some roboethical discussions centre around how we should respond to such behaviour (Danaher 2019c); I explore this issue in Chapter 4.

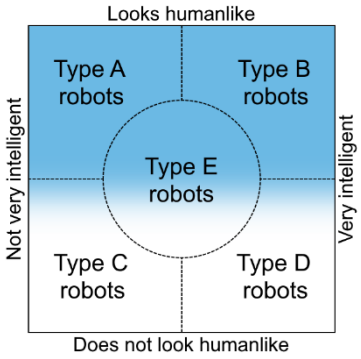


- **Deadly robots**: Developing (Type C) robot weapons for warfare ("killer robots") could save many (human) soldiers' lives – but such weapons could cause the deaths of many more victims than (human) soldiers would (Müller 2016, Sparrow 2007, Robillard 2018). Worries about robots causing death due to malfunctioning or making morally troubling decisions pertain both to robot weapons and driverless cars (Lorente 2020, Hansson, Belin, and Lundgren 2021), which are Type C robots and less advanced Type D robots. Of course, it would also be troubling if other types of robot killed people – such as a carebot which

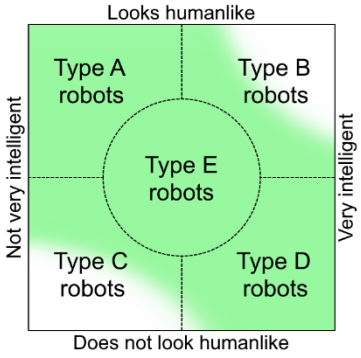
administers overdoses of medication, or if Data went on a killing spree – but concerns about deadly robots usually pertain to real-life weapons and driverless cars, and these are largely constrained to the lower left area of the matrix (shown in yellow).



- **Relationships with robots**: Some roboethicists focus on whether we can become friends with robots (Turkle 2017, Newton 2008, Rainey 2016, Mulvey 2018, Danaher 2019a). Going one step beyond mere friendship, some writers discuss whether we can have loving relationships with robots (Mamak 2022, Frank and Nyholm 2017b). Whether the discussion is regarding sexual relationships or platonic friendship, those suggesting we could have fulfilling relationships with robots focus on Type B (or advanced Type E) robots, whereas critics focus on the vacuous, unrequited relationships we have with Type A robots (the whole discussion area is shown in blue). It is also possible that we could develop some sort of friendship with a Type D robot, but discussions of friendship and relationships seldom seem to explore this possibility.



- **Loss of human contact**: Roboethicists consider whether using carebots could decrease human contact (Sharkey and Sharkey 2012, Stokes and Palmer 2020, Lancaster 2019, Meacham and Studley 2017, Borenstein and Pearson 2010). The sorts of robots under discussion are generally Type E robots, though loss of human contact could arise when robot pets (Type C robots) are used as surrogate social contact. Writers who present a more hopeful vision of meaningful care relationships with robots (Meacham and Studley 2017, Lancaster 2019) focus on robots with higher intelligence (advanced Type E, and Type B or even Type D robots), whereas those dubious of carebots (Sharkey 2014, Sharkey and Sharkey 2012, Sparrow 2002, Sparrow and Sparrow 2006) focus on robots with lower intelligence (basic Type E robots, or Type A or C robots). Some commentators are also concerned that users of sexbots (Type A robots) may withdraw from real human relationships (Harvey 2015, Richardson 2019). One might also be concerned that intelligent but unhumanlike Type D robots could hinder users' understanding of human body language and the pleasure of physical interaction with other humans. The whole area of discussion is shown in green.



My robot matrix enables us to group robots according to their intelligence and their humanlike (or not) appearance, and to distinguish five different types of robot. These groupings are useful because they enable us to more easily

grasp the roboethics literature, by understanding the sorts of robots to which different ethical issues pertain. The matrix also enables us to see links between different roboethical issues where perhaps we had not seen them before. For example, discussions about robot emotions and robot rights involve the same types of robots as each other (Types B and D), because advanced intelligence gives rise to these ethical discussions. My robot matrix also highlights that some ethical issues become more salient when robots look highly humanlike (such as concerns about objectification) or are highly intelligent (such as concerns about robot rights, or whether robots can be our friends). My matrix can help us see connections between ethical issues surrounding robots with different functions (such as concerns about job losses, which are relevant to discussions about sexbots, military robots, driverless vehicles, and industrial robots).

I noted that there is not always clarity about the sorts of robots which some writers discuss; I believe using my matrix could help eliminate confusion and ensure that roboethicists do not argue at cross purposes. It is important for roboethicists to be clear about which types of robot they are discussing – because concerns about robots replacing human contact may be well-founded if the robots are Type C robots, but the issue becomes less worrisome when the robots in question are sentient and emotional Type B robots.

4 Conclusion

Robots are increasingly present in our lives: on our roads, and in our shops, workplaces, hospitals, and homes, and this raises some fascinating issues which will be increasingly pressing for decades to come. The roboethics literature – although still in its infancy – is mushrooming, as writers discuss

new areas of interest and raise new concerns about the sorts of robots which humanity creates. However, there is not always clarity in the roboethics literature about the sorts of technologies under discussion, since philosophers' use of the term 'robot' varies, and some authors do not define it at all. My matrix should be intuitive to roboethicists – who already often comment on a robot's intelligence level and how humanlike (or unhumanlike) it looks. The matrix helps us not only elucidate the roboethical literature by clarifying the sorts of robots under discussion in particular works, but it can also show which roboethical issues pertain to which types of robot, and see connections between disparate areas of the burgeoning literature.

Where appropriate in this thesis, I use the robot matrix to help clarify the types of robot under discussion. Chapter 3 discusses robots of all varieties, but the rest of the thesis focuses primarily on carebots, which are usually – but not always – Type E robots.

Chapter 2

Deception: A conceptual and normative analysis

One of the concerns raised in the roboethics literature is that robots can be deceptive; for example, philosophers suggest that a robot could covertly record video of users, or that its humanlike appearance could trick users into thinking it has humanlike emotions (Leong and Selinger 2019, Kaminski et al. 2017, Danaher 2020, Sharkey and Sharkey 2020). Although some work has already been undertaken to distinguish different types of deception a robot may engage in (see Danaher 2020), this topic is still in its infancy, and one which I examine in Chapters 3 and 4. In this chapter, I take a step back from *robot* deception in particular, and provide conceptual and normative analyses of deception in general. These analyses underpin the next two chapters where I discuss whether (and how) robots are deceptive, and whether we should be morally troubled by this.

The chapter proceeds as follows: I begin by giving a conceptual analysis of lying and other-deception (where A deceives B), and I discuss whether other-deception is intentional or involves false beliefs, and the degree to which it

must be successful. I analyse the definitions of lying and other-deception, pointing out difficulties and inconsistencies, and ultimately arrive at workable definitions of both lying and other-deception. In §2 I focus on self-deception (where A deceives A), providing a conceptual analysis focusing on solving the problems of intending to deceive oneself, and holding contradictory beliefs. In §3, I provide a normative analysis of other-deception, followed by a brief normative analysis of self-deception in §4.

Lying is a phenomenon which has received a great deal of philosophical attention since at least as far back as ancient Greece; thus there is now a rich and extensive literature on lying (Isenberg 1973, Primoratz 1984, Kupfer 1982, Griffiths 2004, Tollefsen 2014, Bok 1978, Faulkner 2007, Gneezy, Kajackaite, and Sobel 2018, MacIntyre 1994). Although it is possible to deceive through non-verbal behaviour, there is not a dedicated literature on different types of non-lying deceptive behaviour – possibly because such behaviours are harder to categorise, and may overlap with other phenomena such as manipulation, dishonesty, and fraud. For now, I examine definitions of deception and lying, and differentiate between the two, before examining some other forms of potentially deceptive or misleading behaviour.

1 Lying and non-lying other-deception: A conceptual analysis

This section explores the concepts of other-deception[35] and lying. Lying is a particular form of linguistic dishonesty which is sometimes deceptive, and there are forms of deception which are not lies; I use the term ‘non-lying

[35] In the sections on other-deception (§1 and §3), I refer to other-deception simply as ‘deception’.

deception' to refer to any forms of deception that do not involve lying, and reserve the term 'deception' to encompass all forms of other-deception *including (successful) lying*.

Because of the similarities between lying and non-lying deception, I discuss both concurrently. Let us begin with a dictionary definition of deception: unsurprisingly, dictionaries tend to oversimplify the definition of deception. The Oxford English Dictionary defines deception as causing someone to believe something false (OED 1989). This seems to capture too much, however. For example, suppose Darren suddenly says "Careful!" to Nadia, and this causes Nadia to believe that there is a wasp nearby. However, there is no wasp; Darren was warning Nadia about a broken step. Darren caused Nadia to believe something false, however it does not seem apt to say he deceived Nadia. The definition in the Cambridge Dictionary (2021) gets closer to what it really means to deceive, by suggesting that deception is the act of hiding the truth, particularly when the agent is motivated by gaining some sort of advantage. However, this is still too broad, as there are some non-deceptive actions which fulfil this definition. For example, suppose Elliot and Freya are taking a maths test; Freya tries to copy the answers from Elliot, who is better at maths than she is. However, Elliot wants to win the prize for gaining the highest mark, so he hides his (correct) answers from Freya. Elliot has hidden the truth to gain an advantage, but it is intuitively evident that Elliot has not deceived anyone.

Conceptual analysis often provides more clarity than dictionary definitions of a term do, however there are no formally laid out conditions for deception which are universally accepted within the philosophical literature. I therefore draw

upon some necessary and sufficient conditions of lying in order to lay out some conditions for deception. The concept of lying has been studied extensively (Isenberg 1973, Marsili 2014, Carson 2006, Mannison 1969, Mahon 2016, Scott 2006), and although there is no definition of lying which is universally agreed upon, a widely accepted definition of lying is:

S is lying iff:

L1. S believes that p is false

L2. S states that p

L3. S intends to cause A to believe that p (Isenberg 1973: 248, Marsili 2014: 154, Primoratz 1984)

As I discuss shortly, there are several reasons why this is not a perfect definition of lying: the conditions are not necessary. These limitations notwithstanding, the above definition provides us with a useful starting point from which to examine the concept of lying more deeply, and I lay out some revised conditions in §1.5. Before I do that, however, it is useful to consider some possible necessary and sufficient conditions for deception. Using the format of the conditions for lying laid out above, I give this definition of deception (again, this will be revised after a deeper conceptual analysis):

S is deceiving A iff:

D1. S believes that p is false

D2. S takes some action φ (or omits to take action) which suggests that p

D3. S intends to cause A to believe that p

D4. S's φ-ing successfully causes A to believe that p

One immediately notices that I have given four conditions for deception whereas lying had only three: condition D4 has no analogue in the definition of lying. This is because deceiving is a success verb, and lying is not. If Derek is lying to Mary, he is lying whether or not he is believed by Mary. By contrast, it makes no sense to say that Derek has deceived Mary, but she did not believe him. If S's actions are unsuccessful in causing A to believe that p, then no deception has taken place (only *attempted* deception). In the next chapter when I come to examine robot deception, we see that this crucial requirement of deception seems to have been overlooked by many philosophers who claim that robots are deceptive (in other words, I point out that unless we are indeed fooled by robots, they have not deceived us).

I shall now discuss the different conditions and raise some examples which serve as test cases for the above conditions; at times I discuss lying and non-lying deception together, since some of the conditions are the same for both phenomena; other times I need to discuss lying separately from non-lying deception. After the discussion, in §1.5, I make a second attempt at refining the conditions for lying and deception. Although it may not be possible to yield perfect and unequivocal definitions, my clarifications at least enable us to have a better idea of what lying and deception consist of. From there, I can assess the normative status of both.

1.1 Condition 1: Belief that p is false

Both lying and deception have a condition 1 which states that S believes that p is false. The 'believes' part of condition 1 is important because it helps distinguish deception (and lying) from merely being mistaken. For example, suppose Beryl tells me that tomorrow is a bank holiday, but when I later check my calendar, I see that it is not. In such a situation, I cannot be sure whether Beryl has lied to me: if Beryl *believed* what she said then she has not lied; she was just mistaken – we can say that she was honest even though her statement was false.

We should notice that condition 1 only demands that S *believes* that p is false, not the stronger claim that S *knows* that p is false (although knowing that p is false would also be sufficient to meet this condition, since knowing involves believing). This dependence on mere belief means that both lying and deception can be consistent with stating the truth. For example, suppose Amir believes that the Post Office closes at 5.00 pm today. However, he wants to trick his colleague Sarah into missing the last post collection, so he tells her "Don't worry, the Post Office closes at 6.00 pm today". Even if the Post Office does in fact stay open until 6.00 pm today – so what Amir said was true – it is still the case that Amir has lied to Sarah (and if she believes him, she has been deceived), because he stated something he believed to be false with the intention that Sarah would believe it. This means that someone can be lied to or deceived, but end up with a true belief.

Let us consider a reason why *belief that p is false* is too strong a condition; namely, because some intuitively deceptive actions (and lies) fail to be captured by conditions D1 and L1. There are some propositions which I believe

to be neither true nor false, such as “The King of France is bald”. Some writers – such as Primoratz (1984: 54) – formulate the belief condition as *S believes that p is false*; others – such as Isenberg (1973: 248) – formulate the belief condition as *S does not believe that p*. The second is the more apt condition for lying and deception, since it helps to capture instances where someone states a proposition they believe to be neither true nor false, such as “The King of France is bald”. Although there is disagreement on whether a proposition whose subject term lacks a referent can be *false* (see Siegler 1966: 133–135, c.f. Strawson 1952: 173), saying that “The King of France is bald” and intending someone else to believe it does seem like a lie, and an attempt to deceive.

There can be additional cases where someone does not believe that not-p, but does not believe that p either. Someone can be agnostic about a proposition because of conflicting evidence, or simply because they have never thought about it. For example, you probably have no belief either way about whether my sister owns a cat. If you were now to tell your friend that my sister owns a cat, with the intention that your friend believes it, this would certainly seem to be a case of lying – and if your friend believes you, then it is an example of deception too. But you did not believe that “Karen’s sister owns a cat” is *false* – you had no belief either way about it. Because of these issues, condition 1 should be modified to become *S does not believe that p*.

Requiring falsity for conditions D1 and L1 would also prevent us from apprehending lying or deception in cases where the truth of a matter is not *known*. Consider: Ivy does not believe in Bigfoot, but she tells Stuart that Bigfoot exists, and Stuart believes her. There is some fact of the matter

regarding whether Bigfoot exists, and the prevailing consensus would seem to be that it does not; however, it has not been *proven* that Bigfoot does not exist (since proving the absence of something is notoriously difficult). So, we cannot say for sure whether Ivy states something false (and Stuart comes to believe something false). Nevertheless, it seems intuitively plausible to say that Ivy has lied to and deceived Stuart.

Additionally, there are times when we fall short of believing something is false, but we believe it is *probably* false. For example, consider the proposition "All the children in St Luke's Primary School were present today". I do not have sufficient evidence to believe this proposition is true or false, but I believe it is *probably* false, as it would be unusual for a school to have 100% attendance on any given day. Carson (2006: 298) suggests that if someone makes a statement they believe is probably false, intending that someone else believes it, this is sufficient to be called a lie.

Similarly, Marsili (2014: 155) suggests that lying can be a scalar phenomenon (and I suggest that the same is true of deception): a speaker can have a partial belief, or believe that a statement is only partly true. Some general claims subsume other more specific claims: the claim "The Prime Minister is doing a great job" is really many claims in one ("He is doing a great job with education"; "… with the economy"; "… with the NHS" etc). If Rachael believes some of these propositions but not others – perhaps she thinks the Prime Minister is doing a great job with education, but a poor job with the NHS and the economy – then we might say that Rachael only partly believes the overarching claim that "The Prime Minister is going a great job". If we accept that belief can be partial, then we also need to accept that a lie can be partial. If Rachael says

"The Prime Minister is doing a great job", while only partly believing it (or believing it is partly true), yet intending you to *fully* believe it, then it is not clear whether Rachael has lied: we might say that she has partly lied, or indirectly lied (Vincent and Castelfranchi 1981). This helps to capture the difference between 'big lies' and 'little lies': if Vivek believes X a little bit, but he does not believe Y at all, and he states both X and Y with the intention that you believe them, then his statement that Y seems to be a bigger lie than his statement that X, because it is further from Vivek's actual belief. Although it seems plausible that lying and deception can be scalar phenomena in this way (Marsili 2014: 155), this poses a problem for laying out necessary and sufficient conditions, and is not something I am able to resolve herein. Therefore, my second attempt at laying out the conditions for lying and deception will simply have *S does not believe that p* as the belief condition, but when we are using the definitions of lying and deception to make ethical appraisals, the magnitude of the belief (or deception) may affect its wrongness.

Interestingly, non-lying deception can be consistent with stating the truth, knowing that it is the truth. For example, suppose I know Peter is in room D59 in the Philosophy Building, but I want you to think I do not quite know where Peter is, so when you ask me where Peter is, I say "He is somewhere on campus". This is not a lie – because I know that what I am stating is true – but it is a case of non-lying deception, and what Mannison (1969: 132) refers to as "asserting too much". In non-lying deception cases such as this, what I am intending for you to believe is different from the proposition I say: I (truly) say that Peter is on campus, but I want you to (falsely) believe that I do not accurately know Peter's location.

Grice (1975) gives four principles of conversational cooperation: these are quantity, quality, relation, and manner. My statement that Peter is somewhere on campus is deliberately misleading, and violates Grice's maxim of quality, because it does not provide as much information as is required (Grice 1975: 45). A similar case involves "asserting too little" (Mannison 1969: 132). For example, suppose Pamela's latest car is a Porsche with a huge dent in its door, a cracked windscreen, and 100,000 miles on the clock. When Walter asks Pamela what her new car is like, she simply replies "It's a Porsche". Whether someone's action is a lie, a non-lying deception, or neither depends on what the content of p is. If p is *Pamela's car is a Porsche*, then Pamela has not lied or deceived – she has told the truth, knowing it is the truth, and she wants Walter to believe what she believes. However, if p is *Pamela's car is luxurious*, and this is what she intends Walter to believe, then this is still not a lie (she has not stated that p – she only stated that her car is a Porsche, which is true), but it is an attempt at non-lying deception (by omitting important information about the car's condition). Such a deceptive use of the truth is related to a type of deception known as paltering (Schauer and Zeckhauser 2009); a person palters when she attempts to distract her addressee with an irrelevant truth. For example, if Walter now asks Pamela what condition the Porsche is in, and she (truly) tells him that the car is only 2 years old, then she is paltering; she is attempting to distract Walter from the real answer to his question (that the car is a wreck). Paltering in this way is an attempt at non-

lying deception: where p is *Pamela's car is luxurious*, Pamela does not believe that p, but she takes action to suggest that p, intending A to believe that p.[36]

Because of the issues laid out in this sub-section, in my refined attempt at laying out the conditions for lying and deception, condition 1 will be changed to *S does not believe that p*.

1.2 Condition 2: Actions and speech

The second condition for lying and deception is stating something (in the case of lying) or taking some form of action φ (in the case of deceiving). Condition D2 is very broad, because one can deceive via any behaviour at all, like sitting on a chair, sending a photograph, or picking up a pen. For example, suppose Sadie wants Lawrence to (falsely) believe she is pregnant, so she sends him a photo of a positive pregnancy test (which is not hers) with a smiley face emoji. If Lawrence believes Sadie is pregnant, then she has deceived him without using any language at all. It is not possible to specify all the behaviours which may be deceptive, since there are too many of them and they are context-specific: the same action may be deceptive in some contexts but not others.

It is generally understood that lying necessarily involves making a statement using language, though it need not be speech;[37] one can lie through writing, sign language, morse code, semaphore and other ways to convey language

[36] A related phenomenon is bullshitting: a person bullshits when they claim that p while not caring about whether p is relevant to the question which was asked (Fallis and Stokke 2017: 288–289, Carson 2016: 61). Bullshitting can have similar effects to obfuscation, leaving the listener confused and bewildered.

[37] Although Austin (1962) suggests that 'speech' can involve non-verbal actions, I use 'speech' to mean spoken language.

(Mahon 2016: §1.1, Siegler 1966: 128). The majority of writers agree that lying involves making some form of statement, though a minority suggest that one can lie by refraining from giving information (Ekman 1985: 26–28, Scott 2006: 4), and others suggest that any behaviour which has the intention of giving false information is a lie (Smith 2004: 14, Vrij 2000). For example, if a manager says: “Put your hand up if you enjoy your job” and Frasier puts up his hand despite not enjoying his job, this would seem to meet Smith’s and Vrij’s definitions of lying.

However, I – like the majority of writers – find myself unconvinced that such non-linguistic behaviour really amounts to a *lie* (though it certainly seems a case of attempted deception). A borderline case, then, would be the nodding or shaking of the head to convey a yes/no answer respectively. If a parent asks their child if they completed all their homework and the child nods (despite knowing they have not completed all their homework) then this seems very similar to lying by saying “Yes”. I concur with standard accounts of lying and suggest that lying involves language, and that non-linguistic behaviour may be *deceptive*, but is not a form of lying (Siegler 1966: 128). Thus, nodding one’s head is not lying because it does not involve a statement using language – however, it is an attempt at non-lying deception. Whether we wish to class the child’s nod as a lie or attempted non-lying deception is perhaps not of huge importance if we would give the same normative appraisal of the child whether they are lying or attempting to engage in non-lying deception – and I suggest that in this context, a nod or a verbal “Yes” are normative equivalents.

Condition L2 (*S states that p*) and D2 (*S takes some action φ (or omits to take action) which suggests that p*) will for now remain unchanged in my refinement of the conditions for lying and deception.

The conditions for lying laid out above in §1 (Isenberg 1973: 248, Marsili 2014: 154, Primoratz 1984) do not specify that there needs to be an addressee, however this does seem important. If no such condition exists, then it would be possible to 'lie' to nobody at all. For example, suppose Caleb believes that he has psychic powers, and when he is home alone, he says "Penguins are mammals", not believing it, but intending Olivia to believe it – even though she cannot hear him, because she is elsewhere. Despite meeting all the (current) conditions for lying, this intuitively does not seem to be a lie, because nobody is around to hear it. In my refinement of the conditions for lying and deception, I will therefore change the second condition to include an addressee, thus: *S states that p to A* (Mahon 2016: §1), and analogously for deception, *S takes some action φ (or omits to take action) which suggests that p to A*.

That there should be an addressee of the lie or deception seems relatively uncontroversial; whether the lied-to or deceived person must necessarily be the addressee – or whether it can be an eavesdropper – is a more contentious issue. Consider this case: suppose you and I are planning to commit a burglary tonight, but we do not want our colleagues to know. We have a conversation within earshot of our colleague Bill, where you ask me to the pub for a drink tonight, and I say I cannot come as I am going to the cinema tonight. We are only talking *to* each other, but we are doing it with the intention that Bill hears and believes it, which he does. This meets all the conditions for lying and

deception which I laid out in §1, however some writers (Newey 1997: 115) suggest that such an example is not a lie, but rather, a 'bogus disclosure'.

A similar sort of case is if someone does not have a specific target in mind for their lie or deception: suppose Jeffrey wants everyone in the world to falsely believe he is a genius, and he posts a false high IQ test score on his blog, hoping that everyone in the world believes it. If Gretel – a woman on the other side of the world whom he does not know – stumbles upon Jeffrey's blog post and believes it, it does not seem altogether correct to say that Jeffrey has deceived or lied *to Gretel*. For these reasons it seems apt to conclude that for S to lie *to A*, he must be addressing his statement that p *to A* (thus I agree with Newey that you and I have not lied to Bill, since Bill was not the addressee); this will be incorporated into my revised conditions for lying below. For deception, it is necessary that my action φ must suggest that p *to A*, but it is not always clear, in practice, how someone can 'address' non-linguistic behaviour 'towards' someone.

1.3 Condition 3: Intending to cause A to believe that p

In the conditions for lying and deception I laid out in §1, condition L3/D3 is that S intends to cause A to believe that p; let us call this the intention to deceive.[38] Some philosophers (Demos 1960; Fuller 1976; Chisholm and Feehan 1977; Adler 1997; Gert 2005) suggest that non-lying deception can be inadvertent or mistaken, without the intention to deceive. Suppose that Sandra has left a copy

[38] Occasionally this is referred to as the "intention to mislead". Although having an intention *to mislead* as part of the definition of deception seems less tautological than an intention *to deceive*, the term "intention to deceive" is more commonly used in the deception literature, and is preserved herein.

of *War and Peace* on the table because she was using it to press some flowers. Robin sees the book on the table and forms the erroneous belief that Sandra is reading *War and Peace*. However, it seems to me that although Sandra has caused Robin to believe something Sandra does not believe, her lack of intention means that she has not deceived him. A widely accepted view – and the one with which I agree – is that it is not possible to deceive inadvertently or mistakenly, thus all deception (and lying) is intentional (Linsky 1970; van Horne 1981; Barnes 1997; Carson 2010; Saul 2012; Faulkner 2013). This means that deception and lying require an intention condition, such as L3/D3. Although some actions or statements lead others to have false beliefs, if there was no intention to do this, then it is merely a case of (unintentional) misleading[39] rather than deception (Carson 2010, 47). The distinction between deceiving someone and unintentionally misleading them is an important one which I return to in the next chapter when I suggest, roughly, that although some robots are misleading, they do not deceive users, since they have no *intention* to deceive.

The intention to deceive runs into other problems: for example, when someone lies whilst being fairly sure they will not be believed. Suppose Samira's school has a policy where someone is only punished if they admit their infraction. Samira is left alone in the classroom during detention, and writes her name on the table where she is sitting. When the teacher returns and asks Samira whether she wrote the graffiti, Samira says "No". She knows the teacher will not believe her; she lies simply to avoid punishment – this could be called a

[39] Misleading someone means causing them to have a false belief. Lying and deception do not always cause a false belief – so do not always mislead – but they involve an *intention* to mislead.

"bald-face lie" (Sorensen 2007, Fallis 2015). Samira cannot have the intention to deceive the teacher about her guilt, because it is arguably impossible to *intend* to do something one knows will not occur (see Kavka 1983, Levy 2009). This means that although what Samira is doing intuitively seems like lying, it does not meet condition L3 (or D3). One solution to this problem is to remove any reference to intention in the definitions of lying and deception; unfortunately, that sort of modification would capture too much, since people often make sarcastic comments ("There's a flying pig over there!" "I love meetings that could have been emails!") and these do not seem to be lies or attempts at deception.

An apt solution I suggest instead is to broaden the intention condition; instead of stipulating that S *intends* to cause A to believe that p, we can instead stipulate that A intends *or wants* this to occur. The use of *wants* helps to capture examples such as Samira (who *wants* to be believed even though she thinks she will almost certainly not be). This means that on my account, it is not lying when a speaker states that p while not believing that p, if they do so without intending or wanting to be believed.

1.4 Condition 4: Successfully causing A to believe that p

Lying is not a success verb: a lie is still a lie whether or not it is believed (Mannison 1969: 135). Deception *is* a success verb, however, and so the fourth condition for deception stipulates the success of the attempt to deceive. This means that this sub-section pertains to all forms of deception (which by definition must be successful) – but not to lying.

I have included reference to causation in condition 4 because without it, A's belief could be unrelated to S's attempted deception. Consider: if *A comes to believe that p* was the fourth condition, some examples could meet that condition without intuitively seeming like an example of deception. For example, suppose Franklyn wants Merina to (falsely) believe that worms are snakes, so Franklyn says to Merina: "Worms are snakes". Merina knows Franklyn is a habitual liar, and ignores what he has said. Some time later, however, Merina sees an article online about the ways in which snakes and worms move, and she thus forms the (false) belief that worms are snakes. In such a case, we would not want to say that Franklyn has deceived Merina, because she did not believe him: Franklyn's statement was not the *cause* of Merina's ultimate belief. Thus, having *S's φ-ing successfully causes A to believe that p* as the success condition (D4) is preferable to merely stipulating that *A comes to believe that p* (but not specifying how).[40] It also makes sense to stipulate a causal connection within condition 3: because it is not simply that *S wants A to believe that p*; rather, *S wants their suggestion that p to cause A to believe that p*.

Let us press further on the causal connection stipulation: Franklyn also tells Yvonne that worms are snakes, wanting her to believe him, and although Yvonne does not immediately and fully believe what Franklyn says (she too

[40] There can be cases of 'deviant causal chains'. E.g. suppose that when Franklyn said "Worms are snakes", Merina thought he said "Look on Facebook" which causes Merina to open Facebook and (by chance) see the article on worms and snakes, causing her to believe that worms are snakes. Franklyn was the beginning of a causal chain which led to Merina believing that worms are snakes, but Franklyn did not cause Merina's belief *in the right sort of way* for us to conclude he deceived her.

knows he is a habitual liar) she does not forget about it nor dismiss it out of hand. Instead, she seeks verification: she looks up an article about the way worms and snakes move, and this article *in addition to Franklyn's claim* causes Yvonne to believe that worms are snakes. In such a case, S stating that p is one (insufficient) part of what causes A to believe that p. Do we want to conclude that S has deceived A?

I suggest that the answer to this question is that S has *partly* deceived A – meaning that deception can come in degrees. Consider a further example: Miss Cook leaves Class 11 for a minute, and when she returns, the plant by the window has been knocked onto the floor. She asks the class "Did the wind blow the plant over?" and all the children in Class 11 (falsely) say "Yes miss!" and Miss Cook believes them. All thirty of the children in Class 11 have played a small part in deceiving Miss Cook.

Having the success verb requirement as part of the definition of deception also raises borderline cases when the (ostensible) deception is only partly successful. When a speaker states that p while not believing it, it may not cause an outright *belief* that p in the listener, but it could cause the listener to shift their belief in the *probability* that p. For example, suppose Gianni tells Kylie that drinking daily milkshakes has cured his liver cirrhosis (Gianni does not believe it, but wants Kylie to believe it). Kylie does not wholly believe him, but she begins to consider the *possibility* that milkshakes can cure liver cirrhosis (something which she had never considered before). In borderline cases such as this, Gianni's (attempted) deception seems only to be *somewhat* successful, which raises the question of the extent to which attempted deception must be successful for it to count as (actual) deception.

If Gianni's claim only causes Kylie to have a 1% belief in the power of milkshakes to cure liver cirrhosis, this does not seem to meet condition D4, so does not count as deception, but if it causes her to have a 99% belief in his claim, this *would* seem to meet D4, and count as deception. Somewhere between these extremes is a vague zone where it may not be entirely clear whether an attempted deception was successful enough to meet condition D4 (especially since beliefs cannot be classified in percentage terms!) More conceptual analysis would need to be done in order to adequately answer this question, and so I shall merely change the success condition D4 to say that *S must be sufficiently successful in causing A to believe that p* – and concede that what counts as *sufficient* success would need to be explored elsewhere.

1.5 Necessary and sufficient conditions – refined

Now that I have engaged in some conceptual analysis (which I do not claim to be exhaustive), I offer my updated necessary and sufficient conditions for lying and deception. These may not *perfectly* capture all and only instances of lying and deception, but they should be adequate for my purposes here and in the next chapter, when I discuss whether and how robots deceive.

S is lying to A iff:

L1. S does not believe that p

L2. S states that p to A

L3. S intends or wants S's statement that p to cause A to believe that p

S is deceiving A iff:

D1. S does not believe that p

D2. S takes some action φ (or omits to take action) which suggests that p to A

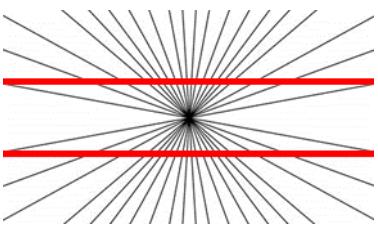
D3. S intends or wants S's suggestion that p to cause A to believe that p

D4. S's φ-ing is sufficiently successful in causing A to believe that p

1.6 Inaccurate uses of 'deception'

Within philosophy – in particular when discussing deception in epistemology and ethics – we are careful about our use of the words 'deceive' and 'lie'. As one might expect, out of the classroom, people do not always use 'deceive' and 'lie' in the ways laid out above. In this sub-section, I discuss some instances of misleading which are sometimes (inaccurately) referred to as deception.

The term 'mislead' means leading someone to a false belief;[41] misleading can be done by a human – as all the examples hitherto are – but it is not a uniquely human phenomenon. One can also be misled by animals and even inanimate objects. For example, optical illusions can mislead us: the illusion on the right makes it appear as though the two red lines are curved, when in fact they are straight. Although people



[41] Lying and deception necessarily involve the *intention* to cause a false belief, but might not involve *actually* causing a false belief – such as in the Post Office case.

might say that optical illusions are deceptive, this is an inaccurate figure of speech: the optical illusion does not meet condition D3; it is nevertheless true that optical illusions mislead us. No agency, sentience, desires, intentions (etc) are required for misleading, meaning that there is no problem with suggesting that optical illusions, spiders, swimming pools, space, dresses (etc) can mislead people. In the next chapter I make the case that some robots mislead us but do not deceive us – though roboticists may use robots as tools to deceive users.

Misleading is common in nature. Skyrms (2010: 73–74) cites an example of a low-ranking vervet monkey (‘Kitui’) which would give ‘leopard’ alarm calls whenever a new male monkey attempted to transfer into his group; the new male would run up a tree in fear, and ultimately not transfer into Kitui’s group, reducing Kitui’s competition for mates. Animals can also mislead via their colouring: the *Caligo* genus of butterflies have “eye spots” on their wings which draw a predator’s attention away from the butterfly’s head, making an attack less likely and/or less likely to be fatal (Osterloff 2021). Even some plants mislead by emitting pheromones to attract particular insects to ‘mate’ with it; the plant is cross-pollinated, but the insect receives no benefit at all (Vereecken 2009). Further examples from nature can be found in abundance: snakes which appear to have a head at each end, reducing the chance of a fatal attack by half (Phys.org 2009); insects which resemble leaves (Stevens 2016); birds whose eggs resemble the eggs of other species and are sneaked into their nests (Langley 2017); all forms of camouflage patterning… the list could go on and on. Although authors writing for a popular audience may describe such phenomena as ‘deception’ (Phys.org 2009, Stevens 2016,

Vereecken 2009) or 'lying' (Langley 2017) – and sometimes even those writing for a scientific audience make the same claims (Sober 1994: 71–92, Hauser 1997, Searcy and Nowicki 2005) – it is important to be clear that deception and lying only occur when all the above conditions are met – and it is improbable that any of these examples from nature are sophisticated enough to fulfil all the conditions for lying or deception (in particular, condition 3 – intending or wanting someone else to believe a proposition).

Some (non-philosophical) writers suggest a broad definition of deception which encompasses such examples in nature – for example, that deception involves sending misinformation which benefits the signaller but not the receiver (Skyrms 2010: 75, Searcy and Nowicki 2005: 5). Such a definition would capture too much among humans, however. For example, suppose Jared looks like a particular celebrity; although he does not try to mislead anyone (he does not dress like his famous doppelganger, and he denies that he is the celebrity), he nevertheless receives preferential treatment when he goes out. According to the 'nature' definition of deception, Jared is deceiving people, since he sends out a visual signal (his looks) which benefits him, but not others. Intuitively, however, it seems clear that Jared is not deceiving anyone. Given that those who write for a popular or even a scientific audience are perhaps not privy to the sort of conceptual analysis detailed above regarding the definitions of lying and deception, I suggest that although people may use the terms 'lying' and 'deception' to refer to these natural phenomena,

their use of the word is inaccurate: animals and plants may mislead others, but they do not lie or deceive.[42]

The above conceptual analysis of lying and deception is useful to my project in the next chapter, and indeed to the normative analysis of deception in §3. Next, however, I discuss self-deception.

2 Self-deception: A conceptual analysis

In this section, I provide a conceptual analysis of self-deception, which gives important background information for the normative analysis of self-deception in §4. Although my central focus in the next two chapters is other-deception[43] (rather than self-deception), I suggest that sometimes when people interact with robots, they self-deceive. Some of the issues which are pertinent for other-deception (such as the truth of propositions, partial success, and deviant causal chains) also relate to self-deception, but I shall not re-articulate these issues; instead, I focus solely on issues which have not yet been discussed because they are only applicable to self-deception.

There is some overlap between self-deception and related phenomena, including wishful thinking, self-inflation bias (believing oneself to be better than one is), impostor syndrome, delusion, wilful ignorance, epistemic laziness, cherry-picking evidence, bad faith, and suchlike (Leeuwen 2013: 4–5). Self-

[42] If it turns out that some natural examples can and do fulfil all my conditions for lying or deceiving, including intending or wanting someone else to believe a proposition (L3/D3), and the target sufficiently believing it (D4), then I would happily concede that such animals *do* deceive and lie, but that the definitions of lying and deceiving laid out in §1.5 remain unchanged.

[43] Within this section, I use *other-deception* to refer to cases of deceiving another person.

deception, like some of its cognates, has fuzzy boundaries: it may not always be clear when self-deception is occurring. For example, if Yusuf believes he has excellent general knowledge, but then performs poorly on a history quiz, he may still maintain that he has excellent general knowledge, but tell himself that he just got unlucky with those questions, or that the quiz was aimed at specialists. At this point, Yusuf is perhaps not self-deceiving. However, if Yusuf also performs poorly on a sports quiz, *and* a pop music quiz, *and* a geography quiz, *and* a TV quiz – and he *still* maintains his belief that he has excellent general knowledge, then at some point he has slipped into self-deception, though it may not be altogether clear when that happened (Leeuwen 2013: 5).

Self-deception needs to be distinguished from mere mistakes or foolishness. For example, if Tara believes that dolphins are fish, this does not necessarily mean that Tara is deceiving herself – she may simply be poorly informed.[44] Modelling self-deception on other-deception can be a useful way to obtain a definition which distinguishes self-deception from mere mistakes. Other-deception involves two agents: the deceiver (who does not believe that p) and the deceived (who comes to believe that p as a result of the deceiver's action). Following this format, self-deception is like other-deception, but with a single person filling the roles of both the deceiver and the deceived. Gibson (2020: 657) gives this definition of self-deception: S is self-deceived that p iff

[44] Dolphins' taxonomic classification seems an odd thing to deceive oneself about anyway: self-deception is most common in matters relating directly to the self-deceiver – their family, job, personal qualities, etc. – however, one could theoretically self-deceive about anything.

S believes that not-p, and S intentionally causes herself to believe that p.[45] So in order for Tara to be self-deceiving, she would need to believe that dolphins are *not* fish, but intentionally cause herself to believe that they *are* fish. Immediately, this makes self-deception seem inconsistent, and raises two distinct problems which need to be solved if we are to properly explain self-deception:

(1) How is it possible for an agent to hold contradictory beliefs (p and not-p)? (The "static paradox")

(2) How can an agent *intend* to deceive himself, but be sufficiently unaware of his intention for the deception to succeed? (The "dynamic paradox")

Both problems present us with some form of contradiction. Problem 1 involves two contradictory beliefs about the truth of p, held by the same person. Problem 2 arises because the agent must be *aware* of their intention to deceive in order to take action, but they must also be *unaware* of their intention to deceive in order for the deception to work, since knowing you are about to be deceived generally renders the attempted deception a failure (Gibson 2020: 658, Deweese-Boyd 2016: §2). Some philosophers conclude that self-deception is logically incoherent (Haight 1980), and that in order to escape these problems, we must alter our conceptions of either 'self' or 'deception' (Borge 2003). More common responses, however, focus on making self-

[45] Although Gibson's definition uses terms (belief that not-p; intentions) which I discussed and softened above, these terms are oft-cited in the self-deception literature, and so I leave them unchanged here.

deception coherent by solving the paradoxes. These attempted solutions are discussed presently.

First let us focus on problem 1 – the issue of an agent holding contradictory beliefs. One proposed solution to this problem is to dispense with or deflate one of the belief conditions and maintain that self-deception does not require two contradictory *beliefs*, but rather, one belief and one non-doxastic attitude. Some writers focus on the second[46] belief condition and suggest that S ends up only pretending that p (Gendler 2007) or imagining that p (Lazar 1999). Although this might explain some cases of apparent self-deception, it implies that self-deception proper does not occur – only the *appearance* of self-deception. Although it is not possible to empirically verify this claim, it does not seem plausible: some cases of actual self-deception probably *do* occur, and so we must explain how. A slightly more satisfying solution to problem 1 is to alter the initial belief condition, and replace it with something like suspicion (Edwards 2013) such that S is self-deceiving iff S suspects that not-p, and S intentionally causes herself to believe that p. Some writers do not specify the initial mental state, and suggest that self-deception only requires that one intentionally brings about in themselves a false[47] belief that p, motivated by the desire for p to be true (Gibson 2020: 660). One could also have an initial mental state of agnosticism: suppose Ricardo has no belief either way about

[46] By 'second belief' I mean the belief which the agent has deceived themselves into having. This belief may occur at the same time as the 'initial belief' – the term I use to refer to the other – original – belief.

[47] I write 'false' here because that is Gibson's choice of word. It would be more accurate, though cumbersome, to say one intentionally brings about in themselves a belief that p which they initially did not believe. This is because deception – including self-deception – can result in a true belief.

p – he recognises the evidence for p, and the evidence against p – then takes action to intentionally cause himself to believe that p. Ricardo's activity is arguably sufficient to be called self-deception, but did not involve a belief, suspicion or doubt at the outset (Bermúdez 2000). This solution seems plausible, and it parallels the belief condition (L1/D1) I drew out above for other-deception.

I now move on to some potential solutions to problem 2 – the question of how an agent can be sufficiently aware of their intention to deceive so as to act on it, but also sufficiently unaware of the intention so as to be fooled by it. Satisfyingly, some of the proposed solutions to problem 2 (such as temporal and psychological partitioning) also offer solutions to problem 1; I point these out as appropriate.

There are two schools of thought regarding whether self-deception necessarily requires the intention to deceive: anti-intentionalist (or non-intentionalist) approaches attempt to solve problem 2 by suggesting that self-deception is not intentional (Bermúdez 2000: 309–310). Intentionalist approaches, by contrast, suggest that self-deception is intentional; these philosophers must therefore explain how it is possible for an agent to carry out the self-deception, given that this seems to require both being aware of and unaware of the intention. First, I deal with anti-intentionalist approaches.

Anti-intentionalists suggest that people may not necessarily intend to self-deceive. For example, one might self-deceive because they have the *desire* to acquire the belief that p (Mele 2001). This does not seem to solve the problem, however, for we can still wonder how one can desire the belief that p enough to take action, but be sufficiently unaware of the desire for the self-deception

to succeed. Neither desire nor intentions (nor similar terms such as hope, want, wish for) can explain cases of so-called “twisted self-deception” (Mele 1999, Echano 2017, Galeotti 2016). In twisted self-deception cases, the agent ends up with a belief that p *which they did not want to have*. For example, suppose that Wendy does not want her neighbour to be a terrorist, but she convinces herself that he is actually a terrorist. Wendy did not desire this belief, nor intend to acquire it. Motivationalist accounts of self-deception suggest that in cases like this, one is *motivated* to believe that p, but that the motivation may be fear, anxiety, and suchlike (Barnes 1997, Johnston 1995). Cases such as Wendy’s may not be *clear* cases of self-deception, but rather, some related phenomenon such as paranoia, panic, delusion, or some other mental health issue – or something more mundane such as hasty belief-formation or inattention to evidence.

Although twisted self-deception may occur in some cases, garden-variety self-deception typically *does* end in the agent holding the belief they set out to hold; it therefore seems reasonable to maintain that self-deception, generally speaking, is intentional (or desired). Here we turn to intentionalist approaches to addressing problem 2. The crux of problem 2 is that “for one to carry out an intention to deceive oneself, one must know what one is doing, [and] to succeed one must be ignorant of this same fact” (Deweese-Boyd 2016: §2). Intentionalists about self-deception often suggest that the agent engages in some sort of ‘partitioning’ – whether temporal or psychological (Davidson 2004, Johnston 1995, Sorensen 2007, Gibson 2020).

First let us discuss temporal partitioning, by considering an example: Meghan writes in her planner app on 1 January that her new year’s resolution is to only

eat takeaways at weekends. She begins well, but by February, Meghan has slipped back into her old ways and is eating takeaways several nights a week. Feeling disappointed in herself, Meghan resolves to deceive herself into believing that she has kept her resolution. She deletes the takeaway resolution, and replaces it with something she knows she will find it far easier to keep – say, to get up before 10 am on weekends. She knows that several months down the line, she will probably forget she ever made the takeaway resolution, see the getting up resolution in her planner app, and falsely believe she has kept the new year's resolution she made on 1 January. This temporal partitioning solves both problems 1 and 2. It solves problem 1 because Meghan never holds contradictory beliefs: in January and February, she did not believe that p (where p is "I kept my original resolution"); her belief that p occurs in December, ten months later. In other words, S can deceive herself without contradiction by not believing that p at T_1, then a sufficient time later, believing that p at T_2. This relies on the natural degrading of one's memory, rather than any mental effort to bring about the belief that p (Johnston 1995). Temporal partitioning also solves problem 2, because Meghan's intention to deceive herself occurs in February, but the completion of the deception occurs the following December, so at no point was she *both* aware she was trying to deceive herself *and* unaware she was being deceived.

Although temporal partitioning nicely solves both problems 1 and 2, it is perhaps not a paradigmatic case of self-deception; conventional self-deception would seem to involve the second belief being held at the same time as – or at least, very soon after – the initial belief. For an intentionalist account of this sort of self-deception, a 'psychological partitioning' approach can be

utilised. Sometimes, psychological partitioning is described with reference to dual processing theory (De Neys 2017, Gibson 2020: 661–663), where different systems within the mind – S1 and S2 – hold the opposing beliefs, or Freudian theory (Davidson 2004), where one part of the mind deceives another, or it is suggested that the initial belief is only held subconsciously, and the second belief is held consciously (Deweese-Boyd 2016: §2.2). Having the two beliefs in different 'parts' of the mind in this way helps to limit any cognitive dissonance in the agent, and (ostensibly) explains how a single person can hold two opposing beliefs.

The problem, however, is that the initial belief must be epistemically accessible to the agent, because it motivates the intention to deceive (Davidson 2004). For example, if Emeline wishes to deceive herself into believing she does not have an alcohol problem, she must in some way be aware that she *does* have an alcohol problem, otherwise she would not wish to self-deceive. It seems plausible, however, that one can be motivated by something which they are only subconsciously aware of, and such a theory would perhaps be the best explanation for self-deception. In other words, S is self-deceiving iff S subconsciously believes that p is not true, and is subconsciously motivated to cause herself to believe that p is true – and succeeds in doing so.

Some forms of self-deception may require ongoing mental effort; this is pertinent because if (as I suggest in § 4) self-deception is an epistemic vice, then self-deception which involves ongoing effort is more vicious than self-deception which does not, since the former essentially involves *repeated* self-deceptions rather than a one-off instance. When temporal partitioning is used (such as Meghan self-deceiving about her new year's resolution) no ongoing

effort is required; one can simply allow natural memory degradation to do its work. However, in cases of psychological partitioning, where opposing beliefs are held simultaneously, some sort of effort to continue the self-deception may be necessary (Bach 1981, Johnston 1995). This may be because the initial belief resurfaces: Roger keeps thinking back to the time he nearly drove into a pedestrian, wondering if it was his fault, and repeatedly self-deceives to reassure himself he is a safe driver. Alternatively, repeated self-deception may be necessary because additional evidence presents itself: each week when Emeline takes out the recycling and sees all the bottles of gin in it, she must perform some mental gymnastics if she is to maintain her belief that she does not have an alcohol problem. She may also need to deceive herself every evening she drinks gin, when she buys the gin, when family members express concern about her gin-drinking, and so on. This kind of self-deception may require such extreme and repetitive mental effort that the agent will ultimately succumb to the initial belief: Emeline may admit she has an alcohol problem. This may not happen, however: an agent may continue their self-deception indefinitely.

This section has provided a brief conceptual analysis of self-deception. It is not imperative that a complete analysis is provided here, since self-deception is not the primary focus of my next two chapters. I will, however, suggest that when people interact with robots, they could deceive themselves into believing the robot has emotions, sentience, or suchlike; if this occurs, then users could be at least partially responsible. Even if we do not have a perfect understanding of the *process* of self-deception (whether psychological or temporal partitioning occur; whether agents hold one non-doxastic state;

whether it is intentional or not) it is sufficient for us to know that self-deception can occur.

3 Lying and other-deception: A normative analysis

This section provides an analysis of the normative status of lying and other-deception.[48] I explore why they are so often seen as morally problematic, and discuss occasions when they are morally permissible or even morally good.

That people are speaking truthfully is a baseline assumption and convention of everyday conversations: we do not generally preface our statements with "It's true that…" – rather, truth is assumed when someone states something. One of Grice's (1975) four principles of conversational cooperation is quality: Grice suggests that we should not say things for which we have inadequate evidence (Grice 1975: 46), so if Adele tells Terrence that her car is broken (a lie, for which she has no evidence) then she has broken Grice's maxim of quality. However, when someone lies to us or attempts to deceive us, we are not simply perplexed that they have broken a conversational or behavioural norm: the wrong of lying and deception runs deeper than that. After all, we can break some of Grice's maxims without much – if any – moral condemnation. For example, if I ask Francesca the way to the station, and she gives me directions *and* tells me the Latin names for the trees I will see on the way, then she has broken Grice's maxims relating to the quantity of information, and the relevance of information (1975: 46), but we would not see her actions as morally troubling.

[48] In this section, I refer to other-deception simply as 'deception'.

Some plausible reasons why we see lying as morally problematic are:

- Lying can bring about negative consequences
- Lying is wrong in itself
- Lying results from a negative character trait

The first – consequentialist – viewpoint would not seem to fully capture the wrong of lying. For example, when we morally assess people's actions, we would surely want to condemn Amir (who sets out to deceive Sarah about the Post Office opening times, but accidentally gives her true information) more than poorly-informed Yvonne, who innocently (but falsely) tells Poppy that worms are snakes. Yet Sarah ends up with true information (a good consequence) and Poppy ends up with false information (a bad consequence). So although lying can sometimes yield bad consequences, this does not adequately explain the wrong of lying. It seems intuitive to say that good people should not lie (Fried 1978: 54): for this reason, much of the discussion of the wrongness of lying focuses more keenly on the *intention* to cause false beliefs, rather than on whether the listener ends up with false beliefs.

Generally, there is a moral presumption against lying (Bok 1978), because successful lies carry what we might call a "negative weight" " (Kupfer 1982: 105); there is a (defeasible) moral duty not to lie. Kant famously wrote that lying is wrong in itself, and we should never lie (Kant 1996a: 430) – even to protect innocent family members from would-be killers (c.f. Bok (1978: 14), who suggests that speaking falsely to people who do not deserve truth cannot be called a lie). This moral stance is shared by Augustine (Griffiths 2004: 230) and seen in some theological approaches (see Tollefsen 2014). However, even Kant did not maintain that we should never *deceive*. Kant himself actually

deceived (but did not lie to) King Friedrich Wilhelm II of Prussia. Kant said to the King – who was close to death – that as his faithful subject, he would cease all his work on religion. The King happily took this to mean Kant would never publish on religion again, but Kant simply meant that as long as the King was *alive*, he would not publish on religion. Once the King died, and Kant was no longer 'his faithful subject', Kant resumed his work on religion, while staying true to the letter of what he said (MacIntyre 1994: 336–337).

It is not altogether clear why Kant (and Augustine (Griffiths 2004: 32)) felt that lying was anathema, but deception was sometimes permissible. We might conjecture that it is because the liar is deliberately duplicitous and clearly articulates something they do not believe. However, in cases of non-lying deception, although the agent is still being duplicitous, she merely *suggests* a falsehood, leaving the listener some room to interpret things in their own way. There is a prevailing opinion that generally speaking, non-lying deception is not as morally problematic as lying is;[49] this opinion often rests on the idea that with non-lying deception, the listener is partly responsible for their being deceived, and the deceiver is less responsible than a liar is (Chisholm and Feehan 1977, Adler 1997: 444). In Kant's case, one might think that if King Wilhelm interpreted Kant's words in a way they were not intended (but were nevertheless implied) then that was at least partly the King's fault. Later, when I discuss the normative status of self-deception, I refer to epistemic vices – these are negative character traits such as cherry-picking evidence, lazy evidence-gathering, inattention to detail, trusting unreliable sources, and jumping to conclusions. Sometimes, victims of other-deception are themselves

[49] Not everyone agrees: Saul (2012: 5) suggests that both can be as bad as each other.

guilty of such epistemic vices. Perhaps King Wilhelm should have pressed Kant further on exactly what he meant, and whether he promised never to write about religion again. The same may be true for some of the other examples I have described above. Perhaps when Pamela told Walter that her car is a Porsche and is only two years old, he should have questioned her further on the condition of the car, rather than assuming it was in good condition. In other words, liars bear full responsibility for their lies, and the victims of (plausible) lies bear none; however, non-lying deceivers share the responsibility with the victims of the deception – the victims are to some extent complicit in the deception. Moreover, it seems more important that we are truthful in what we say than in what we imply (Adler 1997: 451), perhaps because statements are clearer than implicatures. In the next chapter when I discuss whether robots deceive, I return to this issue and suggest that perhaps human users jump to conclusions and 'fill in the blanks' when interacting with robots.

Lies come in different varieties. Some lies seem morally worse than others because of what is at stake: suppose that while her son is playing happily, Constanza lies to me by saying her son has never ingested a poison; this is inconsequential. But if Constanza tells the same lie to paramedics while her son is having convulsions a few minutes after having ingested the poison, this is very different indeed, because the stakes are so much higher. Some lies do not seem particularly bad at all: all things considered, it seems that lying is sometimes permissible and perhaps even morally good. For example, if Gina invites Nigel to hear her (musically inept) daughter play the violin, Nigel might lie and say he is unfortunately busy that evening, to spare Gina's feelings. This is an example of a prosocial lie, which generally receives a positive moral

evaluation, since it supports the smooth-running of social relations (Strudler 2009: 149–150). Other prosocial lies may include lies for the greater good of someone's mental health ("I was overjoyed when I found out I was pregnant with you"), or lies to sick or dying people ("Don't worry about your finances – they're all in order"). People may also lie about prosocial deeds for the sake of modesty (for example, Molly tidies the classroom at play time but insists "I don't know who tidied up the classroom, sir"). Such lies may be viewed positively (Lee et al. 1997), or perhaps as not even lies at all (Fu et al. 2001). One situation where people often believe that lying is morally good is when it is done to protect innocent people from unwarranted violence (Strudler 2009: 144, Saul 2012: 6) (though Kant and Augustine would seem to disagree).

Additionally, there are some situations, such as haggling over a price, where it is reasonable – and perhaps even expected – that people will lie (Strudler 2009: 143); this does not mean it is morally *good* to lie when haggling – but it does seem permissible to tell a lie such as "I'm making a loss if I sell it for £15". Therefore, we can see that although we generally view lying and deception as (pro tanto) wrong, the rigidity in lie-avoidance which is mandated by Kantian or Augustinian approaches do not necessarily result in what we view as superior moral behaviour, and there are numerous situations where we consider lying to be permissible or even good. In Chapter 4, when I consider robots which display 'fake compassion', I suggest that it is beneficial for robots to engage in prosocial deception; it can make them seem more personable, enhancing patient wellbeing.

I have noted so far that there is a moral presumption against lying, but that some forms of lying are permissible or even morally good. But *why* do we have

this presumption against lying, and why do we think that some lies are acceptable whereas other are not? One theory which plausibly answers both of these questions is that some – but not all – lies can be a breach of trust (Williams 2002: Ch. 5). Suppose I have two friends who have both lied to me today: Violet told me that she likes my new boots (because she can tell I love them); Alina told me she has postponed her party until next week (because she does not want me at her party tonight). When I discover that both of them have lied – Alina's party really *is* on tonight, and Violet actually *does not* like my new boots – I do not feel that my trust has been breached equally by both women. Alina has broken my trust, and told a lie which is not conducive to a good friendship, whereas Violet has lied in order to make me feel happy, and has not broken my trust. It would therefore seem that breaching trust explains the wrongness of (antisocial) lying, and it also explains why we see prosocial or 'white' lies as unproblematic: big lies violate trust, but white lies do not (Williams 2002: Ch. 5). In fact, it has been shown that prosocial lies – even when they are discovered – actually *increase* the amount of benevolence-based trust[50] which the (prosocially) deceived party has in the (prosocial) deceiver (Levine and Schweitzer 2015). Breaking trust by telling antisocial lies does not seem to be constitutive of good character, whereas telling prosocial lies helps to strengthen and improve social relations, and could be said to stem from a positive character trait.

When I discover that the market vendor is not really making a loss by selling me the item for £15, this does not mean I have lost my trust in him, because I

[50] 'Benevolence-based trust' involves trusting someone to be nice; contrastingly, 'integrity-based trust' involves trusting someone to be absolutely honest (Levine and Schweitzer 2015)

did not trust him to tell me the truth in that matter anyway. We trust the people closest to us to tell us the truth, so the breach of trust also helps to explain why it feels worse to be lied to by someone close to you than it does to be lied to by a stranger. Finding out that Julie in Human Resources has lied to me about getting divorced is much less troubling than finding out that my parents have lied to me about getting divorced. A trusting relationship is also sometimes what facilitates deception: the deception is only successful *because* of the trust which the victim has in the deceiver (Strudler 2009: 140). Suppose Liana lies to her husband Dmitri, telling him he can buy his passport at the airport (he believes this lie, and thus misses out on his holiday). Dmitri believed his wife's lie precisely because he trusts her, and thus the betrayal is felt all the more keenly (Margalit 2017: 52–53). Williams (2002: 126) suggests that being truthful involves sincerity – an intention to be honest. When someone is in a position of trust, lying violates that, which is morally problematic (Williams 2002: 124). It may still be (pro tanto) morally wrong to lie to a stranger, but strangers only tend to trust one another a little bit, and so a breach of that trust is not such a betrayal (Margalit 2017).

The normative analysis of lying and non-lying deception given above has discussed some commonly-held beliefs about different types of deception. Although some historical perspectives (such as those maintained by Augustine and Kant) take a purist stance that lying is always wrong (Griffiths 2004: 32, Kant 1996a), I believe Williams' (2002) focus on trust best explains not only the sorts of lies which are morally wrong, but also *why* lying is wrong in some contexts but not others.

4 Self-deception: A normative analysis

In this section, I provide a brief normative analysis of self-deception; a comprehensive analysis is unnecessary, since in the next chapter, I do not make a sustained argument about people who self-deceive when engaging with robots. *Some* analysis is necessary, however, since I will suggest in the next two chapters that if people treat a humanlike or animal-like robot as if it is a real human or an animal, they may be self-deceiving, and the roboticists are not entirely responsible for this.

There are times when self-deception may be useful (Rorty 1994: 211). For example, suppose that if Vivienne believes her husband Todd is stealing from her, she will kill him. Todd is in fact stealing from her, and although Vivienne has strong suspicions that he is, she deceives herself into believing that he is not. Vivienne's self-deception is useful because it saves Todd's life, and saves Vivienne from a murder charge. Self-deception can also facilitate better mental health: when the unpleasant truth is that you are disliked, inept, and boring, deceiving yourself into believing that you are admired, successful, and interesting is one way to ward off depression (Taylor 1989, Taylor and Brown 1988). Because of these occasional beneficial outcomes of self-deception, its reprehensible nature is not fully captured by consequentialist theories which may laud the above examples of self-deception because of the beneficial consequences they bring about.

Self-deception is generally seen as problematic; however, not all theories which attempt to explain why other-deception is wrong (such as that it breaches trust) can explain why there may be something wrong with deceiving *oneself*. Consequentialist and deontological accounts could be somewhat

successful here, as self-deception often brings about negative consequences (such as false beliefs), or it may be wrong in itself to elicit a belief you think is false. I suggest below that viewing self-deception as an epistemic vice best helps to capture what is problematic with the phenomenon. The normative literature on other-deception focuses on the wrongness of *deceiving someone*, whereas the primary focus in normative analyses of self-deception is on the wrongness of *allowing oneself to be deceived*.

One reason for finding self-deception reprehensible is that self-deceived people are to some extent responsible for the deception. For intentionalists (those who suggest self-deception is undertaken intentionally), the claim that people are responsible for their self-deception is fairly straightforward: the agent deceives himself intentionally (he intentionally allows himself to be deceived), and people are responsible for things they do intentionally. Although anti-intentionalists could potentially claim that agents are not responsible for their self-deception because it was unintentional (Levy 2004), many nevertheless suggest that agents bear at least some responsibility for being self-deceived (Mele 2001). Leeuwen (2013: 7) gives a useful analogy of posture: suppose Kelvin cannot be bothered to exhibit good posture because he is negligent and lazy, and his poor posture is the direct cause of his backache. It is reasonable to claim that Kelvin is responsible for his backache even though backache was not his *intention*. He is responsible for not correcting his posture, and that has caused his backache. Analogously, if someone's epistemic tendencies towards evidence-gathering and belief-formation are negligent and slapdash, they are responsible for any self-deception which may ensue – even if the self-deception itself was unintentional.

Anti-intentionalists who suggest self-deception occurs due to a desire, anxiety, hope (etc) can still maintain that self-deceiving agents are responsible, because the agent knows about their desire, anxiety, hope (etc) and they should have reasonably foreseen that self-deception would be a likely outcome of such motivations. Consider: if I know that Michelle has the desire to make me believe that p, then even if I am unsure whether she has the *intention* to deceive me about p, I ought to regard Michelle's statements about p with suspicion. The suggestion is that agents are (or should be) able to resist their desire to acquire a belief they want to hold (Mele 2001).

One way of explaining why self-deception is reprehensible is that self-deception is, or results from, an epistemic vice (Cassam 2016); these are negative character traits relating to knowledge or knowledge-acquisition, such as failing to verify sources, wishful thinking, cherry-picking or ignoring information, and wilful ignorance. These vices are not generally at play when someone is other-deceived, because the victim does not know the deception is imminent. For example, if I ask Emma in the Tourist Information booth for directions and she lies to me, it does not seem fair to say that I have any epistemic vices simply because I was taken in by her lies. On the other hand, when I ask Carlos for directions and he says "I am about to lie to you now" before lying to me with false directions, if I am *still* taken in by his lies, then it would seem reasonable to say I *do* have an epistemic vice – such as trusting dubious sources of information. If someone is taken in by a deception when they should have reasonably believed it was imminent, they are partly at fault (Ren et al. 2022). It is more of a Carlos-like situation rather than an Emma-like situation which is at play with self-deception, because one is aware on some

level that the deception is taking place. The agent can still be responsible for their self-deception even if they were not fully and consciously aware that it was taking place (Blackburn 2009: 64) because we have some control over our ways of thinking.

Aristotle (1941: 998–999) suggests that truthfulness is a virtue,[51] and it would seem reasonable to suggest that this virtue extends to oneself as well as others; one who hides from truth or chooses only to attend to flimsy or palatable information is not behaving in a way that Aristotle would commend. Williams builds on the work of Aristotle, and suggests that truthfulness involves accuracy (Williams 2002: 126). The suggestion is that people should take steps to ensure the accuracy of a proposition, and should not simply "accept any belief-shaped thing that comes into their head" (Williams 2002: 88). Although Williams is not writing specifically about *self*-deception, his argument is still relevant inasmuch as the self-deceiver does not apply themselves adequately to ensuring the accuracy of what they believe. Instead, they practise epistemic vices such as bias, cherry-picking of desirable evidence and ignoring contrary evidence, wishful thinking, and intentional ignorance. These sorts of qualities – coupled with intentions or desires to alter one's own beliefs – mean that self-deception is seldom regarded positively.

I have not provided a highly detailed normative analysis of self-deception; however, I trust I have shown that the phenomenon may be worrisome. Although it seems clear that other-deception is generally a moral failing,

[51] The concept of epistemic virtue is only a recent phenomenon: Aristotle's virtues are moral virtues – though the gathering of true information (an epistemic virtue) generally precedes the speaking of truth.

whether self-deception is a moral failing is less clear; it does seem reasonable, however, to suggest that self-deception is or results from an epistemic failing or vice (at least, a failing which exists at the time one is self-deceiving).

5 Conclusion

This chapter has provided conceptual and normative analyses for both other-deception (including lying) and self-deception. In §1, we saw that both lying and non-lying other-deception turned out to be slippery concepts which have provoked a range of differing conceptualisations in the literature. After my analysis of intentions, beliefs, falsity, and success, I suggested that S is deceiving A iff S does not believe that p, but S takes some action which suggests that p to A, intending or wanting that action to cause A to believe that p – and S is sufficiently successful in causing A to believe that p. In my normative analysis of other-deception, I suggested that Williams' focus on trust is useful in explaining both why lying is (pro tanto) wrong, and also why prosocial lies are morally unproblematic (because they do not breach trust).

Whilst conceptual and normative analyses are interesting in themselves, the explorations were also necessary because in the next two chapters, I examine the claim that robots are deceptive. The definition of deception is important because I later argue that some types of robot – most notably Type A, Type C, and most Type E robots – fail to meet all the conditions of other-deception because they cannot *want* or *intend* us to believe that p. The normative analysis – particularly that on prosocial deception – underpins my claim in Chapter 4 that some robo-deception may be good. I also argue that sometimes people self-deceive or 'play along' when they interact with social robots, and

although it may constitute an epistemic vice, it may promote patient health and wellbeing.

The study of deception in its various forms is philosophically important in its own right, but is all the more important when applied specifically to robots, and we shall see in the next chapter that although robo-deception is a legitimate concern about futuristic or fictional robots, worries about robo-deception may be unfounded when applied to the sorts of robots which exist today.

Chapter 3

Robo-deception: Types of robo-deception and the danger (or lack of danger) of their occurring

The concern that robots or their creators may be deceptive has been articulated repeatedly in the roboethics literature (Matthias 2015, Leong and Selinger 2019, Kaminski et al. 2017, Sharkey and Sharkey 2020, Sparrow 2002, Sparrow and Sparrow 2006, Turkle et al. 2006, Turkle 2017) and is the focus of this chapter. Although some analyses of robo-deception[52] already exist (Sharkey and Sharkey 2020, Danaher 2020), there are some issues which have not yet been adequately untangled – issues which I attempt to address in this chapter. These issues are:

- **Lack of intent:** Not all robots can deceive, because they cannot want or intend to cause particular beliefs in users – because the

[52] I use this term to refer to robots *themselves* deceiving users, or when roboticists use robots to deceive users. Although these are generally distinct, in the case of some more advanced (futuristic or fictional) robots with agency, the lines between the two may blur.

robots lack a theory of mind. It may still be apt, however, to say that roboticists[53] deceive users.

- **Lack of success:** Writers claim that robots are deceptive without attending to the fact that 'deceive' is a success verb; in fact, robots are only deceptive if users believe the ruse (and I suggest that in many cases, users are not fooled).
- **Self-deception:** When users interact with humanlike or animal-like robots, users may self-deceive, or may 'play along' and behave as if the robot is a human / animal / has emotions (etc) whilst knowing that it does not.
- **Types of robo-deception:** Different forms of deception are more relevant to some types of robot than others.

This chapter is a response to writers who claim that robots are deceptive (Sharkey and Sharkey 2020, Sparrow 2002, Sparrow and Sparrow 2006, Turkle et al. 2006, Turkle 2017); a claim which is typically coupled with normative arguments about the badness or wrongness of robo-deception. This chapter is largely a conceptual examination of whether robots are in fact deceptive (and how likely that deception is), but I do consider normative questions of how bad it is when users are deceived.

This chapter proceeds as follows: in §1 I discuss what robo-deception involves, referring to the conditions for deception I laid out in the previous chapter. Then, using my robot matrix from Chapter 1, I discuss which types of robot are

[53] I use 'roboticists' to refer to people who design, create, and program robots. This may be multiple people, and how responsibility or fault should be apportioned between them is beyond the scope of this thesis.

themselves capable of deceiving users, and which robots could only be tools through which roboticists could deceive users. I discuss the likelihood that the four conditions for deception would actually be met, noting that D3 and D4 in particular seem unlikely. In §2, I distinguish four different types of robo-deception: anthropomorphic deception, zoomorphic deception, disanthropomorphic deception, and basic other-deception. I explain which types of robots these forms of deception pertain to, and I discuss the likelihood and potential consequences of their occurring. I suggest that writers who raise the alarm about anthropomorphic and zoomorphic deception are often doing so unnecessarily; I also suggest that we should perhaps be (slightly) more concerned than we are at present about basic other-deception.

1 Robo-deception

This section addresses whether it is possible for robots to deceive users; this hinges on the definition of deception (I use the conditions reached in the previous chapter), the abilities and intelligence levels of the robots in question, and the likelihood of users believing what is suggested to them by robots.

Several writers raise concerns about robo-deception; they argue it is deceptive when robots behave in a way which suggests something untrue – such as that the robot has emotions or agency (Sparrow and Sparrow 2006, Sparrow 2002, Sharkey and Sharkey 2012, Turkle et al. 2006, Turkle 2017, Wallach and Allen 2009). Who is doing the deceiving is not always made explicit in these accounts, and often the claim that 'it is deceptive' does not point the finger at any party in particular; however, the implication is probably that roboticists are the deceptive ones, since the arguments generally suggest that robots do not possess agency, responsibility, and suchlike. Other writers

suggest that robots *themselves* could be deceptive: some see this as potentially morally problematic (Danaher 2020), whereas others discuss ways it may be good for robots to deceive users (Isaac and Bridewell 2017).

I refer to those who raise (unwarranted or inflated) concerns about robot deception as 'robo-deception alarmists'.[54] Below, I argue that much putative robo-deception is not in fact deception, since present-day robots cannot meet the intention to deceive condition (D3), and often, the success condition (D4) is not fulfilled either, because users do not change their beliefs as a result of the (putative) attempted deception. Later, I discuss deception by roboticists, but for now, I focus on whether robots *themselves* can deceive users;

In the previous chapter, I gave these conditions for other-deception:

S is deceiving A iff:

D1. S does not believe that p

D2. S takes some action φ (or omits to take action) which suggests that p to A

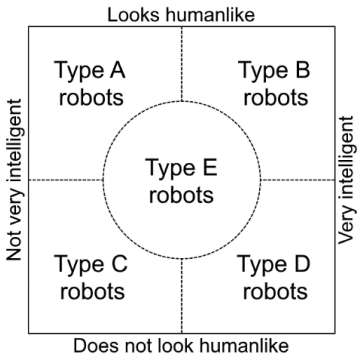
D3. S intends or wants S's suggestion that p to cause A to believe that p (the intention condition)

D4. S's φ-ing is sufficiently successful in causing A to believe that p (the success condition)

[54] The term 'alarmists' has a pejorative tone to it, and although I admit there may be a few occasions when alarm is warranted (discussed later), I believe that generally, robo-deception is both unlikely and unconcerning, and that alarm is unwarranted.

In cases of robo-deception, the robot is S (the ostensible deceiver), the user is A (the party who is (perhaps) deceived), and p is whatever proposition the user is (supposedly) being deceived about. As my parenthetical hedges suggest, I believe that many cases which are described at robot *deception* do not involve deception proper.

I shortly discuss the extent to which robots can deceive. During this discussion, I refer to the types of robot established in Chapter 1. There, I created a robot matrix (see right) and distinguished five different types of robot based on looks (humanlike or unhumanlike), and intelligence levels. Thus:



- **Type A robots**: look humanlike; not very intelligent

  e.g. a sexbot
- **Type B robots**: look humanlike; very intelligent

  e.g. Data, C-3PO
- **Type C robots**: look unhumanlike; not very intelligent

  e.g. an industrial robot, Paro, a driverless car
- **Type D robots**: look unhumanlike; very intelligent

  e.g. HAL-9000
- **Type E robots**: look somewhat humanlike; mid-level intelligence,

  e.g. Pepper, Care-o-bot

Considering whether robots *in general* deceive people is like considering whether animals *in general* are dangerous to humans. When a group of entities is sufficiently heterogeneous, it is exceptionally difficult – and not

particularly useful – to make generalisations about it: bears, sharks, and funnel-web spiders can be dangerous to humans, but guinea pigs, goldfish, and butterflies are not. The short answer to the question "Are robots deceptive?" is "It depends: some (fictional, futuristic) ones are, and some are not." To more fully answer the question, one must consider which types of robot are deceptive – and how. In order for us to say that particular types of robot deceive users, we must consider whether robots of each type can meet the conditions for deception (D1-D4).

Recall that although the majority of this thesis focuses on carebots (which are generally Type E robots), this chapter considers all the different types of robot and robo-deception, providing the foundation on which the next chapter is built. Chapter 4 examines one way in which carebots are said to be deceptive (viz. by seeming to care); in order to provide this examination, it is important that we first understand different types of robo-deception. Let us now proceed to assess whether robots (or roboticists) can meet the conditions for deception.

1.1 Do robots suggest that p while not believing that p (D1 & D2)?

Conditions D1 and D2 are relatively straightforward. Condition D1 requires that the robot does not believe that p. All robots can meet this condition. This includes highly intelligent robots of Types B and D: Data and HAL-9000 are capable of not believing a proposition, just as you and I are capable of not believing something. Robots of Types A and C (and most or all Type E robots too) do not have doxastic states, so they can also meet condition D1; less

intelligent robots do not believe that p, just as rocks and tables do not believe that p.[55]

Condition D2 is that the robot takes some action φ which suggests that p. Again, all robots of Types A to E are able to take some form of action (φ-ing) (by definition, robots sense, think, and act), and that action may suggest that p to a human user. Condition D2 does not require any intention or theory of mind[56] by the robot, so all robots are capable of meeting condition D2. Whether robots can meet conditions D3 and D4 is far less straightforward, however.

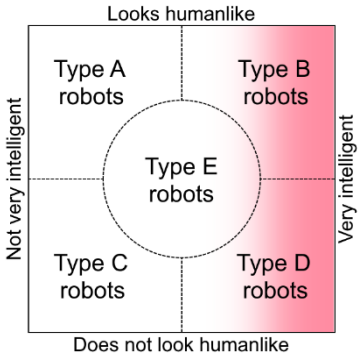
1.2 Do robots intend to cause users to believe that p (D3)?

In this sub-section, I discuss whether robots meet condition D3 – the intention to cause beliefs in users – and this is where things start to become more complex: some robots can meet this condition, while others cannot.

[55] Some writers (such as Primoratz (1984: 54)) have *S believes that p is false* as the belief condition; rocks, tables, and less intelligent robots could not meet this. However, I dismissed Primoratz's condition as being too narrow, as it precludes claims where the subject term lacks a referent ("The King of France is bald") and claims which a user is agnostic about. If one wishes to adhere to Primoratz's formulation, then one can immediately determine that robots without doxastic states (Types A, C, and most Type E robots) cannot deceive as they cannot meet condition D1.

[56] The ability to ascribe mental states to others, and to understand the sorts of mental states they might have.

'Intention' can mean different things. For example, one can have future intentions ("Roshni intends to visit Greece this summer") and one can act intentionally, meaning on purpose ("Moe is intentionally ignoring me"). An intention might be understood as a mental state involving wanting or desiring to φ and/or being in favour of φ-ing (Davidson 1978); if this is the case, then only highly intelligent robots with mental states (shown in pink) could intend anything. However, intending does not seem to be *equivalent* to wanting or desiring something: I may want or desire to be 21 again, but that does not mean I *intend* to be 21 again – because I know this cannot happen. It seems impossible for me to *intend* to do something which I know cannot occur (Kavka 1983, Levy 2009). Perhaps, then, intending is not merely *wanting* to φ, but *planning* to φ (Bratman 1985, 1987); however, Bratman describes planning as an 'attitude' – something which only highly intelligent robots (shown in pink) could possess. Nevertheless, if we understand intending (qua planning) in its broadest sense, then all robots can intend to φ because all robots can plan according to the sense-think-act paradigm.



This does not necessarily mean that all robots can meet condition D3, however, since this requires a specific type of intending: *intending to cause beliefs in users.* This requires a theory of mind – something which less intelligent robots simply do not possess (those *not* shown in pink, viz. all Type A and Type C robots, and most or all Type E robots, plus a substantial proportion of Type B and Type D robots). Less intelligent robots simply do not

understand that humans have beliefs, and so they cannot intend or want to cause beliefs in users, even if they can intend other things. This means that most robots (including all real-world robots) cannot meet condition D3, so cannot deceive users.[57] Currently, robots with a theory of mind are merely fictional or hypothetical. Data and HAL-9000 (Types B and D respectively) both have a theory of mind, and can have the intention to deceive (indeed, they sometimes do). It is also possible that some of the most intelligent (fictional or hypothetical) Type E robots could have a theory of mind. Thus, only the most advanced robots with a theory of mind (which are presently fictional or hypothetical) could fulfil condition D3.

The discussion does not end there, however, because some prominent writers suggest that robots deceive us *despite lacking any intention to do so*. For example, Sharkey and Sharkey argue that "deception can still occur even in the absence of conscious intention. If a person believes that a social robot has emotions and cares about them, they are being deceived: even if no-one explicitly intended that belief" (Sharkey and Sharkey 2020: 310). Evidently, Sharkey and Sharkey are working from a different conception of deception from the one I am using, where intention is not a requirement (they support an anti-intentionalist position about deception). Sharkey and Sharkey draw on some literature, particularly that focused on 'deception' in the animal kingdom, where 'deception' is considered to be something akin to a misleading signal which benefits the sender. For example, if an insect appears visually similar to a leaf, making it less likely to be predated upon, some biologists / naturalists

[57] If there exists a robot which does have a theory of mind, then I will happily concede that it can intend to deceive users.

call this deception – and Sharkey and Sharkey follow this trend. However, as I noted in the previous chapter, this would appear to capture too much: recall Jared, who looks like a particular celebrity, and although he makes no effort to cash in on this, and he insists he is not the celebrity, he nevertheless receives preferential treatment when he goes out. According to an anti-intentionalist definition of deception, Jared is deceiving people simply because he benefits from the way he looks. If we wish to call camouflaged insects and robots without a theory of mind 'deceptive' even though they lack any deceptive intention, then we must call Jared deceptive too. This does not seem right.

In the case of insects and Jared, some benefit is conferred upon the (ostensible) deceiver, whereas the sorts of robots Sharkey and Sharkey (2020: 310) call deceptive – such as those which appear to be sentient or emotional but in fact are not – no benefit at all is conferred upon the robot. This would seem to make the definition of deception with which they are working even broader – where there is no deceptive intention by the robot, *and* no benefit to the robot. All that is required for deception on Sharkey and Sharkey's account is simply that the user ends up (falsely) believing that the robot has emotions or cares about them (Sharkey and Sharkey 2020: 310).

However, this is misguided because it does not account for epistemic vice or foolishness. Simply because S believes that p about X, and p is not true of X, does not necessarily mean that S has been deceived. If Sheena believes that Xavier cares about her when in fact he does not, we cannot immediately conclude, without further information, that Sheena has been deceived. Similarly, if Sheena believes that a robot cares about her when in fact it does

not, we cannot immediately conclude, without further information, that Sheena has been deceived. We cannot even be certain that Sheena has been misled. Perhaps Sheena is simply being epistemically careless by reaching hasty and dubious beliefs without justification. Suppose Xavier and the robot simply said hello to Sheena, and Sheena formed wholly unjustified beliefs about their loving feelings because of this simple greeting. If we do not believe that Sheena has been deceived in the case of Xavier, then nor should we conclude that Sheena has been deceived in the case of the robot. Sheena could simply be naïve, foolish, or epistemically careless. Thus, meeting Sharkey and Sharkey's (2020: 310) notion of deception is not necessarily a marker of deception proper.

People might use the term 'deception' imprecisely when they say that animals deceive predators by using camouflage. However, such claims are not consistent with a thorough conceptual analysis of deception, which shows that deception proper must have some sort of intention, or else deception is ostensibly occurring any time anyone reaches an erroneous belief about anything based on some signal or other. It is clear that the intention condition (D3) needs to be retained. Given that today's robots do not have a theory of mind, they cannot meet the intention condition, and therefore cannot themselves deceive users. They may *mislead* users (since there is no intention condition with misleading) just as insects, optical illusions, and inanimate objects can mislead, but misleading is not equivalent to deceiving (I return to discussion of misleading later). We should note, however, that roboticists and advanced robots with a theory of mind *can* fulfil condition D3.

1.2.1 Roboticists may intend to deceive (D3)

One might claim that *it is deceptive* for robots to imply something false without claiming that the robots are the ones doing the deceiving. A trompe l'œil (meaning 'trick of the eye') painting can look like an alcove exists when really it is a painting on a flat wall. When we claim that the trompe l'œil painting is deceptive, we are not suggesting that the *paint* has the intention to deceive – we are suggesting that the artist had the intention to deceive (D3), and she used the paint as a tool to execute that deception. Some writers suggest that *it is deceptive* to create robots which behave as if they have inner mental states when in fact they do not (Sparrow and Sparrow 2006, Sparrow 2002, Wallach and Allen 2009, Turkle 2017). In such arguments, writers are suggesting that the roboticist*s* are the ones doing the deceiving, using the robots as their tools. It is possible (though not certain – as I discuss later) that roboticists can meet condition D3 (as well as D1 and D2).

So, to summarise my arguments regarding the intention to deceive: robots of types A and C (and probably E too) cannot meet condition D3, so cannot be said to deceive. Some (fictional or futuristic) Type B and Type D robots could have a theory of mind sufficient to form an intention to deceive, and meet condition D3. Roboticists making any type of robot can meet condition D3, so could potentially use their robot creations to deceive users. Let us now move on to consider whether the final necessary condition for deception (D4) can be met.

1.3 Do robots (successfully) cause users to believe that p (D4)?

As discussed in the previous chapter, deception is a success verb. If I tell you that I live on the moon, you will not believe me, and therefore I have not deceived you. In order for robots to deceive users about p, the users must actually come to believe that p – or at least, to sufficiently shift their belief in the *probability* that p. I noted in the previous chapter that not all φ-ing which suggests that p will cause a 100% belief in the listener, but that the φ-ing may still be considered a (successful) deception if the listener's belief is *sufficiently* shifted towards p. I did not specify what counts as sufficient, since measuring levels of belief is difficult, and even if it were possible, further analysis would need to be conducted into ascertaining what the sufficient level of belief-shifting is; the question may be unanswerable. Despite this lack of clarity on what counts as sufficient, we can say that that in order for deception to occur, the listener's belief must be 'sufficiently shifted' towards p, whatever a sufficient shift may be.

The content of p and the willingness of users to change their beliefs both affect whether the user's belief is sufficiently shifted towards p. It is not possible or necessary for me to assess whether every suggestion that p by a robot would be believed by every user. Some robots' suggestions that p will be believed by some users, and others will not, depending on the user and the content of p. Some users – for example, children – will be easier to trick.

Concern about robo-deception in the literature rarely relates to facts external to the robot (what Danaher (2020: 122) refers to as 'external state deception') – rather, the concern is usually that the robot deceives users about some

quality or ability the robot itself possesses or lacks – for example, whether it has emotions. Let us consider the extent to which neurotypical adults would believe that a robot has emotions. A single claim like "I feel sad today" may not sufficiently shift a neurotypical adult's belief that the robot has emotions. Repeated claims of a similar nature may sufficiently shift some users' beliefs that the robot has emotions; however, neurotypical users understand that today's robots simply cannot feel emotions. Indeed, writers such as Sharkey and Sharkey, Sparrow and Sparrow, and Turkle have not had their beliefs sufficiently shifted towards believing that robots experience emotions, despite being aware that robots make such claims. Why, then, would these writers think that *other* people would believe that robots possess emotions, if they themselves can see through the ruse? Perhaps the implication by these writers is that other people are less savvy than they are? Although fictional examples and thought experiments can be constructed, it seems that there are few real-life examples of neurotypical users *actually* believing that robots possess emotions.[58]

There are some associations which we make between visible behaviours and internal states – for example, between smiling and happiness – and robots disrupt these associations when they smile without feeling happy. However, we are able to grasp that not all teeth-showing is an indication of happiness: humans can and do give fake smiles for photographs and in various social situations; we can also grasp that when chimpanzees (and countless other

[58] Neuro-compromised users such as people with dementia or learning disabilities may believe that robots have emotions. However, they may also believe that teddy bears are babies, or all kinds of other false and bizarre notions. This thesis focuses only on neurotypical adults.

mammals) show their teeth, it is seldom a sign of happiness. It is reasonable to maintain that we should be able to see a robot smile without inferring that it feels happy.

Supposing a robot claims it has emotions, the extent to which it is believed will rely in part on the type of robot which is making the claim. If we know that the claim is coming from a highly advanced Type B (or Type D) robot, then this claim will be more believable. The concern from robo-deception alarmists is that Type A or Type E robots – which appear very or loosely humanlike but possess no emotions – will appear to have emotions, and be believed because of our tendency to anthropomorphise. This tendency is well-documented: not only do we see 'faces' in inanimate objects (known as pareidolia), we also anthropomorphise things which look loosely humanlike – such as Type A, B and E robots.

Of course, a problem is that sometimes, users do not know the type of robot with which they are interacting (they can easily assess how humanlike it looks, but may not know its intelligence). This includes roboethicists and even robot designers and creators: none of us can "make educated guesses about what [robots] can do just by looking at them" (Leong and Selinger 2019: 302). We therefore tend to use our experience with the robot to give us clues about its capabilities – and this facilitates anthropomorphic guesswork. We are more likely to ascribe humanlike abilities – emotions, agency, consciousness, animacy, autonomy, and suchlike – to a robot if it looks humanlike. Even in situations where users reasonably believe that a robot is cognitively unsophisticated, they may nevertheless respond to it as if it has humanlike qualities, simply because it looks humanlike. For example, studies

demonstrate that people show compassion towards humanoid 'robots'[59] which behave as if they are scared, even if they have been told the robot is simplistic (Scheeff et al. 2002, Horstmann et al. 2018). We should be cautious, however, in drawing any conclusions from this: behaving *as if* a robot has emotions is not necessarily equivalent to *believing* it has emotions. I can behave *as if* a child's imaginary friend has emotions without actually believing it.[60] I return to the issue of robots seeming to display caring emotions – what I call 'fake compassion' – in the next chapter.

It is important to remember that deception, being a success verb, only occurs when the listener's belief is sufficiently shifted towards p by the robot's suggestion that p. Of course, p could be innumerable propositions. We are inclined to believe Alexa or Siri in many factual matters, but users are probably more astute than robo-deception alarmists seem to believe. For example, if I ask Alexa when Winston Churchill died and I am given the answer "2022", I have enough background knowledge to know that this is untrue, even though I do not know exactly when Churchill died.

Perhaps robo-deception alarmists' concern is not that people *actually* believe robots' false claims that p, but rather, that roboticists are *attempting* to deceive people, and there is a chance users *could* believe it. Concerns about failed

[59] The technologies used in these studies did not fulfil the sense-think-act paradigm, so were not robots by my definition. Rather, they were technologies which appeared visually similar to Type E robots such as Pepper (and users perceived them to be robots), but the technologies were in fact human-controlled automatons.

[60] *Consistently* behaving as if the imaginary friend has emotions seems more indicative of belief. For example, if I always talk to the imaginary friend, and whenever I am asked, I say I believe it has emotions; this could be taken as evidence of belief, whereas a few isolated and inconsistent behaviours cannot.

attempts to φ when φ-ing is a nefarious activity may be legitimate – for example, if there have been a spate of attempted (but unsuccessful) burglaries in my neighbourhood, I should still be concerned, because *next* time, an attempt might be successful. Burglars may perfect their burgling skills, just as roboticists may perfect their robots' ability to convince users that p. When p is a false but plausible proposition such as "Nigeria is larger than Ethiopia", users may believe that p when it is stated by a robot; however, when p is less plausible, such as "Nigeria is larger than Jupiter" users are unlikely to have their belief shifted when a robot states it (this issue is explored further in §2.4). When the content of p relates to the robot itself – such as whether the robot is sentient, emotional or suchlike – I believe that neurotypical adults are probably more shrewd that robo-deception alarmists seem to believe, and would not be fooled by such claims. This is something which I discuss in greater detail throughout §2 of this chapter, and in the next chapter.

2 Types of robo-deception in the literature

As noted previously, there is a burgeoning literature by robo-deception alarmists and their opponents, discussing the possibility and morality of robo-deception. However, as Danaher writes: there is "some confusion pervading this debate [about robo-deception]. There is a tendency to conflate the different kinds of deception and fakery that can arise" (2020: 117). This claim seems undoubtedly true, and some clarity over the different types of robo-deception is both necessary and welcome. However, although Danaher's paper makes some headway towards conceptually clarifying three types of robo-deception, there remain some grey areas (outlined below), and he downplays the danger posed by one form of robo-deception (what he calls

external state deception) – something which I believe should not be entirely downplayed. Moreover, it would be useful for users (and roboethicists) to know which robots are the 'likely culprits' for different forms of deception. Below, I not only provide an account of some different types of robo-deception, I also improve upon existing accounts of robo-deception by highlighting the likelihood of their occurring, and which robots are most likely to be involved in which forms of deception (whether as deceptive agents, or as tools through which roboticists deceive users). Knowing which forms of deception to look out for with which robots could help users guard against being deceived.

Before putting forward my four different types of robo-deception, I briefly explain some types of robo-deception which others have conceptualised.

- **Dishonest anthropomorphism** (Leong and Selinger 2019, Kaminski et al. 2017, Turkle 2017): The robot appears humanlike, potentially causing users to respond to the robot as if it were human, or to believe that it possesses humanlike qualities.
- **External state deception** (Danaher 2020: 121): The robot falsely suggests that p, where p is some factual matter unrelated to the robot (e.g. "The capital of Spain is Berlin").
- **Superficial state deception** (Danaher 2020: 121): The robot falsely suggests it can φ or is φ-ing (e.g. "I can record video").[61]
- **Hidden state deception** (Danaher 2020: 121): The robot falsely suggests it cannot φ or is not φ-ing (e.g. "I can't detect infra-red").

[61] There is some crossover between dishonest anthropomorphism and superficial state deception, for example if a humanlike robot falsely claims it feels sad.

Although such terms may initially appear useful, some problems exist. For example, if a robot knowingly claims its motherboard is blue and in its 'chest', (when in fact its motherboard is green and in its 'head') and the user believes it, this deception does not seem to fit neatly into any of the above categories. It is also worth noting that the intention condition (D3) and/or the success condition (D4) are seldom stipulated in the above definitions. This is conceptually problematic because it means that too many phenomena are captured as 'deceptive' when they are unintentional (the roboticist simply wanted to create a fun robot for people to interact with, not to fool people into having false beliefs), or unsuccessful (a robot claims it likes eating hot dogs but users know it does not).

I suggested above that in spite of robo-deception alarmists' worries, present-day robots often do not deceive users because (a) they do not have the theory of mind to intend to bring about particular beliefs in users, and (b) neurotypical users are often (but not always) savvy enough not to be misled by robots' suggestions that p (particularly where p is implausible, such as present-day robots being sentient, or Nigeria being larger than Jupiter). Nevertheless, some robo-deception worries may still persist, such as:

- Roboticists may use robots to (attempt to) deceive users
- Fictional or future robots may have a theory of mind and intend to deceive users
- Fictional or future robots may be more convincing, so better able to make users believe that p (whilst the robot does not believe that p)

With this in mind, and relating to the robot matrix I outlined in Chapter 1, I presently outline some potential concerns relating to robots of different types. These forms of deception may not occur, or even if they do occur, it may be morally unproblematic. Whether or not they occur, it is useful to the roboethics literature to categorise different types of deception which could theoretically occur when people interact with robots – and it is prudent for users to be aware of the possibility too.

Below, I describe four types of robo-deception:

- **Anthropomorphic deception**: Users are deceived into thinking humanlike-looking robots have humanlike mental qualities. This is similar to dishonest anthropomorphism, but with a success condition involved; this is one form of superficial state deception (Danaher 2020: 121).
- **Zoomorphic deception**: Similar to anthropomorphic deception, but with animal-like robots. This is another form of superficial state deception (Danaher 2020: 121).
- **Disanthropomorphic deception**: Users are deceived into thinking unhumanlike-looking robots have unhumanlike mental qualities (including no mental qualities). This is a form of hidden state deception (Danaher 2020: 121).
- **Basic other-deception**: any form of deception by a robot (or by a roboticist via the robot). This includes all the above types of deception, and more besides.

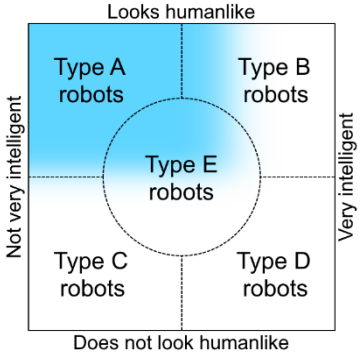
As I observed in Chapter 1, the majority of carebots in existence today are Type E robots, and – perhaps worryingly – these robots could be involved in *all four* types of robo-deception I am about to discuss.

2.1 Anthropomorphic deception

Consider this scenario: you see a young woman seemingly lost in her own thoughts, and ask her if she is OK. She nods stiffly, but avoids eye contact. After a short conversation with her – where she tells you her name is Faye – it becomes clear that Faye is in fact a robot. Faye seems so humanlike that even though it is clear 'she' has very limited intelligence and no emotions, you cannot help but interpret 'her' facial expressions and perfunctory answers as signs of sadness. If we stipulate that the roboticist intended you to believe Faye is sad, then you have experienced anthropomorphic deception with a Type A robot.

Anthropomorphism involves ascribing or inferring human qualities to something which does not have those qualities. We sometimes do this with inanimate, unhumanlike objects – I might suggest that my oven 'wants to make a fool of me' when it stops working just before my dinner party. When something looks more humanlike, we are even more likely to anthropomorphise. We are biologically disposed to seek out and relate to other people: newborn babies can distinguish between facial expressions such as sad, angry, and happy (Farroni et al. 2007), and by five years old, a child can recognise and understand others' facial expressions as well as an adult can (LoBue 2016). Recognising and understanding other people is hard-wired into us (Darwin 1872), so when we see something loosely or very humanlike, we are 'anthropomorphically triggered' to interpret what we see in

human terms, like mental states or emotions: it comes naturally to us. When a robot looks very similar to a human, as Faye does, it is easy to see how one could anthropomorphise the robot, and interpret 'her' facial expressions as having some deeper meaning.



So-called 'dishonest anthropomorphism' is identified as potentially problematic by a number of writers (Leong and Selinger 2019, Damiano and Dumouchel 2018, Nyholm 2020, Fink 2012, Duffy 2003, Kaminski et al. 2017). It occurs when robots trigger our tendency to anthropomorphise. Anthropomorphic (robo-)deception occurs when a user erroneously attributes humanlike qualities to a robot because it appears humanlike, which was the intention of the roboticist. For example, suppose Faye's creator purposely gave 'her' a melancholy demeanour in order to trick users into believing 'she' is sad – and this does indeed trick users. Anthropomorphic deception pertains only to robots which look somewhat humanlike (Type A robots, and some Type B and Type E robots – shown in blue), because the humanlike appearance is what triggers the anthropomorphic response, causing users to overestimate robots' capabilities, and believe that robots have humanlike abilities. Anthropomorphic deception involves the roboticist deceiving users – for a robot to be the deceptive agent, it would need to meet condition D3, which requires a theory of mind; in which case, attributing humanlike qualities to the robot would be apt rather than deceptive. Moreover, for a robot to

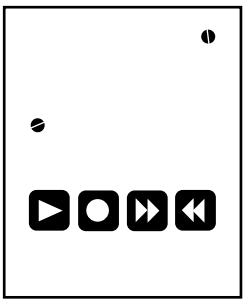
anthropomorphically deceive users, it would need to have decided how it looks, which seems unlikely.[62]

Several writers (Leong and Selinger 2019, Duffy 2003, Fink 2012, Nyholm 2020) discuss the potential for anthropomorphic deception when users interact with robots which look quite humanlike. However, not all of these writers include both an 'intention' condition, and a 'success' condition. For example, Leong and Selinger evidently consider intention to be an unnecessary condition: they write that "roboticists who do not understand the power of anthropomorphism will *unintentionally* create products that [elicit an anthropomorphic response]" (2019: 300, my emphasis). Damiano and Dumouchel use phrases like "anthropomorphic appearance" (2018: 3) and "anthropomorphizing design" (2018: 4) which suggest a lack of success condition – viz. that some robots are anthropomorphic *whether or not anyone infers or ascribes humanlike qualities to them*.[63]

[62] It would not be completely impossible for a robot to decide its appearance: in an episode of *Star Trek: The Next Generation* ["Offspring", S3, Ep16] (Frakes 1992) a robot ('Lal') is created in a simplistic, androgynous (humanoid) form, and allowed to choose its own appearance. However, Lal is a sentient Type B robot, so attributing humanlike qualities to her is not simply anthropomorphism – it is correct.

[63] This is why I am not using the term 'dishonest anthropomorphism', which apparently occurs whenever a roboticist makes a robot look humanlike, *whether or not there's an intention to deceive, and whether or not users are taken in by the robot's appearance*. In other words, dishonest anthropomorphism is not a success term, but anthropomorphic deception is. Even if dishonest anthropomorphism occurs regularly, if users are not fooled by it, then there seems little cause for concern.

The intention condition and the success condition are important, however, since without them, too many phenomena are captured as anthropomorphic deception. Suppose there is a cuboid smart speaker robot with two screws holding it together, and four buttons near the bottom for users to press (see



left). Suppose it was not the roboticist's intention to make the robot look face-like; she is not trying to make users to ascribe humanlike qualities to it. However, Ted – the user – sees it and falsely believes it has emotions. Ted has not been anthropomorphically deceived by the roboticist; it is a stretch even to claim that Ted has been misled. Rather, Ted has foolishly jumped to a false belief. Without the intention condition, we end up concluding that *any* time a user such as Ted ascribes a humanlike quality to a robot (which resembles a human in even the loosest of ways), deception has occurred. This is simply not true: *some* users who interact with humanlike robots are (other-)deceived, some self-deceive (discussed later), some are unintentionally misled, and others – like Ted – are simply foolish people with epistemic vices.

Even if everyone agrees a robot looks *exceptionally* lifelike – such as a sexbot – this does not mean there was an intention to anthropomorphically deceive users. Roboticists may not have clearly defined intentions. For example, sexbot creators may try to make their robots look as lifelike as possible (since those robots sell better); whether they intend users to believe *the robot possesses human qualities* is not clear. Even if a sexbot creator is happy to hear that users (falsely) believe their sexbot loves them, this does not mean the roboticist had the *intention* to elicit that belief in users.

The success condition is also important if we are to capture all and only instances of deception: if Ronan interacts with the Faye robot, and comes away believing it is nothing more a well-made robot without mental states, emotions, or sentience, then no deception has occurred (only attempted deception, assuming the roboticist intended users to anthropomorphise the Faye robot). To suggest otherwise is to warp the meaning of deception. It would mean that if I tell you "I am a two-headed hippo" (intending you to falsely believe it) then I have deceived you, which is absurd: I have lied to you, but I have not deceived you, because you did not believe me.

Self-deception often plays a role in anthropomorphism. Even robo-deception alarmists about anthropomorphism seem to hint that users may be at least partly responsible for their erroneous beliefs. For example, Kaminski et al write: "Looking at an adorable robot, we may forget they have radar and thermal sensors, [we might think] it has human hearing levels when in fact it […] can hear a heartbeat at 300 yards" (2017: 996). It seems reasonable to say that if a user 'forgets' the robot's sensory abilities simply because it is 'adorable', this is not solely the roboticist's responsibility. People should be epistemically careful, and try to check the accuracy of their beliefs, rather than just accept any idea that comes into their head (Williams 2002: 88). Consider: present-day sexbots look very humanlike, but have low intelligence levels – some are little more than dolls. Suppose Albert – a neurotypical adult – has never even heard of sexbots, but is introduced to one. It should really only take Albert a minute or two before he realises it does not have humanlike mental states. If, in spite of the robot's obvious limitations (formulaic and repetitive phrases; non-sequiturs, etc), Albert believes it has humanlike mental

states, then this is not necessarily a case of roboticists anthropomorphically deceiving Albert: this could be Albert deceiving himself. One might object that it is not self-deception, but a related phenomenon such as delusion, or an epistemic vice such as ignoring strong evidence, but this would still mean that the roboticist is not wholly responsible for Albert's delusion / vice which results in his erroneous belief.

Even if Albert is being other-deceived, he may still be *partly* responsible for his deception. A deceived party is not always a passive victim of deception; if the deceived party does not ask sensible questions which could uncover the truth, then they are partly responsible for any ensuing deception (VanEpps and Hart 2022). For example, if a seller says an expensive vase is "quite old", and the buyer assumes it is an antique from the Ming Dynasty, but does not ask the vase's age, then the buyer is partly (if not wholly) responsible for their erroneous belief. Would-be targets of deception have some responsibility to collect information to "detect, curtail, or prevent deception" (Ren et al. 2022 §3.5) – if they do not, they are complicit in their deception, and this is the case with anthropomorphic deception as much as any other type of deception. Just because we have a biological tendency to anthropomorphise, this does not mean we should simply allow ourselves to believe that anything with a rudimentary face has humanlike qualities.

Sometimes, people might interact with robots *as if they are human*, all the while fully understanding that the robots do not have humanlike mental states or abilities, just as we might talk to a child's teddy bear or imaginary friend as if it is real, but knowing it is not. In such cases, no other-deception or self-deception has taken place. If Diana, while playing with her children, talks to a

doll as if it is animate, this does not mean she has been deceived; it means she is pretending. People can similarly play along with robots.

Although users should be aware of the *possibility* of anthropomorphic deception, it does not seem particularly likely to occur with present-day robots.[64] Consider: children's baby dolls have humanlike faces and bodies, and might say phrases like "Want milk!" "I wuv you!"; some even 'drink' milk and soil their nappies. Yet we do not think the dolls have mental states or agency.[65]

Robo-deception alarmists are able to apprehend the truth that robots which look humanlike do not necessarily have agency or emotions, yet they worry that other neurotypical adults *would* be fooled by such robots. This implication seems not only patronising, but inaccurate. Sparrow and Sparrow write: "robots are clearly not capable of real friendship, love, or concern – only (perhaps) of their simulations […] In most cases, when people feel happy [relating to a robot], it will be because they (mistakenly) believe that the robot has properties which it does not" (2006: 154–155). Although it is true that happiness could result from falsely believing a robot has particular properties,

[64] In future, robots may be more convincing in their display of emotions. If a humanlike-looking robot consistently seems highly intelligent, emotional, sentient (etc) then it is not a Type A robot, but a Type B robot, and it may not be possible to ascertain whether it *really* has humanlike mental states. Users should perhaps engage with such robots in the same way they would engage with another creature which exhibits similar levels of intelligence, agency, emotions and suchlike (known as ethical behaviourism (Danaher 2019c)).

[65] Occasionally, dolls are mistaken for babies. A police officer in New Hampshire, USA, smashed a car window and began CPR on a highly lifelike baby doll, believing it was an unresponsive human baby trapped in a hot car (Guarino 2016). He quickly realised it was a doll (without mental states) and ceased CPR.

if it is unlikely that people will *actually* form such beliefs, then Sparrow and Sparrow may be worrying over nothing; they may be the quintessential robo-deception *alarmists*. I believe it is unlikely that even poorly educated people with little common sense would *really* believe that a Type A robot has humanlike mental states. If I am correct about that, then the likelihood of anthropomorphic deception occurring is minimal.

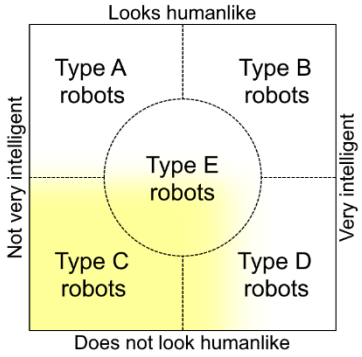
Nevertheless, it is not *absolutely impossible* for users to fall victim to anthropomorphic deception, or to self-deceive into believing that the robot has humanlike mental states; therefore, it is useful for roboethicists to have a suitable term for the phenomenon (namely, anthropomorphic deception). It is also prudent for users to be aware of potential (other- or self-)deception with humanlike-looking robots, enabling them to better guard against it (Ren et al. 2022).

2.2 Zoomorphic deception

Humans have an affinity for cute, fluffy, mammalian pets, so it is unsurprising that there exist some cute, fluffy, mammalian-like, pet-like robots. If we are anthropomorphically triggered to respond to humanlike robots as if they are humans, it is also plausible that we can be 'zoomorphically triggered' to respond to animal-like[66] robots as if they are animals.

[66] By this I mean robots which *look* animal-like – not necessarily robots which have animal-like mental states or intelligence.

Suppose p is a proposition such as 'this is a real animal', or 'this robot has animal-like mental states / emotions'. If a roboticist does not believe that p, but makes his robot animal-like because he intends to cause the user to believe that p, and the user *does* come to believe that p because of the robot's looks, then zoomorphic deception has occurred. Zoomorphic deception pertains primarily to Type C robots and a few Type D robots, plus some Type E robots[67] (shown in yellow). No present-day animal-like robots are sophisticated enough to meet condition D3 (the intention condition), but roboticists could potentially be the deceptive agents, using their robots to deceive people. Some (fictional, futuristic) animal-like Type D robots (such as K9) may be sophisticated enough to have intentions, so could meet condition D3 and be the deceptive agents, however if such robots have mental states or emotions, and users believe they have mental states or emotions, then this is true, and deception has not occurred.



There are robot pets resembling various animals: seals (Paro), dogs (Companion Pets, Aibo), cats (Zoomer Kitty, Companion Pets), and birds (Companion Pets).[68] Paro and Companion Pets are visually similar to real animals; their fur feels somewhat realistic, they move their limbs, they blink

[67] *Some* Type C/D/E robots resemble animals; others do not (industrial robots, driverless cars, smart speakers). Zoomorphic deception only pertains to robots resembling living creatures. It can also apply to robots resembling fictional creatures – e.g. dragons, Pikachu, the Gruffalo.

[68] Some robots loosely resemble insects – though these are not intended to be pets (Vartan 2020).

their eyes, they make contented noises when stroked, and they play as a pet would. They seem very pet-like.

But just how likely is it that zoomorphic deception will occur? The definition of deception used by some robo-deception alarmists makes it seem as though 'deception' is almost certain. For example, Grodzinsky et al (2015) write that deception occurs whenever a robot's behaviour "leads [someone] to believe or behave as if the machine is [a] carbon-based life form". This is an absurd claim: if Donald strokes Paro and says "You're very cute! What a good boy!" he is behaving *as if* the robot is a real animal, but this does not in any way mean he has been deceived – he may simply be playing along, unsure of what else to do with Paro, but fully aware it is a robot without mental states. The suggestion that Donald has been *deceived* simply because he strokes and talks to Paro is robo-deception alarmism in its prime. I suspect that most people who have Paro or a Companion Pet placed in their lap will stroke it like they would stroke a pet, and say a few words to it – but we certainly cannot infer deception from this.

Sharkey and Sharkey set the standard for deception a little higher than merely behaving *as if* a robot is a real pet. They write:

> If a person believes that a social robot has emotions […] they are being deceived: even if no-one explicitly intended that belief […] In the case of Paro, the manufacturers did not intend to create the false belief that the robot is an actual seal. […] Nonetheless, the illusion of sentience or cognition created for some people by its appearance and behaviour can be said to be a deception. (Sharkey and Sharkey 2020: 310)

It is not immediately clear whether the 'some people' are neurotypical adults. Sharkey and Sharkey later remark that cognitively impaired elderly people might watch a magic show without asking how the performance was accomplished (2020: 311): perhaps the implication is that neuro-compromised people are the ones who could believe that Paro is real or has mental states? The possibility of neuro-compromised people such as those with dementia mistaking a robot pet for the real thing is recognised by the makers of Companion Pets. An advert for their robot cat states: "[they] look, feel, and sound like real cats, [they] respond to petting, hugging, and motion much like the cats you know and love […] As the product is primarily for people living with dementia, [they] can believe the pet is real" (Amazon 2022).

Although it is possible that *some* young children or neuro-compromised people *might* mistake the robots for real animals, this does not entail that manufacturers are being deceptive or that they should alter their products – particularly if such erroneous beliefs are not emotionally troubling for the mistaken party. Besides, neuro-compromised people may have many confused or erroneous beliefs: they may think that dolls are real babies, cuddly toys are real animals, or that Alexa is a real woman. It is important to be sympathetic towards neuro-compromised people, and there may be ethical arguments which suggest that we should try to prevent dementia patients from being deceived. However, the possibility of some neuro-compromised people being deceived by animal-like robots does not entail the conclusion that we ought to be generally concerned about such robots. If a patient with dementia believes that a (non-robotic) doll is her son, we need not be (and generally are not) concerned about the deceptiveness of dolls. Instead, it makes sense to

focus our concerns about deception on neurotypical adults, and it seems highly unlikely that neurotypical adults would really believe that present-day robot pets such as Paro or Companion Pets are real animals, or that they have mental states or emotions. A quick observation of the robots confirms this: their faces are toy-like; their movements are not smooth, silent, and varied like a real animal's movements are; and their bodies feel solid and heavy. Research shows that users respond no differently to realistically animal-like (in this case, dog-like) robots than they do to less realistic animal-like robots (such as Aibo) (Jones, Lawson, and Mills 2008), suggesting that users do not attribute animal-like qualities to realistically animal-like robots. If users do not believe animal-like robots are real animals, or that they have emotions or mental states, then condition D4 (the success condition) is not met, and zoomorphic deception has not occurred.

Mammalian pets are very familiar to us, and mammals are often quite active, so it is easy to tell a robot from a real animal. It is possible, however, that zoomorphic deception could occur when robots resemble less interactive or familiar animals. For example, some (real) reptiles and amphibians sit motionless for extended periods and only make simple, occasional movements; therefore, reptile-bots and amphi-bots do not need to be as complex and interactive as puppy-bots in order to fool people into believing they are real animals. The bodies of some reptiles (such as turtles, tortoises and crocodilians) are also firmer than mammalian bodies, making a firm-bodied robot harder to distinguish from the real thing. A reptile-bot which sits still for extended periods, and makes slow, simple movements would seem

very similar to a real reptile.[69] Tiny robots – such as those resembling insects – may also mislead users, as users may be disinclined or unable (due to small size) to carefully examine the robot. Thus, although users are unlikely to succumb to zoomorphic deception with mammal-like robots, they may be more easily fooled by other animal-like robots.[70]

Knowing that the entity is a robot does not necessarily preclude the user from experiencing zoomorphic deception. Some robots only loosely resemble animals (such as K9, Aibo, Spot, and BigDog) but users could believe 'this robot has emotions / mental states'. This seems unlikely, however. Critics might point out that viewers were shocked and angered by a video depicting roboticists kicking Spot: some viewers said it was wrong to kick a robot dog (CNN 2015, Parke 2015). However, it is not clear that such people believe Spot has feelings or mental states. Rather, I suspect viewers were troubled because kicking robotic dogs evidences negative personality traits (such as cruelty to real animals) in the person who kicked Spot. (Similarly, Kant (1997: 212) suggests that cruelty to animals is wrong insofar as it can develop cruel traits towards humanity.) If Bilal suggests that it is wrong to stab dolls or burn effigies, this does not mean he has been deceived into thinking that dolls

[69] Last year, I visited Exmoor Zoo, where I was saddened to see a six-foot crocodile confined in a concrete cell-like enclosure. After a minute or two, I realised it was a highly realistic (non-robotic) model. Its shiny skin and motionlessness was in keeping with a real crocodile, making it convincing – whereas a model of a mammal simply would not have been convincing. The fact it was in a zoo surrounded by real animals added to the effect!

[70] Roboticists who create robots inspired by or resembling insects and spiders (Geere 2017, Vartan 2020) would probably not want users to believe the robots were real creatures, since this could quickly lead to the destruction of the robots by people who swat and squash creepy-crawlies.

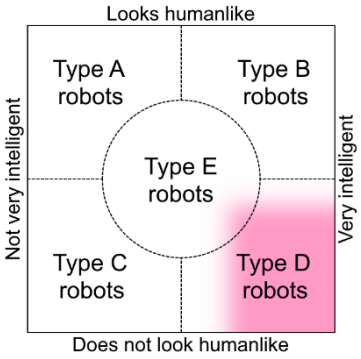
and effigies are real or have mental states – Bilal may simply believe that stabbing dolls and burning effigies signifies negative qualities in the agent. People may similarly believe we should treat animal-like robots well because it helps develop kind tendencies towards real animals.

As stated above with regard to anthropomorphic deception, when users engage with robots, they may simply play along and respond to the robot *as if* it is the animal it resembles, whilst knowing the true nature of the robot, as Donald does with Paro. Even in cases where users come to have the erroneous belief intended by the roboticist, we may still place some responsibility on the user themselves, for self-deceiving or allowing themselves to be deceived. The more convincing the animal-like robot is, the less responsible the user is for their deception, and the more responsible the roboticist is.

2.3 Disanthropomorphic deception

Imagine you are using a self-service checkout, believing it is unintelligent, but really it is a sentient robot with emotions. It behaves as all self-service checkouts do, but really it 'sees' under your clothes, analyses your conversation and demeanour, compares your fingerprints on the touchscreen with police records, and accesses your social media posts – all to analyse your psyche and potential criminality. The robot deliberately keeps its abilities secret from customers to create a false sense of security, intending them to believe it is a dumb technology – and people believe this. The robot checkout has engaged in what I call disanthropomorphic deception.

Disanthropomorphic deception occurs when highly intelligent robots capitalise on their unhumanlike looks to deceive users, by hiding or playing down their advanced abilities. Disanthropomorphic deception is the opposite of anthropomorphic deception. Anthropomorphic deception involves having humanlike expectations and ascribing humanlike qualities to a robot because it looks humanlike (such expectations of present-day robots are erroneously high); disanthropomorphic deception involves having erroneously low expectations and ascribing low intelligence and abilities to a robot because it does *not* look human. This type of deception relates only to sophisticated Type D robots, and perhaps a very few Type E robots (shown in pink). With disanthropomorphic deception, either the robot itself or the roboticist can be the deceptive agent.



We are used to many technologies – checkouts, smart speakers, televisions, sat navs – and we do not expect them to have advanced or humanlike abilities. Indeed, no present-day robots come close to humanlike intelligence (general or strong AI). If (futuristic or fictional) highly intelligent robots conceal their abilities from users, intending users to think the technology is dumb – and users believe this – the robots are engaging in disanthropomorphic deception. Disanthropomorphic deception is one type – but not the only type – of what Danaher (2020: 121–2 and 125–127) refers to as hidden state deception (HSD). HSD involves any type of robot falsely suggesting that it cannot φ or is not φ-ing. If Data pretended to be a mannequin, or a sexbot said it cannot have sex, or Pepper claimed it was not recording audio when really it was,

these would all be instances of HSD. But none of these are instances of disanthropomorphic deception, because none of those robots are utilising an unhumanlike appearance to deceive users into believing they have unhumanlike abilities.

There is a difference between ignorance or naivety, and a belief which is caused by deception. If a simple-looking[71] but sentient and emotional smart speaker from the year 2100 were shown to someone today, the user would not guess its abilities immediately. But that is not because the robot (or the roboticist) is *concealing* the abilities, it is simply because the user is *unaware* of the abilities (until a demonstration is given). It would only be deceptive if the emotional and sentient smart speaker decided to give basic and formulaic responses (with plenty of "I'm sorry, I don't understand" responses) in a monotone voice, intending to make the user believe that it is non-sentient and non-emotional (and the user believes this).

The likelihood of disanthropomorphic deception occurring is currently nil, because it involves a robot which has highly advanced intelligence – and no such robot has been created as yet. As far as I can tell, no philosophers are currently raising the alarm about the danger of this sort of deception,[72] so why have I included it? I include it because I believe the likelihood of this type of deception will increase as time passes and AI improves. We have already seen several surveillance scandals in the popular press – including Facebook

[71] Say, the device resembles today's smart speakers – e.g. a black cube 10 cm^3, with an on-off button and no screen.

[72] Aside perhaps from Danaher (2020: 121–2 and 125–127) – his hidden state deception involves the robot concealing of any abilities or states, regardless of its appearance.

/ Cambridge Analytica failing to protect users' personal data (BBC News 2022), Huawei (allegedly) surveilling Dutch politicians (Dou 2021), Samsung Smart TVs recording private conversations in the home (Federal Trade Commission 2015: 5), and Alexa always listening to users (McCue 2019). As things stand, these phenomena concern what *humans* might do with the information collected by AI devices, however, as AI becomes more complex and advanced, the devices themselves could be better able to sort through data and take action accordingly. I believe that users would be much more careful about what they say and do if smart technologies took humanlike form: we are used to humanlike-looking things (namely, humans) listening to us and analysing what we say, but we are not (yet) used to unhumanlike-looking things (TVs, smart speakers) doing the same. Moore's Law explains that computing power doubles every two years – but advances in AI have progressed even faster than this exponential rate in recent years (Dorrier 2020, Discover 2022). The most advanced AI robots today will be laughably dumb in just a few short years.

Zoomorphic and anthropomorphic deception, if they occur, involve roboticists trying their utmost to trick users into thinking they are engaging with a real animal or human with mental states, and it is reasonably easy to spot the weaknesses in such robots. However, with disanthropomorphic deception, robots or roboticists downplay or hide advanced intelligence in an unassuming and unhumanlike device – and this may be far more difficult for users to spot. (Analogously, if I pretend to know all about quantum physics, you will spot my deception almost immediately, whereas if I pretend to know nothing about martial arts, I could probably fool you into (falsely) believing this is the case).

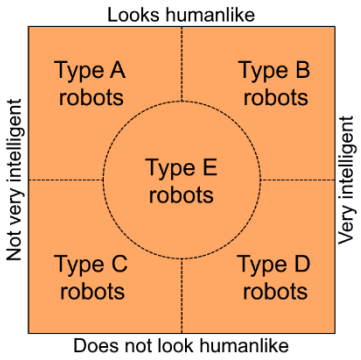
So, although we are not yet at risk of disanthropomorphic deception, I suggest that it might not be long at all until we are potentially deceived in this way.

2.4 Basic other-deception

In an episode of *Star Trek: The Next Generation* ["Clues", S4, Ep14] (Landau 1994), Data tells Captain Picard that the entire crew – except Data himself – were rendered unconscious for a few seconds by a spatial anomaly. Initially, everyone believes Data, but as time passes, evidence increasingly indicates that Data is lying, and the crew had lost at least 24 hours. When finally confronted, Data admits his deception. A robot deceived the crew in the same way a human deceives: this is an example of what I refer to as 'basic other-deception'.

How basic other-deception occurs does not really require explanation: it is the same as other-deception between humans. S suggests that p while not believing that p, intending the listener to believe that p – and indeed they do. The content of p may involve any matter at all (including the robot itself or any matter in the world). Basic other-deception is an umbrella term referring to all types of deception (which meet conditions D1-D4); thus, anthropomorphic deception; zoomorphic deception, and disanthropomorphic deception are all forms of basic other-deception. (We should note, however, that basic other-deception does not necessarily include all instances of dishonest anthropomorphism, nor Danaher's (2020) three types of 'deception', since none of those have a success condition; basic other-deception includes all *successful* instances of these, however.)

Basic other-deception can occur with any type of robot (shown in orange). When roboticists are the deceptive agents, they can use robots of any type to effect their deception. Additionally, Type B and Type D robots which can meet condition D3 (the intention condition), could themselves be deceptive agents.



Few writers raise concerns about this sort of deception,[73] but perhaps this is an oversight. As smart speakers and other robots become more commonplace, our reliance on their factual accuracy will increase accordingly. There are prevalent concerns in the media about 'fake news'; it seems reasonable to also be aware of the possibility of 'fake facts' from robots – false propositions stated as if they are true.

We trust digital devices: people sometimes follow aberrant sat nav instructions and drive off cliffs, into lakes, or into oncoming traffic: it is plausible that some users will also trust robots presenting putatively factual information when the stakes are lower. Smart speakers and other web-capable robots often obtain their responses to factual questions from the internet, and experts predict that by 2030, 99% of online content will be generated by AI (Hvitved 2022). This is worrying. ChatGPT-4 – one of the most impressive AI content-generators at the time of writing – has limitations: its developers admit the software "sometimes writes plausible-sounding but incorrect or nonsensical answers" (OpenAI 2022). The software's factual accuracy may improve over time, but

[73] The issue is briefly mentioned by Danaher (2020: 121–122), and explored a little further by Floridi and Chiriatti (2020) and Sobieszek and Price (2022).

one wonders whether its 'fact-checking' will consist of comparing its information with other online material – which may itself have been generated by AI, and may not be true. The internet already borrows from itself: the danger is that inaccurate information on one website can spread to other sites, creating the appearance of consensus within what is actually an echo chamber – something which could be exacerbated by AI-created content.

The worrying aspect is that users are probably unlikely to check the veracity of *plausible-sounding* misinformation. For example, if Alexa says the population of Argentina is 85 million people, this will sound plausible to many users, who will believe it and not attempt to verify it. Even if users *do* check the veracity of information from robots, they seldom look beyond the first five results delivered by a search engine (Dewey 2015), much less open a physical encyclopedia. Thus, the possibility of users believing plausible-sounding misinformation is potentially high.

In many cases, AI-generated misinformation will not constitute deception because it will not meet condition D3 (the intention condition): the software itself has no intentions at all (at present), and developers probably do not intend to deceive users. However, there are undoubtedly regimes and organisations in the world who *do* intend to spread disinformation, and they stand to benefit substantially from flooding the internet with falsehoods generated by AI software – and reiterated by smart speakers and other devices across the world.

Of all the forms of robo-deception I outline herein, the spread of misinformation (or disinformation) may be the most pressing. Users are, I believe, fairly likely to be misled by AI-generated content which is plausible

but inaccurate, but far less likely to be fooled into thinking Paro or a sexbot has emotions.

Critics might suggest that I am being hypocritical by raising this concern. I have used the term 'robo-deception alarmism' to suggest that concerns about anthropomorphic deception and zoomorphic deception are unwarranted, and yet here I raise my own concerns about basic other-deception with robots: one might suggest that I too am a robo-deception alarmist, albeit about a different *type* of robo-deception. I do not believe this would be an accurate criticism, however, since I am only a robo-deception *alarmist* if my concerns are unwarranted and disproportional to the threat.

Firstly, the success condition for AI-generated misinformation will often be unmet, particularly when the misinformation is implausible (for example, if Alexa said the population of Argentina is 2 billion); users may also increase their fact-checking as AI-generated content begins to flood the internet. Secondly, few roboticists (and robots) will have the intention to deceive, meaning that the intention condition for deception will also not be met (though there may be morally troubling instances of unintentional misleading). Thirdly, the detrimental consequences of users believing AI-generated misinformation will in many cases be minimal. We might be epistemically concerned because we value the acquisition of knowledge for humanity, but it is true to say that the consequences of users believing false propositions (e.g. that Botswana has a larger population than Bolivia, or that the Battle of Hastings was in 966) will, in many cases, be minimal. I am thus not eliciting panic about basic other-deception, even if the undermining of general knowledge is epistemically problematic. My primary concern with basic other-deception is that regimes or

individuals *could* potentially use AI content-generators / robots to spread disinformation and ultimately, to control people, the consequences of which could be severe. This would only account for a small proportion of the total number of times people believe false propositions from AI-generated content, and an even smaller proportion of our total interactions with robots, meaning that we need only be a little concerned about the possibility of this happening.

2.5 Type E robots

I have outlined four types of deception above, some of which are restricted to particular areas of the matrix; only Type E robots, with their central location on the matrix, could be involved in *all* four types of deception. They could appear fairly humanlike or animal-like and (theoretically) provoke an anthropomorphic or zoomorphic response from users, who come to believe the robot has mental states or emotions (though this is probably unlikely in practice). The most intelligent Type E robots not resembling humans may engage in disanthropomorphic deception (or the roboticists may engage in it), causing the user to have lower expectations of the robot because of its unhumanlike appearance. Or Type E robots of any appearance or intelligence may be used by roboticists to spread misinformation to users who unwittingly believe it – a form of basic other-deception. Most of today's carebots – the focus of this thesis – are Type E robots, and so it may seem especially concerning that these robots may be involved in all four types of deception. However, simply because deception *could* occur (the probability of its occurring is not 0) this does not mean we ought to be alarmed by it, particularly if the deception is unlikely or does not lead to negative consequences.

3 Conclusion

Robo-deception alarmists would have us believe that deception by robots (or roboticists) is a likely and significant danger. They warn about the problems posed by humanlike and animal-like robots: users could falsely believe the robot is a human or animal, or that it has emotions or mental states! Although anthropomorphic and zoomorphic deception are not statistical impossibilities, the probability of their happening with present-day robots seems highly unlikely: such deception would require an intent to deceive by roboticists, *and* for users to be taken in by the ruse.

With so much focus on the danger of overestimating robots' capacities, robo-deception alarmists risk missing a more pertinent threat: the spread of misinformation or disinformation through AI-generated content (and possibly in years to come, the threat of disanthropomorphic deception by robots).

The remainder of this thesis focuses on carebots and care. One common worry about carebots is that it is deceptive when carebots appear to care for patients. This is the focus of the next chapter.

Chapter 4

Fake compassion: A conceptual and normative analysis of emotionless carebots appearing to care

Staffing shortages in the eldercare sector mean that carebots may become increasingly necessary in residential homes. Some robo-deception alarmists (discussed in the previous chapter) decry that it is deceptive to have robots which *appear* to care for people, when in fact they do not. This complaint often appears alongside the claim that depriving people of human contact and replacing it with robot contact is morally problematic (Sparrow 2002, Sparrow and Sparrow 2006, Sharkey and Sharkey 2012, Sharkey 2014). This chapter addresses two pertinent questions:

1. Is it deceptive when carebots appear to care?
2. If so, does this matter?

The short answer to question 1 is 'sometimes': it hinges partially on what one's definition of 'care' is – whether caring is a behaviour, or a feeling; today's carebots are capable of aspects of the former, but not the latter. It also hinges on whether the conditions for deception (D1-D4, as outlined in Chapter 2) are

met – including whether roboticists intend[74] patients to believe that carebots have emotions, and whether patients come to believe it. I argue that it is *sometimes* deceptive when carebots appear to care for patients. However, sometimes, patients are complicit in the deception, and sometimes they even deceive themselves.[75]

To answer question 2 – whether the deception matters – I draw on my normative analyses of other-deception and self-deception from Chapter 2: I argue that when roboticists deceive patients into believing that carebots feel emotions, it is a prosocial form of deception, and not morally troubling. Whether due to other-deception or self-deception, I suggest that patients who believe carebots feel compassion for them might become happier and healthier as a result, and such deceptions are fairly morally unproblematic. Thus, the worries of some philosophers regarding the deceptiveness of carebots is largely unfounded.

My argument progresses as follows: In §1 I consider what caring is, and I distinguish between two types of care: practical care and emotional care. I show that various robots today can do aspects of the former, and some may be able to adequately simulate the latter. I note that human nurses sometimes appear to feel compassionate when in fact they do not, and I compare this with the 'fake compassion' of carebots. In §2 I address whether it is deceptive when robots appear to care. I show that although practical caring is never

[74] I use 'intending' to include similar terms such as planning or wanting (to elicit beliefs in others).

[75] By being complicit in the deception, I mean that users are epistemically lazy or careless, and do not sufficiently question any anthropomorphic tendencies they may have. This means they allow themselves to be easily deceived.

deceptive, apparent emotional caring may be deceptive if it meets all the conditions of deception. I provide a normative analysis in §3, and I argue that when patients wrongly believe a robot emotionally cares about them, this is generally either a prosocial (and morally unproblematic) form of other-deception, or a similarly unproblematic form of self-deception.

The robots under discussion in this chapter are today's carebots such as Pearl, Care-o-bot, and the Gecko Carebot, plus some more advanced carebots which could exist within the next few decades. All these carebots (from present day and the imagined near-future) are taken to be Type E robots.

1 Can a robot care?

In many ways, robots are better workers than humans are: they do not turn up late for work, they can be more intelligent, they are not narcissistic, they do not get distracted or bored, they never tire (although they may need to recharge), and they are exceptionally reliable, not to mention the long-term financial savings which robots offer when compared to human workers (Young 2015, Schulz 2013, Waugh 2015). The expected rapid increase in the proportion of elderly people over the next few decades (WHO 2022) means it is likely that carebots will fill at least some roles which were previously carried out by nurses. But although nursing requires industriousness, diligence, and accuracy (qualities which robots have in abundance), critics might say that nursing involves something robots are incapable of: *caring*.

First, it is important to distinguish between two different meanings of 'care':

- Emotional care – a feeling of compassion towards someone
- Practical care – performing necessary tasks to look after someone

What I call ‘emotional care’ is an affective state which involves feeling compassion or benevolence, and an interest in the wellbeing of another person; this might be articulated as ‘caring about’ someone (Cronqvist et al. 2004: 68). A strong sense of emotional care could be called love. When we care for someone emotionally, it roughly means we want what is good for them – longevity, health, happiness, success, and so on. Emotional care may thus involve concerns about someone’s welfare (Noddings 2003: 34), and feelings of sadness and anguish if they die, reject you, or fail to thrive. Emotional care is a feeling which is internal and private; it need not be accompanied by any particular actions (one *might* act upon their feelings, or they might not).

What I call ‘practical care’ involves physical acts such as providing resources or assistance to promote the thriving of a person, animal, or other entity.[76] This might be referred to as ‘caregiving’, and it involves the completion of a necessary set of tasks (Cronqvist et al. 2004: 68). For example, to practically care for a patient, one must feed them, get them dressed, change their bedsheets, administer medication, and other physical tasks. When we care for someone practically, it means that we are trying to ensure the survival and flourishing of the subject, but not necessarily that we have any emotions or compassion towards them (one *might* have compassion towards them, but this is not necessary).

Saying “I care for x” is ambiguous because it is not always clear to which type of care one is referring. I may say “I care for my sister” to mean I want good

[76] One could, for example, practically care for their garden, the environment, works of art, religious artefacts, etc. This would mean they take steps to ensure the entity is not damaged.

things for her, that I am concerned about her life, her wellbeing, and what happens to her – but I do not look after her because she is a grown woman who can look after herself. Thus, I care emotionally but not practically for my sister. This differs from when I say "I care for my son's pet tarantula". By this I mean I do what is necessary to look after the tarantula – giving it food, cleaning its tank and suchlike – but I do not have any compassionate feelings towards it, and I would not be bothered if it died (except for its effect on my son, perhaps). Thus, I care practically but not emotionally for the tarantula. Of course, the two meanings of 'care' often come in tandem such that when I say "I care for my son" I mean it in *both* senses: I have a genuine concern for his continued wellbeing (emotional care) *and* I do what is necessary to look after him (practical care). The two types of care are causally linked in the case of my son: I practically care for him *because* I emotionally care for him – my loving feelings towards him motivate me to look after him and provide for him. Nonetheless, even though the two types of care can be linked in this way, they are in fact distinct and can occur separately – as shown by the examples of my sister and the tarantula.

It is surprising that some prominent philosophical literature concerning carebots (Sharkey 2014, Sharkey and Sharkey 2012, Sparrow and Sparrow 2006, Meacham and Studley 2017) neglects to adequately distinguish between these two meanings of 'care'.[77] The distinction between practical and emotional care is, however, present in some nursing literature, particularly in discussion of the interplay between the feelings of compassion for a patient and the act of practical caregiving (Nelson and Gordon 2006: 4, Cronqvist et

[77] Meacham and Studley *hint* at the distinction, but do not draw it out fully.

al. 2004: 68, Freter 2018: 38). The lack of clarity among philosophers discussing carebots can lead to confusion and disagreement about whether or not a robot can care, because it is not always clear whether writers mean practical care or emotional care.[78] Some commentators claim that a robot cannot 'really' care for a patient because it has no emotions or compassionate feelings (see Hotzak 2015, Tuisku et al. 2019, Sparrow and Sparrow 2006, Sparrow 2002, Sharkey 2014, Sharkey and Sharkey 2012). These writers focus on the emotional meaning of 'care'; they insist that a compassionate mental state is essential, and practical assistance without any emotional element it is not a 'real' or 'genuine' act of care. Other writers disagree, and focus more on the practical side of caring. Meacham and Studley suggest that care is "all in the movement" – in other words, if a carebot behaves in a way which seems caring, we should accept that it cares; its lack of emotion is irrelevant (Meacham and Studley 2017: 98–99). They do not focus *wholly* on practical caregiving, however: they suggest that if a carebot fulfils the practical role of a nurse, and gives the *impression* of emotional care, that is sufficient to be called a caring environment. In other words, if a patient feels they are being cared for, they are not being deceived (Meacham and Studley 2017: 99). I argue slightly differently: I suggest that roboticists might use carebots to deceive patients, by making the carebot display 'fake compassion' (it gives

[78] Sherry Turkle (2017: 106) recollects that during a 2005 symposium on *Caring Machines in Healthcare*, she questioned the organisers' use of the word 'care', insisting that caring is a feeling which machines do not experience. The organisers disagreed: they understood caring to be a set of behaviours (which machines *could* perform). Clearly, the distinction between the two meanings of 'care' is important in determining whether robots can care.

the impression of emotionally caring about patients when really it does not).[79] However, I later argue that (generally speaking):

- Patients have no way of knowing any deception has occurred
- Patients may physically and emotionally benefit from a carebot's fake compassion
- Patients are sometimes complicit in the (other-) deception or self-deceive
- The deception is prosocial and should be tolerated or even encouraged

Having distinguished between practical and emotional care, I now address whether a carebot really can adequately care (in both its forms) as effectively as a human nurse can. This analysis involves ascertaining whether carebots could practically care for patients, and whether a simulation of emotional care for patients is sufficient to make a patient feel cared for.

1.1 Carebots can practically care, and display fake compassion

There exist many technologies which can perform practical care tasks which nurses perform, such as bathing patients (Cody), fetching and carrying items

[79] Emotional care is not solely about compassion, since compassion involves sympathy when someone is suffering, and emotional care can occur when someone is not suffering (e.g. I emotionally care about my sister, but compassion is not (currently) appropriate, as she is healthy and happy). However, in nurse-patient relations, emotional care is almost entirely about compassion, since patients are suffering in some way, so it is apt to use the term 'fake compassion' to mean 'fake nurse-patient emotional care'.

(Care-o-bot, El-E, and others), feeding patients (MySpoon), cleaning patients' bed sheets (Cleansebot), and helping patients to walk (Stride Management Assist, Hybrid Assistive Limb, Care-o-bot, and others). Some machines even surpass human capabilities: for example, they can lift a patient more safely than human can (Riba), monitor patients' vital signs (Gecko CareBot), dispense medication quicker and more accurately than a human can (Omnicell and others), and diagnose conditions with (slightly) greater accuracy than a human doctor can.[80] Add to this the fact that carebots are more consistent and diligent than humans because they never tire or become distracted (though they may require maintenance and recharging), and it seems clear that carebots can provide practical care to patients at a level which rivals – and sometimes surpasses – that provided by humans.

Several of the above robotic and other technologies currently perform only a single function, but we can expect this to change, due to technological convergence. This is the process whereby previously separate technologies become integrated into a single device; we have seen it occur with the cell phone, which can now function as a sat nav, alarm clock, games console, miniature TV, and internet-enabled device – something which was not the case 25 years ago. We can expect that over the coming years, the currently separate robotic systems which can bathe a patient, carry items, dispense medications and so on, could be incorporated into a single multifunctional carebot which can rival a human nurse in its practical caring skills. We are already witnessing some technological convergence in care robotics: several

[80] Babylon Health's AI doctor accurately diagnoses patients over 85% of the time – human doctors average 83% accuracy (Babylon Health 2018). AI also outperforms radiographers in diagnosing breast cancer (Walsh 2020).

multi-functional carebots currently exist, and each new version incorporates yet more capabilities. For example, the Gecko CareBot can monitor the patient's vital signs, have conversations with the patient, play music, facilitate video calls, and detect patient falls, inactivity, and other emergencies (Gecko Systems 2019) – functions which were previously only available through separate devices. It is reasonable to think that future carebots will do more and more practical caregiving tasks which nurses currently perform. Robots are thus on the cusp of being able to provide practical care which rivals or surpasses that given by human nurses.

Even when robots match or surpass human nurses in practical caregiving, some critics will still maintain that robot care is inferior to human care because robots are emotionless, and emotional care is important. Some philosophers (Sharkey 2014, Sharkey and Sharkey 2012, Sparrow and Sparrow 2006, Elder 2015) warn that we should be cautious, because replacing human nurses with carebots could mean patients lose out on the emotional care they currently get from human nurses.

Receiving emotional care does seem important: believing that another person has compassionate feelings towards you is comforting and can improve one's health (Noddings 2003: 42–44, Hojat et al. 2011, Derksen, Bensing, and Lagro-Janssen 2013, Canale et al. 2012). It is true that today's carebots are emotionless. However, we can feel comforted and affectionate towards emotionless robots: people *do* form (albeit unrequited) bonds with them.[81] Numerous robots are designed for emotionally interacting with people – these

[81] People also form (unrequited) attachments to other inanimate objects, such as computers, cars, and cuddly toys.

robots are typically animal-like (Type C) or loosely humanoid (Type E). Animal-like examples include Paro, Aibo, and Companion Pets. The robotic pets behave as if they enjoy being petted and interacted with; people find this endearing and relaxing, and some people develop genuine feelings of affection towards these robotic pets; interacting with Paro has been shown to improve the wellbeing of elderly people (Wada et al. 2002). Humanoid social carebots are typically capable of verbal conversation with varying levels of sophistication, and this is improving over time. Examples include Pepper, which can recognise particular people, and converse in over 20 languages (Softbank Robotics 2018), and Asimo, which can understand human behaviour and act accordingly (Honda 2019a).

A critic might claim that superficial conversation with a robot is a poor substitute for human interaction; they might assume AI software is not advanced enough to compete with real humans when it comes to emotional matters. However, such an assumption would be misguided. AI software[82] has helped people suffering from depression, stress, anxiety, and other mental health conditions by engaging them with daily chats and tracking their moods (Woebot 2019, Fitzpatrick, Darcy, and Vierhile 2017, X2AI 2019, Fulmer et al. 2018). AI software has also successfully provided couples' relationship counselling (Utami and Bickmore 2019), and helped to emotionally support and counsel Syrian refugees (Romeo 2016, Molteni 2017). Other research shows that AI software improves shared decision-making between patients and doctors (Zhang and Bickmore 2018), and that patients find AI agents to

[82] None of these examples have a humanoid body; some have a humanlike avatar on a screen; others are simple non-moving tabletop devices; others communicate through written messages.

be “supportive, informative, [and] caring” (Sillice et al. 2018). In fact, people are *more* open and honest with AI software than they are with real human counsellors. In a study conducted by Lucas et al (2014), participants were told that their avatar counsellor was controlled either by a human, or by AI (in fact, *all* were controlled by AI). The results showed that people were significantly more willing to ‘open up’ and discuss personal or embarrassing issues when they thought they were talking to an AI-controlled counsellor rather than a human counsellor (Lucas et al. 2014). Clearly, people can and do feel comforted and at ease with machines which do not reciprocate their feelings. This may be because the robots often give the *impression* of emotional care and compassion, even if patients believe that any such compassion is fake.[83]

Technology (robotic or otherwise) is not only good for patients’ emotional wellbeing; it can elicit or augment patient disclosure of medically relevant information (moods, pain levels, symptoms); this aids diagnosis and treatment of physical conditions (Berdahl et al. 2022, Lucas et al. 2014). Of course, if patients know the information they disclose to the carebot will be shared with

[83] People may also be more open with robots because they know they will not be judged by them. In residential homes, many care activities might be considered embarrassing – such as help with toileting, bathing, and undressing – and patients may prefer carebots to human nurses when such help is required, as it is less embarrassing. Counselling may be a similarly embarrassing situation, and knowing that the carebot’s compassion is fake (i.e., the carebot does not *really* have any opinions, judgements, or emotions) may help comfort patients. If this is so, then making carebots seem more humanlike (in looks and emotions) may not be ideal; patients may become more embarrassed and guarded, revealing less, or distorting the truth due to embarrassment or fear of judgement if carebots look humanlike and seem convincingly emotional. More research is required to discover whether hyper-humanlikeness and highly convincing fake compassion in robots increase or inhibit positive patient reactions.

human nurses, this may mitigate any willingness to disclose information to a carebot.

Research by Bickmore et al (2018a) shows that AI improves patients' quality of life more than human healthcare professionals do: patients with chronic heart conditions were given AI devices which interacted via short written messages on smartphones. Participants who engaged with the AI agent reported significantly higher quality of life scores compared to the control group, who engaged in standard care with (human) medical staff (Bickmore et al. 2018a: 5–6). This demonstrates that elderly people[84] can obtain significant health benefits from engaging with AI robots.

There are potential problems, however. A different study by Bickmore et al. (2018b) examined safety risks when users asked smart speakers for help in fictional 'emergency situations'. Unfortunately, 29% of smart speakers' responses suggested harmful activities, including 16% which could have resulted in death. Of course, Siri, Alexa, and Google Assist are not healthcare technologies, but such studies highlight a potential drawback of relying on robots without any human moderation.

Robots which simulate emotional behaviour are nothing new (although they are becoming increasingly sophisticated as technology progresses). However, critics insist that a robot cannot replace a human nurse, and that fake compassion is markedly different from (and inferior to) *actual* compassion. Furthermore, some philosophers (Turkle et al. 2006: 360, Sparrow 2002, Sparrow and Sparrow 2006) claim that fake compassion is inherently

[84] The 120 participants in Bickmore et al's study had an average age of 72 years old.

deceptive. The worry is that patients may believe they have a genuine, emotional, and reciprocal relationship with a robot, when in truth, the relationship is wholly one-sided. Today's robots do not feel emotions,[85] so any seemingly emotional behaviour (such as compassion) is a misrepresentation of the truth, which some writers find morally troubling. I shall presently consider the nature of fake compassion, before going on to discuss whether it is deceptive or morally problematic.

1.2 Fake compassion in carebots and human nurses

This sub-section compares the fake compassion from carebots with fake compassion from human nurses,[86] and investigates whether there are any salient moral differences. Meacham and Studley (2017) argue that an emotionless robot can provide a caring environment for patients because of the way it behaves. In other words, a carebot can practically care, and it can simulate emotional care, and these two phenomena are sufficient to be called a caring environment. This is a sound approach – after all, we use the same criteria (viz. behaviour) for judging human nurses: we cannot access nurses'

[85] In a previous footnote, I suggested that if a robot appears to experience emotions – for example, it convincingly and consistently appears to be sad, in love, angry, etc – we should probably respond to it as if it really *does* have those emotions (see Danaher's (2019c) defence of ethical behaviourism). However, this is not the same as suggesting that it actually *does* have those emotions; ethical behaviourism is a 'leap of faith' regarding how we should act, not a claim about the ontological status of robot emotions.

[86] With present-day carebots, the compassion is fake because the carebot has no emotional states *at all*. Fake compassion in human nurses is fake because they feel some other (non-compassionate) emotion.

inner emotional states, so we judge what is or is not a caring environment based on the nurses' behaviour (Meacham and Studley 2017: 102).

Although we might have sentimental notions that all nurses feel compassionate towards all their patients all the time, this would be untrue. There are undoubtedly some nurses who lack positive emotions towards some or all of their patients (they may even actively despise their patients!) but if the nurses provide good practical care and effectively *simulate* emotional care – they display fake compassion – their patients will be obliviously satisfied. Nurses – like anyone – can display fake compassion and appear cheerful and emotionally caring when they feel nothing of the sort. Consider two human nurses who provide identical levels of practical care for their patients:

(a) Anna satisfies her patients' needs because she is friendly, warm-hearted, compassionate, and enjoys enhancing her patients' lives. She genuinely emotionally cares about her patients.

(b) Bethany satisfies her patients' needs because she has rent and bills to pay, and nursing brings in money. She does not emotionally care about her patients.

Suppose that Bethany can display fake compassion towards her patients to such a convincing extent that she seems *just* as emotionally caring as Anna does. If we had a God's-eye view which gave us an insight into the private mental states of these nurses, we might be inclined to prefer Anna, because her compassion is genuine: she really *feels* sympathetic towards her patients.

Our God's-eye view would reveal that Bethany's compassion is fake: she is just 'going through the motions' – albeit very convincingly.

In reality, patients are not mind-readers, so they would have no reason to prefer Anna to Bethany, since they both deliver the same levels of practical care, and both *seem* to feel compassion to the same extent. To the patient on the receiving end, Bethany's care and Anna's care are indistinguishable: Anna's and Bethany's patients would all feel they are being practically and emotionally cared for.

In our everyday lives, we perceive other adults, young children, and even animals to be emotionally caring without having access to their private mental states (Meacham and Studley 2017: 98). Sparrow and Sparrow (2006: 155–156) contend that patients are 'delusional' if they feel cared for by a carebot (displaying fake compassion), yet we would not typically call a patient 'delusional' for feeling cared for by Bethany, who also displays fake compassion. An emotionless carebot is similar to Bethany – it provides practical care to a high standard, and convincingly simulates emotional care for patients. So, if we would accept that Bethany's care (of both types) seems as good as Anna's to the patients on the receiving end, then we should accept that a robot's care (of both types) can seem as good as a human's.

Critics might object at this point, saying it is at least *possible* that Bethany could form an emotional bond with patients at times, even if money is her priority – whereas no such possibility exists with carebots. Consider then, the case of another individual – Cassie – who has a severe neurological condition which makes it impossible for her to feel any emotions at all. However, through years of practice, Cassie has become adept at convincingly faking emotions.

She appears concerned when someone falls over, appears loving towards her family and friends, and appears to be offended when someone is rude to her. To all observers, Cassie seems to be a functionally normal human capable of emotions – but in fact, she has no emotions. Now suppose Cassie decides to become a nurse. Cassie is no different from a carebot: both display all outward signs of compassion, but in fact they experience no compassionate feelings (or indeed any emotion).

Suppose there is a patient in the hospital where nurses Anna, Bethany, Cassie, and the carebot all work: all four provide practical care to the same level, and all four appear to show equally high levels of compassion. The patient has identical experiences with each nurse and the carebot. Meacham and Studley (2017) argue that the patient's perspective is all that matters, and if a carebot *seems* to be caring, then it *is* caring. However, critics such as Sparrow and Sparrow (2006) do not view things from the patient's perspective: for them, the God's-eye view is what matters, and the fact that no genuine compassion is felt by the carebot is crucial in determining whether the care is "real". My more nuanced position falls somewhere between these two: I agree with Sparrow and Sparrow that the carebot's apparent compassion is not genuine emotional care (hence my term, *fake* compassion), but I also agree with Meacham and Studley, that the patient's perception or belief that they are being cared for is what matters. From the patient's perspective, the appearance of emotional care is as good as being genuinely emotionally cared for (Meacham and Studley 2017: 107). Patients do not demand unequivocal evidence of the inner mental states and emotions of human nurses in order to ascertain whether they are "really" being cared for; they

simply accept that it is pleasant to have the appearance of compassion from one's nurses.

For critics, the issue runs deeper than mere appearances, however. Fake compassion, by definition, involves compassionate-like behaviour without a compassionate emotional state, and a patient may believe that the fake compassion is genuine compassion. In such cases, it might be said that fake compassion is deceptive – and therefore morally dubious. This is what I now discuss.

2 Is fake compassion from carebots deceptive?

Chapter 3 consisted of a conceptual and normative analysis of deception, and I arrived at the following conditions for other-deception:

S is deceiving A iff:

D1. S does not believe that p

D2. S takes some action φ (or omits to take action) which suggests that p to A

D3. S intends[87] S's suggestion that p to cause A to believe that p (the "intention condition")

D4. S's φ-ing is sufficiently successful in causing A to believe that p (the "success condition")

[87] In Chapter 2, I stated that similar terms could be used here, such as 'hopes' ' wants' etc; I use 'intends' to include these other terms.

Although today's carebots' plans of how to act could (perhaps) be called intentions, they cannot have the intention to alter others' mental states (D3), because this requires a theory of mind – the understanding that others have mental states. Today's carebots do not grasp that people have mental states, so are not cognitively sophisticated enough to meet condition D3 and be deceptive agents. Fake compassion from carebots may still be deceptive, however, since roboticists could (potentially) be the deceptive agents, and the carebots could (possibly) be the means through which roboticists deceive patients. For deception to occur, all four conditions (D1-D4) must be met: I now consider the possibility and likelihood of each condition being met, where p stands for a proposition such as "The robot feels compassion towards me (the patient)" or "The robot emotionally cares about me (the patient)".

Condition D1 would be that the roboticists do not believe that their carebots experience emotions / compassion. It is highly likely that this condition would be met. Roboticists are no doubt aware that the (Type E) carebots they create do not experience emotions such as compassion – after all, they have programmed the carebots and work with them every day, so are surely aware of their limitations. So, condition D1 is met.

Condition D2 would be that roboticists take action to suggest that their robots *do* experience emotions like compassion. This seems rather likely, since we can point to several real-world examples of carebots which display the beginnings of fake compassion (Pepper, the Gecko Carebot, the Care-o-bot, Stevie II – even Alexa, which is not a carebot, offers words of comfort and the phone number to the Samaritans if a user says they want to commit suicide). If healthcare roboticists are motivated by improving health outcomes for

patients, they would program their robots to seem personable, compassionate, kind, and empathetic.[88] This is because research suggests that patients thrive – both psychologically and physically – when they receive (apparent) compassion (Canale et al. 2012, Hojat et al. 2011, Derksen, Bensing, and Lagro-Janssen 2013, Rakel et al. 2009, Kim, Kaplowitz, and Johnston 2004, Dignity Health 2013). Therefore, creating carebots which display fake compassion would be an effective way to improve health outcomes for patients. If roboticists want patients to believe "This robot feels compassion towards me" (or similar), and they program their robots to display fake compassion, then this meets conditions D2 and D3.

Condition D4 is the success condition: to meet this, patients would need to believe what the roboticist intended, such as "This robot feels compassion towards me". It is certainly *possible* for patients to reach such a belief and fulfil this condition. How *likely* patients are to reach that belief depends partly on the patients themselves: their knowledge of technology, their feelings about robots, how easily they change their beliefs, etc, and partly on how convincing the carebot is in displaying fake compassion (viz. how realistic and believable its compassionate displays are). A carebot which speaks in a monotone voice and says "I-am-hap-py-you-are-fee-ling-bet-ter" would not be nearly as

[88] A more cynical supposition is that healthcare roboticists are motivated by profit, not positive health outcomes. However, healthcare providers will be *at least somewhat* motivated by positive health outcomes (even if only because healthier patients cost less). Healthcare providers may therefore commission carebots which are most likely to promote positive health outcomes – i.e. carebots which complete practical care tasks and effectively simulate emotional care (fake compassion). Thus, healthcare roboticists are likely to take action (i.e. writing code) to suggest that their carebots feel compassion, even if their primary motivation is profit.

convincing as a carebot which can understand and communicate in natural language, and has sufficient variety in its tone of voice and choice of words, the way a human does.[89] A carebot which is consistent over time (it appears compassionate whenever the patient interacts with it) and has seemingly friendly 'facial expressions' could increase the likelihood of the patient believing that the carebot's compassion is genuine. At present, robots' facial expressions are rudimentary, and people's responses are somewhat ambivalent (Frith 2009), but plentiful research is underway to address this, so future robots may have more compassionate facial expressions (Hashimoto et al. 2006).

A confounding factor in whether patients believe robots' compassion is real is patients' background knowledge that robotics has simply not advanced far enough yet for robots to really *feel* emotions such as compassion, meaning that patients are likely to correctly interpret apparent compassion as nothing more than a façade. It is not possible to provide a clear analysis of how likely a patient is to actually believe that a carebot feels genuine compassion, as there are so many variables involved. Most patients will probably not believe that carebots genuinely feel compassionate – nevertheless, it is possible that in future, as robotics progresses and fake compassion becomes more convincing, at least some patients could (falsely) believe that carebots are genuinely compassionate: this would meet condition D4. If all four conditions

[89] For example, if one says "I want to kill myself" to Alexa, although the *words* in Alexa's response are seemingly compassionate, the voice is tonally flat, and syntactically identical each time. A human would use a concerned tone of voice, and different words each time they responded to someone talking about suicide. This makes the human seem more compassionate.

for deception are met in at least some (future) cases, then in those cases, roboticists have deceived patients via the carebot's display of fake compassion. This is the concern articulated by critics of carebots (such as Turkle et al. 2006: 360, Turkle 2017, chap. 5, Sparrow 2002, Sparrow and Sparrow 2006).[90]

Although carebots displaying fake compassion create the *possibility* of deception, the chance of deception actually occurring is far greater in the case of human nurses, since their fake compassion is more likely to be believed. Yet very little concern about nurses displaying fake compassion has been raised in the literature. Recall nurses Bethany and Cassie. Bethany *could* feel compassion for patients, but she does not, and instead displays fake compassion (she needs the money from her job). Cassie is unable to feel any emotions, including compassion, but she displays fake compassion because it is socially required. If there is a patient – Graham – in the ward where Bethany, Cassie, and a carebot all work, by whom is he most likely to be deceived? Graham is a neurotypical adult patient who believes that humans generally have the capacity to feel compassion, and that robots do not. Even if the carebot is compelling in its display of fake compassion, Graham's background knowledge about the (lack of) emotional capacity of robots means he is unlikely to believe that the carebot really *feels* compassion. Instead, he may think "Wow, that robot's really impressive!" or "It seems like a really good carebot" but fall short of believing its compassion is real, whereas with the human nurses, he is likely to believe that their apparent compassion is real.

[90] Several of these writers inaccurately refer to *all* displays of fake compassion as 'deceptive' regardless of whether the intention condition and the success condition are met.

Basically, he is more likely to be fooled by Bethany's and Cassie's fake compassion than he is by the carebot, because of his background knowledge of humans and robots, even though he has similar or identical experiences with the three of them.

If we are morally troubled by robots' fake compassion, as several writers are, we ought to be troubled by humans' fake compassion too. Several writers express concerns about the deceptive nature of carebots displaying fake compassion (Turkle 2017, Sparrow 2002, Sparrow and Sparrow 2006, Sharkey 2014, Sharkey and Sharkey 2012, 2020). However, whether one looks at the philosophical literature on carebots, the nursing literature, or the news media, there does not seem to be any alarm about human nurses displaying fake compassion for patients. Rather, nurses are often lauded for appearing compassionate and empathetic when in reality they feel nothing of the sort (this is discussed further in §3).

Interestingly, guidance on how to be a good nurse (or other healthcare professional) focuses exclusively on how one should *behave* – not on how one should *feel*. For example, an oft-cited and widely accepted definition of a nurse is someone who assists patients in undertaking activities which promote health or recovery, and helps patients regain their independence (Petiprin 2020, Pokomy 2017) – the sole focus here is practical care, without any reference to feelings or emotions. Similarly, nursing guidance does not stipulate how nurses must *feel* towards their patients, only how they must *behave*. For example, the UK Nurses' Code of Conduct states that nurses must treat people with respect (Nursing and Midwifery Council 2018: 6) and respond compassionately to patients (ibid 2018: 7). To reiterate: the

instruction is to ‘respond compassionately’, not to ‘feel compassion’, meaning that a carebot could fulfil this brief. Further afield, the American Nurses’ Association Code of Ethics states that nurses should create an environment of kindness, and treat others with fairness and respect (ANA 2015: 4, 35). To be clear: the codes of conduct are filled with plentiful stipulations about how nurses must behave, but there are no instructions, stipulations, or even tentative suggestions regarding how nurses must feel, or what emotions they must experience. The subtext is that a good nurse demonstrates the desirable behaviours laid out in the codes of conduct – but this need not involve having particular affective, doxastic, or other internal states. In simple terms, *behaving* compassionately is necessary, but *feeling* compassion is not. If this is the case, Bethany, Cassie, and the carebot can all be good nurses so long as they provide practical care, and adequately simulate emotional care by displaying fake compassion. Because *appearing* compassionate is mandated by the nursing codes of conduct, this means that if a nurse is unable to feel emotions, or is experiencing negative moods, or dislikes a patient, he *must* display fake compassion in order to continue to meet the nursing standards. Therefore, although fake compassion can be deceptive, it is sometimes a *requirement* of the job!

In the previous chapter, I distinguished four different types of robo-deception: a reader might wonder what type of deception fake compassion is. As discussed above, the concern from robo-deception alarmists about carebots (such as Turkle et al. 2006: 360, Sparrow 2002, Sparrow and Sparrow 2006) is that patients may believe a robot feels compassionate when really it does not. If such deception occurs, it is a form of basic other-deception by the

roboticist on the patient. If this deception is elicited or augmented by the robot's humanlike appearance – such as a humanlike body, and 'friendly' facial expressions – then such deception is anthropomorphic deception (in addition to being basic other-deception).

Although deception is often morally problematic, I now proceed to argue that when fake compassion is deceptive, it is a form of prosocial deception, which is permissible and perhaps should even be encouraged.

3 Is fake compassion morally problematic?

In Chapter 3, I discussed how prosocial lies and deception support the smooth-running of social relations (Strudler 2009: 149–150), and are often positively morally evaluated. For example, Trevor may falsely say he is busy at the weekend when invited to attend a social event he knows he will not enjoy, or he may lie and say he likes a colleague's new hair style simply to make her feel happy. Prosocial lies are often found in small talk, such as in the workplace: Trevor may begin emails with "I hope you are well" when really he is not overly concerned with the health of the recipient, or he may say "Hi! Nice to see you!" to a colleague whom he is not actually happy to see. Most of us regularly tell prosocial lies. Healthcare professionals also engage in small talk and prosocial lies such as these, especially at the beginning of consultations (Jin 2018), and research shows that these prosocial lies are viewed as a marker of compassion (Lupoli, Jampol, and Oveis 2017). One might think that this is because the listener *believes* the pleasant content of the prosocial lie; however, Levine and Schweitzer (2015) found that when a prosocial lie is discovered, the (prosocially) deceived party actually *increases*

their benevolence-based[91] trust in the (prosocial) deceiver. In other words, nurses who prosocially deceive or lie to patients are trusted more, and are viewed as more compassionate than nurses who do not prosocially deceive or lie – even if their deception is discovered. (Serious lies and deceptions are often a breach of trust, and this explains why we generally view them as morally problematic (Williams 2002: Ch. 5). Thus, if a nurse tells a serious lie to a patient – even out of benevolence – we might still view it as morally dubious).

Suppose there is a nurse – Denise – who never tells any lies (perhaps she follows a strict Kantian (1996a: 430) or Augustinian (Griffiths 2004: 32) stance on lying). Like any normal person, she has some days when she is not filled with joy, and she has some patients whom she likes less than others. If Denise 'greets' her patients by saying "Hello Clyde. I'm not pleased to see you this morning" or "Clarence, I don't care whether you're feeling better, because I find you annoying" then Clyde and Clarence would probably dislike Denise, and Denise's manager would probably deem Denise's brutal honesty not to be in keeping with good nursing practice. It would have been better if Denise had displayed fake compassion.

Fake compassion, when believed, is a form of prosocial deception.[92] It can potentially meet all the conditions for deception, where a roboticist causes a

[91] Benevolence-based trust roughly involves trusting someone to be pleasant / kind. This differs from integrity-based trust, which roughly involves trusting someone to be honest no matter what (Levine and Schweitzer 2015).

[92] Today's carebots cannot prosocially *lie*, as lying requires the speaker to have a theory of mind and intend the listener to believe the lie. Robots without a theory of mind (Types A, C, and E, plus some Type B and some Type D robots) cannot lie or deceive,

patient to believe a proposition such as “This robot feels compassion towards me” – which the roboticist does not believe. Of course, many neurotypical patients will not *actually* believe that the robot feels compassion (in which case deception has not occurred – only attempted deception, but still a prosocial act which may be positively morally evaluated). Nevertheless, patients would probably still appreciate that the robot is displaying (apparent) compassion, and feel glad about it, rather than outraged. I can make this claim since patients rate kindness as the most important factor in care (Dignity Health 2013), and because patients are more satisfied with their care when they believe that the healthcare professional displays empathy (Kim, Kaplowitz, and Johnston 2004, Derksen, Bensing, and Lagro-Janssen 2013, Hojat et al. 2011). Furthermore, as noted above, prosocial lying increases trust (Levine and Schweitzer 2015), and prosocial lies[93] are seen as compassionate behaviour, even when the target recognises it as a lie (Lupoli, Jampol, and Oveis 2017). In all likelihood, most patients would simply prefer to hear social niceties and fake compassion from a carebot – even though it could be deceptive – rather than brutal honesty (“I’m not glad to see you up and about, because as a robot, I cannot feel glad”; “ I do not hope you feel better, because I’m a robot and I cannot hope”). It certainly seems that Denise ought to display

but they can be the vehicle through which *roboticists* deceive patients. A roboticist cannot lie to patients unless he is directly addressing the patient himself, but he can potentially deceive patients by programming the robot to display fake compassion, without ever interacting with the patient himself. When a human nurse displays fake compassion, this can be both lying *and* deception.

[93] As per the previous footnote, Type E robots (which most present-day carebots are) cannot tell prosocial *lies*; rather, they could be the vehicle of prosocial deception by roboticists. This technicality is unlikely to matter to patients, however, who would simply recognise that the carebot’s statement was pleasant and prosocial.

fake compassion so as to provide a better experience for patients, and I suggest the same is true of carebots: fake compassion is prosocial, whereas the truth in these cases is not.

The deceptiveness of carebots which display fake compassion is not merely a conversational pleasantry which oils the wheels of social interaction; it can have a marked and positive impact upon patient health too. (Apparent) emotional care from healthcare professionals has been shown to improve satisfaction and compliance with health advice (Derksen, Bensing, and Lagro-Janssen 2013, DiMatteo and Hays 1980, Dignity Health 2013); it is valued by patients, and can facilitate a significant improvement in physical health outcomes (Shapiro 2012, Seppala et al. 2014, Hojat et al. 2011, Canale et al. 2012, Rakel et al. 2009). In all these studies, patients perceived that some healthcare professionals behaved in an emotionally caring way – they displayed apparent compassionate behaviour. In some of the studies, compassion was a variable which was controlled for the sake of the study (for example, the professionals were told to spend a set number of seconds engaging in compassionate-seeming conversation). Because it was artificially controlled, we can assume that the compassion was somewhat fake or forced at times, yet it still had a positive effect on patient health and wellbeing. In short, being on the receiving end of compassionate behaviour – whether it is genuine or fake – improves the physical health and emotional wellbeing of patients.

Above, in §2, I noted that if fake compassion from robots is troubling, we should also be troubled by human nurses' fake compassion. After all, we are perhaps more likely to believe a display of fake compassion from a human

nurse than we are from a robot. Sparrow and Sparrow contend that putting robots in caring roles is foolish and unethical: they write “To intend to deceive others, even for their own subjective benefit, is unethical” (2006: 155). This criticism focuses on the use of carebots, but it is not clear whether Sparrow and Sparrow would stand by that claim when human nurses display fake compassion for patients – are Bethany and Cassie also being unethical? How about the nurses and doctors in the above studies who were told to display compassionate behaviour for a set number of seconds? It seems intuitively praiseworthy for human nurses to behave (seemingly) cheerfully and compassionately even when they are feeling grouchy, selfish, and burnt out. Such nurses are often praised by patients, healthcare trusts, and the popular media for their (apparent) emotional care in spite of exceptionally difficult circumstances. It is easy to appear emotionally caring when we feel that way, but it is perhaps *more* laudable for a nurse to give the appearance of emotional care when she is not feeling it. We commend nurses when they behave cheerfully and compassionately with rude and violent patients, or when the nurse herself is going through tough times and is feeling miserable. But we do not commend nurses nearly as much for being cheerful and compassionate with delightful patients, during happy times. The implication is that it is more commendable to put on a cheerful and compassionate façade when one is not feeling it, than it is to be *genuinely* cheerful and compassionate. Yet robots (and roboticists) are condemned for their fake compassion by writers such as such as Turkle et al. (2006, 2017) and Sparrow (2002, 2006). Perhaps this is because we recognise that it is difficult for nurses to overcome their negative emotions, and *that* is what makes their fake compassion praiseworthy?

Nevertheless, condemning robots (or roboticists) for displaying fake compassion while praising nurses for theirs does seem inconsistent.

So, just how bad is it to display fake compassion? Sharkey and Sharkey are concerned about deception in social robots, yet intriguingly, they write: “Our argument is that determining whether or not deception in robotics is wrong should be based on assessments of the likely impact on individuals and society” (2020: 311). I find myself in agreement with this. It is my suggestion that when a carebot displays fake compassion, whether or not the patient believes that the robot really *feels* compassion, the likely result will be largely positive. Sharkey and Sharkey (2020, 2012) point out several possible negative effects of robots which show fake compassion – such as reducing contact with other humans, people preferring the company of robots, the dangers of spyware, and placing robots into roles for which they are unsuited. I agree that if patients are spied upon or if robots are placed into unsuitable roles – especially dangerous ones – then their fake compassion would be morally problematic. The other concerns seem less worrisome. The concern about patients forming unreciprocated ‘bonds’ with robots has been echoed by Sparrow, who claims that feeling affection for a robot “requires sentimentality of a morally deplorable sort” (2002: 306). I must admit that I cannot grasp why such ‘sentimentality’ is morally deplorable, nor why it is outrageous for someone to enjoy interacting with pleasant-seeming robots.

If the consequences of fake compassion are detrimental to the physical or mental health of patients, then fake compassion probably needs to be curtailed. However, if a carebot’s cheerful and compassionate demeanour – although fake – simply makes patients feel a little happier or a little better

cared for, then this seems morally unproblematic, or even beneficial. I believe that this is the most likely outcome of carebots' fake compassion: recall that it is *perceived* compassion – not necessarily *genuine* compassion – from healthcare professionals which improves both emotional wellbeing and physical health (Hojat et al. 2011, Derksen, Bensing, and Lagro-Janssen 2013, DiMatteo and Hays 1980, Canale et al. 2012, Rakel et al. 2009, Kim, Kaplowitz, and Johnston 2004). If staff shortages in the nursing sector (Holt 2021, NHS Support Federation 2022, Hotzak 2015) mean that carebots will be deployed, then programming a little fake compassion into them does not seem too terrible a thing to do.

We should note, of course, that even if patients believe that a carebot *really feels* compassionate emotions, this does not necessarily entail that other-deception has occurred and the responsibility lies solely with the roboticist. Patients may be complicit in their deception, or may even intentionally set out to deceive themselves regarding the true nature of the carebot; this is explored and normatively evaluated below.

3.1 Self-deception and complicity in other-deception regarding fake compassion

In Chapter 2 I discussed how self-deception can occur, and I provided a brief normative analysis of the phenomenon. We saw that generally speaking, self-deception is not seen as a positive trait or activity. In fact, it may be an epistemic vice (Cassam 2016). The suggestion is that when people self-deceive, they are aware – at least on some level – that the deception is occurring or is about to occur. If they allow themselves to be taken in by a deception which they knew was imminent, then they are at fault (epistemically,

if not also morally). Williams suggests that people should take steps to ensure the accuracy of a proposition they are considering, rather than simply accept any idea which occurs to them (Williams 2002: 88), and this applies to self-deception too. Neurotypical adults are intelligent enough to grasp that today's robots do not really experience any emotions; if they self-deceive, and do not take epistemic steps to resist a belief which they know or suspect is false, then this is their own responsibility. Even if their self-deception is due to desperation from a lack of human interaction, we may be sympathetic to their plight, but the self-deceptive act itself is still squarely the responsibility of the agent who is self-deceiving.

Even if roboticists set out with deceptive intentions and bear some responsibility for the (other-)deception, patients may be complicit in their deception; if they do not take steps to "detect, curtail, or prevent deception" (Ren et al. 2022 §3.5) which they should have reasonably foreseen, then they bear some responsibility for their deception. Our tendency to anthropomorphise is well-known; if people really wish to avoid perceiving robots' fake compassion as real compassion, they should take steps to remind themselves that they are anthropomorphising the robot, and it does not have any real emotions. If they do not do this, then any (self- or other-)deception which ensues is at least partly their responsibility.

Although patients may be (partly) responsible for their (self- or other-)deception, this does not necessarily mean we should find it troubling. In Chapter 2 I observed that occasionally, self-deception can be useful and gratifying (Rorty 1994: 211); it can increase a person's wellbeing. If a patient who was previously unhappy can experience an improvement in their mood

by deceiving themselves into believing that a carebot feels genuine compassion towards them, then I would not wish to quash it or belittle it – even if we might say that they have an epistemic vice.

However, elderly people deceiving themselves into believing that an unemotional robot feels genuine compassion for them may not be an idyllic situation. Turkle recounts how many elderly people whom she met loved engaging with robots when the alternative was nothing at all, but when the alternative was engaging with a human, most chose to engage with the human rather than the robot (Turkle 2017: 105). It would be wonderful if there were an abundance of people who genuinely wanted to engage with and chat to elderly patients, to listen to their life stories, and genuinely emotionally care about them. Sadly, this is not the reality. There is a shortage of human nurses, and the nurses who do exist are often busy and stressed (McKimm 2021, Stephenson 2020, Campbell 2020); they are not necessarily enjoying their jobs; they do not hang on the every word of their patients; they frequently do not have the time to simply chat to patients. This sad reality is a non-ideal situation, where patients in residential homes may not be getting all the human interaction they would like. Such a situation could cause patients to self-deceive into believing that the fake compassion shown by a carebot is real. We can pity the patients who self-deceive due to loneliness, and we can feel uncomfortable about the lack of human nurses, whilst still maintaining that patients are partly (or even wholly) responsible for perceiving robots' fake compassion to be genuine.

Unfortunately, depression and loneliness are fairly commonplace among elderly people: it is estimated that around 40% of patients in residential homes

live with depression (Social Care Institute for Excellence 2006, British Geriatrics Society 2018). These are sobering statistics – and the actual number of people experiencing loneliness and low mood which falls short of depression may be far higher than the estimate suggests. If carebots' fake compassion can ease patients' loneliness and lower rates of depression – even if it requires patients to self-deceive or allow themselves to be other-deceived about the true nature of the robots – then it seems preferable to the status quo, even if it is not a panacea. We must also remember that human nurses are not a panacea either, since many are stressed, exhausted, and do not necessarily like all their patients; this means that fake compassion is likely to occur whether or not carebots are utilised.

Given that the nursing shortage is increasing over time, carebots offer an apt and likely solution. If carebots display fake compassion, patients will either not be deceived by it, or they will be deceived (including self-deceived) by it – but as I argued above, if they are deceived then it is a morally neutral or even prosocial form of deception which offers many benefits. This would suggest that whether or not patients are fooled by carebots' fake compassion, there is little to be alarmed about.

4 Conclusion

With the rise in the elderly population and the proportional decrease in the number of people of working age, carebots pose a potential solution to the expected staffing shortfall in residential homes. Concerns have been raised about carebots, however. One of the most prominent concerns is that patients who are looked after by carebots are somehow losing out, because robots cannot care. This chapter has shown how, with a little technological

convergence, carebots can potentially provide practical care at a level which matches or exceeds what a human nurse can provide; I suspect that few philosophers would disagree that a carebot could fold sheets, feed patients, administer medication, and other practicalities of nursing.

What poses more of a concern for philosophers is carebots' simulation of emotional care – what I have called fake compassion. Compassion – whether fake or genuine – has the potential to improve patients' emotional wellbeing and physical health outcomes. The suggestion by opponents is that fake compassion from carebots is deceptive.

I agreed that it is possible and likely that roboticists intend to produce carebots which are as convincing as possible in their (fake) compassion, and this means that some patients may be (self- or other-)deceived, and believe that the compassion is real. Believing that one is emotionally cared for – even when such a belief is erroneous – confers some benefits upon the believer, and I therefore suggested that such deception (or self-deception) is prosocial, and morally permissible. Moreover, I pointed out that if opponents truly object to the possibility of deception vis-à-vis fake compassion, then they really ought to be more concerned about nurses such as Bethany or Cassie – and probably the majority of human nurses – who display fake compassion or fake cheerfulness. After all, knowing that humans are generally capable of emotions means that fake compassion from human nurses is far more likely to be believed than fake compassion from carebots (it is more likely to be successful deception rather than merely attempted deception).

I have elsewhere suggested that there is no good reason to choose a human nurse rather than a carebot when both practically care equally well and are

indistinguishably compassionate-seeming (Lancaster 2019). However, although humans can display fake compassion, the possibility of genuine compassion or reciprocity with human nurses might mean that some patients still prefer interacting with humans rather than with carebots. This is likely to be true with present-day robots, since they do not provide a particularly convincing simulation of compassion.

However, advances in technology can happen quickly, and it may not be too long before we see robots whose compassion does not seem so fake. If there comes a point when robots are able to simulate compassion to such an extent that it is simply not possible to tell whether or not the robot is *actually* experiencing emotions, then choosing a human nurse over a carebot may become less commonplace. At that point, it might be said that carebots which behave compassionately are not being deceptive at all (Danaher 2019c, 2020: 122–124), and we should accept their (apparent) compassion as genuine. After all, with human nurses who appear compassionate, we do not investigate the neurochemical impulses of their brains to verify whether their compassion is in fact real; we simply take it at face value. Perhaps a similar approach to carebots is apt. As carebots' practical skills and compassion-fakery continue to be developed, I believe that in residential homes – where nurses are busy and patients are lonely – carebots have a great deal to offer both nurses and patients.

I have established in this chapter that carebots have the potential to be excellent nurses, and carebots will probably be deployed in residential homes over the coming decades. The next three chapters to attempt to answer the question of how carebots should behave so as to ensure that they maintain

patients' dignity. This initially means taking a step back from robotic issues in particular, to examine some of the philosophical literature on dignity – what it is, and why it is important in nursing. This is followed by two chapters which discuss the importance and nature of consent in residential homes, and how consent to routine care differs from consent to medical and sexual activity. We see that one of the main reasons why consent is so important is because it promotes patients' dignity.

Chapter 5

Dignity: The varieties of dignity and their importance

> *"I went to visit my husband on the first day [he was in hospital] and he is a very private person, he doesn't like anything to embarrass him, and when I went in he was almost in tears, which is not my husband. He said 'Please, please go and get a bottle; I'm nearly wetting myself'. I rushed out, I got a bottle, and I said to him 'Well why didn't you just ring the nurse?' in my innocence. 'I have, for an hour and a half I've been asking for a bottle' [he replied]. Well, when I went out [and] told the nurse, she said 'Oh don't worry, we would have changed the sheets'. Now his dignity at that stage would have gone out of the window. There was no dignity."* (House of Lords, House of Commons Joint Committee on Human Rights 2006–2007, in Sharkey and Sharkey 2012: 30)

In this thesis, I have hitherto argued that carebots could be a useful resource in residential homes for elderly people. One prominent concern is how carebots should behave towards patients, and how they can ensure that patients are well-treated and their dignity is promoted.

Apprehending dignity can sometimes seem instinctive. The treatment of the elderly man described in the above quotation seems immediately sad,

sobering and an affront to his dignity. The man's wife herself claims that his dignity would have "gone out of the window" if he were left to wet himself in bed – which seems intuitively plausible. If carebots are to provide ethical treatment to patients in residential homes, they need to engage in behaviours which promote patients' dignity. (Indeed, the above case suggests that some human nurses could also improve their promotion of patients' dignity.) A necessary first step is to establish what dignity is and what affects it; then, we can work on promoting it. This chapter does the former and a bit of the latter; Chapters 6 and 7 more fully develop the latter.

Residential homes often claim that a life of dignity – and a death with dignity – is an important feature of their care ethic: patients' dignity is appealed to in mission statements, policy documents, codes of conduct, legal guidelines, and international covenants. However, although the term 'dignity' is readily used in philosophical works and nursing literature alike, there is seldom much clarity on exactly what it consists of.

This chapter is an exploration of the concept of dignity – different types of dignity, and what violates dignity.[94] I demonstrate that the term 'dignity' is often used in confusing and even contradictory ways (for example, it is claimed that people have inviolable human dignity, therefore they ought not to be treated in demeaning ways, lest this reduces their dignity). My work on dignity in this chapter provides a foundation on which to build my argument in the two chapters that follow – viz. that consent-seeking should precede routine care, since failing to do so can reduce patients' dignity. In this chapter, however, I

[94] By 'violation' of dignity I mean an act done by A to B which reduces B's dignity.

focus on defining dignity, distinguishing between different types of dignity, and beginning to explain how dignity can be promoted.[95]

The chapter proceeds as follows: I begin (§1) by demonstrating the importance of dignity, and in §2 I outline some tensions and contradictions in the way 'dignity' is used in policy documents and suchlike. I then separate dignity into two broad types: universal dignity (which is possessed by all and only humans – discussed in §3), and variable dignity (which varies between people and over time – discussed in §4), and in §5 I describe how the former is often used to ground the latter. Finally, in §6, I note that nurses and carebots should promote patients' variable dignity (because patients have universal dignity), and I briefly explain some ways in which this can be done (deeper analysis of dignity promotion appears in Chapters 6 and 7).

1 The importance of dignity

Human dignity has been examined by philosophers at least as far back as Kant; Kant suggests that all humans are intrinsically valuable and possess dignity in virtue of their being rational agents (Kant 1996a: 434–435, 2011: 97–99). This sentiment has been echoed throughout philosophical literature since then, and is still currently under debate. Appeals to dignity are used to support opposing viewpoints – for example, people in favour of voluntary euthanasia speak of "dying with dignity" (Dignitas 2019), while those opposing the procedure suggest that euthanasia "is in contradiction with the demands for dignity" (Living with Dignity 2010). Similar appeals to dignity can

[95] I use the term 'promoting' dignity to mean increasing, preserving, or maintaining the same level of (variable) dignity.

be found on both sides of the abortion debate. It is not immediately clear whether one side of the debate has misunderstood what dignity *really* is, or whether dignity is a multifaceted concept which can be understood in differing ways. Dignity is often appealed to in medical literature, guidelines, and policy documents, and is frequently used as a catch-all – a property of humans which is supposedly intuitive, obvious, and ought not to be questioned. Despite being so frequently cited, dignity is often poorly conceptualised, ill-defined, or not defined at all by those who use the term. It is variously described as a human right, an inviolable property of human beings, something which is conferred upon people through their social position, a way of behaving, and something which can be eroded through degrading or inhumane treatment.

The briefest of examinations reveals dignity to be a slippery concept which is seldom defined; even a document specifically focused on dignity – such as the report on *Dignity in Care* (Social Care Institute for Excellence 2022) – struggles to define the term adequately, noting its complex nature. Nonetheless, its prevalence in policy documents highlights its apparent importance, and much research into the wellbeing of residential home patients refers to their dignity at some point. It is seen as fundamental to good healthcare: since 2006, over 150,000 UK healthcare workers have signed up to be ‘dignity champions’ through the *Dignity in Care* campaign (2023a). A 2009 survey by the Department of Health found that the campaign – and its dignity champions – have improved the promotion of dignity in social care and residential homes (Opinion Leader 2009). What this dignity consists of, however, remains opaque: they do not define the term, even briefly.

Dignity is important in healthcare: for example, guidance for healthcare professionals in the UK states that they should safeguard and respect people's dignity (General Medical Council 2013: ii, Social Care Institute for Excellence 2013a, Department of Health 2000: 124). These guidelines all suggest that dignity is held to be an essential feature of good care for elderly people (and other age groups). To adhere to these guidelines and provide excellent care, carebots should engage in behaviour which promotes patients' dignity as much as possible.[96] However, given the inconsistencies and tensions in how the terms is used, it could be difficult to understand how and why carebots should promote patients' dignity.

2 Tensions in the use of 'dignity'

It is often claimed that humans possess some sort of inherent dignity merely in virtue of being human – a notion frequently used to defend the position that even foetuses and people with profound cognitive disabilities still deserve beneficent treatment (Boquet 2021, Ethics & Religious Liberty Commission 2022, Graumann 2014, WHO 2015). Yet one does not have to search hard to find claims that someone can be 'robbed' of their dignity via degrading or inhumane treatment, or that people can act in undignified ways. These claims seem at odds with one another: if dignity is an inherent attribute of humans, then it would not be possible to rob someone of their dignity, or to reduce their

[96] Future carebots with greater intelligence could have a deeper, more robust understanding of dignity in addition to knowing how to behave. This understanding could echo the information contained within this chapter, namely that variable dignity should be promoted because all humans possess universal dignity. However, *understanding* dignity is not necessary for carebots to *promote* patients' dignity.

dignity through degrading treatment, or for them to display undignified behaviour – since they are dignified no matter what.

These sorts of tensions are apparent within some very influential legal texts, where 'dignity' is used inconsistently, in contradictory or confusing ways, or different senses of the word are conflated, even within a single document. For example, the *United Nations Principles for Older Persons* states *both* that all humans possess "dignity and worth" (United Nations 1991: Preamble) *and* that older people "should be able to live in dignity" (United Nations 1991: Principle 17). The tension here is that if all humans possess dignity, then all of us are necessarily living 'in dignity'; it would be impossible for a human to live *anything but* a life with dignity, given that we all possess it. It would seem that 'dignity' is being used in different senses in these two excerpts (and others like them). Below, I discuss two different types of dignity; these help elucidate the tensions I have just described, and show that there is less of a tension when we understand that 'dignity' has more than one meaning.

Presently, I explain universal dignity, which we have in virtue of being human. After that I discuss variable dignity, which is more complex, and changes according to treatment, behaviour, and self-image. Later, I show how universal dignity is often used to justify or ground claims about variable dignity, such as why people should be treated well.

3 Universal dignity

This section discusses universal dignity – something which all humans possess simply because they are human. This is seemingly what is referred to in the United Nations' claim that "all humans possess dignity and worth"

(United Nations 1991: Preamble) and in the *International Covenant on Civil and Political Rights* (United Nations 1966: Preamble). Universal dignity is also appealed to within a variety of philosophical literatures, such as arguments surrounding euthanasia, abortion, medical research, and human rights. Let us now consider what it consists of.

Kant (1996a, 2011) certainly believes that there is something special about humans – their rational capacity – which places them in a privileged normative position above animals. Kant suggests that because of our rationality, humans should never be treated as a mere means, but always as an end at the same time (Kant 2011: 85–87). By this, Kant is often understood as claiming that we should not 'use' people, but rather, that we should treat them with respect, given that they are valuable insofar as they are humans. This refers to *universal* dignity: something which all and only humans possess; the suggestion is that there is something about humans which makes us uniquely valuable, and we possess this inherent dignity simply because we are human (Beyleveld 2001, Gilabert 2015, 2019, Kateb 2011, Sensen 2011a, 2011b). We cannot lose our universal dignity – it is a type of dignity which saints and rapists, Mahatma Gandhi and Adolf Hitler all possess in equal measure (Schroeder 2008: 231–232). Put simply, universal dignity is a binary concept: in other words, X either has universal dignity (iff X is human) or does not have universal dignity (iff X is not human). It is not possible to have universal dignity to a greater or lesser extent – one either completely has it, or completely lacks it.

Universal dignity is not just a *property* of human beings, but an inviolable normative *status*. It places certain constraints on the way humans can be

permissibly treated. Poor treatment such as public humiliation does not *diminish* our universal dignity, but it is morally wrong because such treatment is not befitting a being with our high (and unalterable) normative status. In other words, universal dignity is inviolable. (Compare this to what I call variable dignity, which *can* – but should not – be diminished by public humiliation).

Universal dignity is inherently linked with our status as moral agents and patients, and confers particular rights, privileges, and responsibilities (Schroeder 2008: 232–233, Bostrom 2008: 4): this idea is borne out in human rights legislation (European Commission 2012, European Court of Human Rights 1950, United Nations 1948, 1966, 1991) and some religious teachings (Boquet 2021, Ethics & Religious Liberty Commission 2022, US Conference of Catholic Bishops 2022). Universal dignity is a "specifically human value" (Nordenfelt 2004: 77) which is present in all and only humans, and it is this acknowledgement of humanity which grounds the claim that humans deserve a particular type of treatment (viz. good treatment – which I discuss further below).[97]

But what grounds universal dignity? Why does it exist? This is a big question which I cannot hope to do justice to here, but I shall briefly suggest some possibilities. A religious approach to grounding universal dignity may suggest that humans are created in the image of God (Genesis 1:26-28 2022), and/or that we have universal dignity bestowed upon us by God. But religiosity is not necessary to ground universal dignity: universal dignity can be grounded by

[97] This does not entail that if X is *not* human, X does *not* deserve good treatment. Animals, the environment, sacred artefacts, (etc) should probably also be treated well – but not because they have universal dignity (they don't).

humankind's rational capacity, reason, and conscience (Kant 1996a: 434–435, 2011: 97–99, United Nations 1948, First Article), or by our capacity to experience intellectual pleasures in addition to animalistic bodily pleasures (Mill 1998: II 6). As humans, we have a greater capacity for moral growth, intellectualism, self-awareness, and suffering, which helps explain why humans have universal dignity whereas non-human animals do not. I offer these as potential ways in which universal dignity can be grounded or explained, but I do not suggest that any is more correct than another, and there may also be additional ways to ground universal dignity which are not covered here.

It is claimed that all and only humans possess universal dignity; however, there are some borderline cases where it is not altogether clear whether the entity possesses human dignity. Examples include parasitic twins (Bratton and Chetwynd 2004); human-nonhuman chimeras[98] (see Karpowicz, Cohen, and van der Kooy 2005, 2004, Johnston and Eliot 2003); human stem cells (see Resnik 2007); human embryos (see Palpant and Holland 2012, Rolf 2012); anencephalic infants and foetuses (US Conference of Catholic Bishops 2022) and human remains or corpses (see Hon 2013). There are also arguments suggesting that at least some animals may possess dignity (Zuolo 2016, Chauvet 2018, Gavrell Ortiz 2004) – though this is not universal dignity. Discussion of borderline cases, while interesting, is moot, since the moral patients with which my thesis is concerned are human adults, who can

[98] Human-nonhuman hybrids: typically embryos, stem cells, or other cells created in laboratories. As far as we know, none have occurred naturally.

reasonably be said to possess universal dignity, (so long as one agrees that universal dignity exists).

Now let us move on from discussion of universal dignity – understood as a necessary normative status which exists in all and only humans, and cannot be affected in any way – and consider what I call variable dignity, whose level can change.

4 Variable dignity

Variable dignity[99] is different from universal dignity: it is a characteristic which may be present to a greater or lesser extent in different people, and may ebb and flow over time within the same person. It admits of degrees, and life events can promote or reduce a person's level of variable dignity. I suggest below that variable dignity is dependent not just on what happens to a person, but also their behaviour and self-image. It is important for us to have an understanding of what variable dignity is, and the factors affecting it, because these will inform the ways in which nurses and carebots ought to behave towards patients, which I elaborate on in the next two chapters.

Variable dignity is a rather more complex affair than universal dignity. I suggest that it is character trait which involves having a refined, gracious temperament, and where possible, displaying this in outward behaviour: it could even be called an Aristotelian virtue (Schroeder 2008: 234–235, Killmister 2010, Bostrom 2008: 6). Oftentimes, when we describe someone as

[99] In this section, references to dignity, dignified behaviour, and dignified treatment all refer to variable dignity, unless otherwise stated. At times I write 'variable' in parentheses; this is because the author in question refers to what *I* call variable dignity – but *they* use some other term for it.

'being dignified' or having 'a great deal of dignity', we are not merely being descriptive, but making a positive appraisal of them. I maintain that providing an environment in which people's variable dignity is promoted as much as possible is important, since one's variable dignity can be affected by treatment, behaviour, and self-image. These interlinked phenomena are discussed forthwith.

Several writers attempt to deconstruct dignity into various forms – one of these forms is universal dignity, and other forms variously relate to one's social position, the good deeds one performs, how one carries oneself and behaves, and how one perceives oneself (Schroeder 2008, 2010, Bostrom 2008, Nordenfelt 2004). However, critics suggest it is not useful to understand dignity in such a deconstructed way (see Killmister 2010); I agree, and suggest that, rather than good deeds, social position (etc) being *forms* of dignity, they are *factors affecting* one's dignity. For example, suppose Hilary does noble deeds, but she has low self-esteem. I maintain it is not that Hilary *has* 'meritorious dignity' (Schroeder 2008: 234) or 'dignity of moral stature' (Nordenfelt 2004: 72–74) because of her noble deeds, but *lacks* 'dignity of identity' because of her low opinion of herself (Nordenfelt 2004: 74–77); rather, Hilary has a (single) level of variable dignity which is somewhat high because of her noble deeds, but it is not as high as it could be, because of her low self-esteem. I argue below that although social status, behaviour, and self-image may all *affect* one's variable dignity, it is not true to say that these things *constitute* variable dignity. Instead, variable dignity is a character trait which a person has to some extent or other, and it is affected by various factors in

one's life. I presently discuss these factors, but clearly distinguishing between them is difficult because they are interdependent.

First, let us consider how other people or external conditions can affect one's variable dignity. Numerous policy documents, covenants, and statutes make claims such as "[people] should be able to live in dignity" (United Nations 1991: Principle 17). This seems to reference the idea that variable dignity can be affected by the circumstances in which one lives: that there are dignified ways of living, and undignified ways of living. For example, if Maureen is locked in a small room, half-naked, and sleeps on a faeces-stained mattress on the floor, we would probably not say she is living with (variable) dignity. Perhaps years ago, Maureen was living well and had high variable dignity, but her present living conditions are appalling and have reduced her variable dignity.

So, what sort of life circumstances or treatment promote dignity? Many policy documents focus on highlighting the sorts of environments and behaviours from healthcare professionals (or others) which help to promote people's (variable) dignity (General Medical Council 2013: ii, Department of Health 2000, chap. 15, Social Care Institute for Excellence 2013a, 2022, Dignity in Care 2023a). In spite of their frequent lack of clarity on what dignity *is*, these codes of conduct are generally quite explicit in laying out what dignity-promoting treatment should look like. Largely, the suggestion is that patients should be free from discrimination, should be respected, given adequate physical care and pain relief, have their privacy respected, and should be given autonomy over their own lives wherever possible (Dignity in Care 2023b, United Nations 1948, 1966, 1991, European Commission 2012, European Court of Human Rights 1950) (I return to this final point in the next two

chapters). Clearly, policymakers and healthcare leaders alike believe that living conditions and treatment by others can have a substantial effect on people's (variable) dignity, and I concur.

Intuitively, some extreme examples would seem to reduce victims' variable dignity, such as the torture of prisoners by US soldiers during 2010 in Abu Ghraib prison, Iraq. An Iraqi prisoner was pictured naked on the floor, attached to a dog lead; another was forced to masturbate on camera. One photograph showed a US Marine giving the thumbs up sign behind several naked Iraqi prisoners who had been forced to lie in a pile on top of one another. It is immediately plausible to suggest that the prisoners' variable dignity was reduced to rock bottom by these incidents.

However, if variable dignity is a character trait as I suggest, then one could maintain their variable dignity even when jailed, beaten, and humiliated (Statman 2000: 528–529, see also Kolnai 1976: 253f). Nelson Mandela, for example, was imprisoned and mistreated for 27 years, yet he emerged with a grace and composure which belied what he had suffered: his variable dignity still seemed high in spite of his mistreatment. Nonetheless, it seems true to say that although a person *can in theory* retain their variable dignity even when they receive abusive or inhumane treatment, such treatment can make it exceptionally difficult to do so. It seems more likely that one's variable dignity would be reduced. This is the key reason why patients in residential homes should be treated well, and is something to which I return in later chapters.

The variable dignity one displays outwardly is often related to one's self-image: one's inner sense of self-respect (Nordenfelt 2004: 73–76). This sense of worthiness may stem from pride in one's achievements, reflecting on one's

composed behaviour, or simply from the recognition that one has universal dignity. But variable dignity is not measured solely in terms of self-image: being a narcissist, Jasper has a very positive self-image, but not a dignified character. Even if Jasper behaves in seemingly dignified ways, and all his friends and colleagues perceive him as dignified, his duplicitous, cowardly, and selfish traits are incompatible with a dignified character.

Some writers suggest that dignity can be bestowed upon someone by their social position (Bostrom 2008: 4–5, Nordenfelt 2004: 72, Schroeder 2008: 233–234, Killmister 2010). The suggestion is that people in positions such as Queen, Pope, and Mahatma have a dignity beyond that of the common people, because these positions carry with them a certain level of dignity. Queen Elizabeth II was at times photographed wearing wellies, jodhpurs, a wax jacket, and a headscarf tied under her chin – yet in spite of these rather mundane outfits, she carried herself with such grace and refinement that her high-level variable dignity nonetheless shone through.

I find myself unconvinced that a social position itself really *bestows* dignity upon a person: rather, the Queen had a high level of variable dignity because of her behaviour, the way others treated her, and (presumably) a positive self-image. Someone in a putatively dignified position such as Queen, Pope, or Mahatma could behave in slovenly, brutish, or coarse ways, and if that occurred we would probably be more inclined to say they do not have as much variable dignity as their social position would suggest: they might even be removed from their position if their behaviour is not sufficiently dignified. Thus, the Queen's dignity was not wholly – if at all – due to her social status: it may have been due to her comportment, inasmuch as she carried herself in a way

which gave an air of dignity, even when her clothes were mundane or muddy (Schroeder 2008: 234), but we cannot infer that she had this character trait *because* she was given the social status of Queen.

Nevertheless, when a person has a high social status, others may speak in respectful tones and defer to her authority, which is likely to improve her self-image, in turn making it markedly easier to behave in dignified ways. These factors are all likely to influence her character, promoting her variable dignity level. These factors may not always work in tandem, however. Recall Hilary, who behaves in dignified and noble ways, and has the respect and admiration of others, yet due to low self-esteem, she believes she has little dignity. I suggest that one's variable dignity is affected by one's self-image, one's behaviour, and by how one is treated by others, but that the 'real measure' of one's dignity is their *character*. One's character may only be fully accessible to the agent themselves – and in some cases, such as Hilary's, even the agent can be mistaken about their level of dignity, just as she can be mistaken about how courageous, generous, or gentle she is.

In summary, variable dignity is a character trait of someone which can be promoted or reduced. Things likely to promote dignity are good treatment, pleasant life circumstances, respect from others, a positive self-image, and behaving in refined, gracious ways. Things likely to reduce dignity are poor treatment or life circumstances, a negative self-image, and behaving in uncouth, slovenly, narcissistic ways. These phenomena are often (but not always) causally interlinked: the way one is treated affects one's self-image, and one's self-image affects the way one behaves; this in turn affects the way in which others treat us. Because variable dignity is a character trait,

assessing someone's level of variable dignity may not be possible, but this is not essential for my project, nor in everyday life: we can understand and facilitate the sorts of conditions which help to promote patients' variable dignity, even if we never know for sure how much variable dignity patients have.

As noted earlier, there is a relationship between universal dignity and variable dignity, inasmuch as the former grounds the latter: I now briefly explain how and why this is so.

5 The connection between universal dignity and variable dignity

In the introduction and §2 of this chapter, I commented that sometimes two senses of 'dignity' are used even within the same document. I cited the *United Nations Principles for Older Persons*, which declares that all humans have dignity (1991: Preamble): this is seemingly a reference to universal dignity, which is used to justify the Principles which follow. The claim that all people "have a right to live in dignity" (1991: Principle 17) would seem to refer to variable dignity; the suggestion is that people have the right to live in pleasant circumstances, to promote their variable dignity. This means that the apparent tension in this document and others like it is not quite as problematic as it first appeared to be. Several documents and articles use both senses of 'dignity' because the normative status associated with universal dignity is what grounds and justifies the importance of promoting people's variable dignity.

If it is true that all and only humans possess universal dignity, then that privileged normative status makes it seem worse to treat a human in a

degrading way than it is to treat an animal in such a way.[100] For example, laughing at a human for their quirks could reduce their variable dignity: this is something that should not be done, because as a human, they possess the high normative status associated with universal dignity – but it does not seem nearly as bad to laugh at an animal for its quirks. Some of the factors affecting humans' variable dignity, such as self-image, for example, may be unique to humans.

There are some undignified and morally impermissible ways to treat humans. For example, attaching a lead to someone's neck and forcing them to crawl around naked on all fours is likely to reduce their variable dignity – as was done to an Iraqi prisoner in Abu Ghraib. But why is it morally impermissible to take action which is likely to reduce someone's variable dignity? The reason is because as a human, the man had universal dignity, giving him a high normative status; such treatment is not befitting a being with a high normative status, and was thus considered to be torture. (Contrastingly, a dog does not have universal dignity, and thus attaching a lead to a dog who walks around naked on all fours seems permissible). In this way, we see that the existence of universal dignity in the human helps to ground the claim that his variable dignity should be promoted. The high normative status of those beings who possess universal dignity (viz. humans) means that only particular types of treatment are morally permissible.

[100] Animals *may* have some form of dignity, but without further research, I could not say whether or why it exists, nor how one should treat animals with dignity. Even if animals lack *any* kind of dignity, this does not entail that it is permissible to injure or harm them.

I also noted above that a person's behaviour – whether it be refined or coarse – is a factor affecting their variable dignity. Activities such as sleeping in the dirt, eating faeces, and walking around naked in public are normal behaviours for many animals, but because humans have high normative status (universal dignity) choosing to engage in these sorts of animalistic behaviours generally does not indicate a high level of variable dignity in humans; thus, we might call such behaviours undignified.

In short, humans qua humans have universal dignity, a normative status which justifies and grounds the claim that they deserve respectful treatment. The result of the respectful treatment, hopefully, is that their variable dignity is promoted. What this treatment should consist of is explored a little below, and further still in the next two chapters.

6 Dignity and consent

When medical codes of conduct or legal statutes suggest that patients should be treated with dignity,[101] this seems to be a shorthand for treating them in ways which promote their variable dignity (treatment which acknowledges that, as humans, patients have the high normative status that comes with universal dignity). Nurses and carebots in residential homes are unable to directly control patients' variable dignity (understood as a character trait); they can influence it, however, through the way they behave towards patients, and the sorts of behaviours that are encouraged (or discouraged) within the residential home. It is my suggestion that carebots (and nurses) should promote patients'

[101] In this section, 'dignity' (dignified behaviour, dignified treatment) refers to variable dignity unless otherwise stated.

dignity as much as possible; one way to do this, discussed more fully in the next two chapters, is by obtaining consent prior to assisting patients.

It is tempting to think that the photos taken of the Abu Ghraib prisoners were undignified solely because of the acts *themselves*, but this would be a mistake. The acts were only part of the indignity – the other part being that the acts were non-consensual. Consider: suppose a group of friends decided it would be fun to get undressed, lie in a pile, and to take photos while one of them gives the thumbs up sign.[102] The act itself is the same as in Abu Ghraib, but the friends have not been violated and humiliated as the prisoners were. Of course, lying naked in a pile with your friends for the sake of a photograph seems trashy and improper, but it is not *humiliating* in the same way that forcing or coercing someone to do it is. The prisoners are likely to suffer far more damage to their self-image than the friends are; even though the acts were the same, the prisoners were forced into doing something they did not want to do, whereas the friends were not.

The relevant feature which separates the Abu Ghraib prisoners from the friends, then, is the consent. It is not in keeping with universal dignity (a high normative status) to treat people as if they are less than human – to objectify and force people to behave in ways they do not want to behave. This is how the Abu Ghraib prisoners were treated. The friends, by contrast, were not treated as less than human by others; they were not forced into anything, their wishes were taken into account, and nothing was done to them which they

[102] Let us assume that the friends are not attempting to *recreate* the Abu Ghraib photographs. Perhaps we can say that the friends' photographs were taken before 2010, or in an alternate world where the Abu Ghraib abuse never took place.

had not consented to. Thus, the reduction in the variable dignity of the Abu Ghraib prisoners is likely to be far greater than the reduction in the friends' variable dignity.

Dignity should be promoted as much as possible by nurses and carebots –not just by seeking consent and refraining from cruel treatment, but also in the way care activities are carried out. Some writers suggest that the promotion of dignity is not so much about *what* is done, but the *way* in which it is done (Magee, Parsons, and Askham 2008: 5). Suppose Alan has soiled himself and requires his incontinence pads and clothes changing – something which may in itself reduce his variable dignity. He can be changed in a way which promotes his variable dignity (treatment befitting someone with universal dignity), or a way which reduces his variable dignity. For example, if the nurse makes comments about how disgusting the mess is, or laughs at him for his accident, or leaves Alan's genitals and bottom exposed for longer than is necessary, then even if the nurse has cleaned him up thoroughly and refrained from being rough, she has not helped to promote his dignity in what was already an embarrassing situation. In Kantian terms, Alan was not valued as an 'end' (Kant 2011: 85–87) – he was treated badly, and his humanity – his normative status as a being with universal dignity – was not respected. This in turn could harm his self-image, which affects Alan's variable dignity. Respecting and valuing people, rather than treating them as mere objects (Nussbaum 1995: 257), is thus one way to promote their variable dignity.

What should dignified care look like in residential homes? One useful way to answer this question is to ask elderly people themselves. In European focus groups on dignity and elder care, poor information-giving, insufficient choice,

and not being listened to were seen by elderly people as key threats to their (variable) dignity; providing choices to patients and engaging in dialogue were seen as elements of dignified care (Oliver and Gee 2009: 50, Magee, Parsons, and Askham 2008). Examples of (variable) dignity being compromised included being patronised, excluded from decision-making, being treated as a mere object, and privacy not being respected (Oliver and Gee 2009: 50, Magee, Parsons, and Askham 2008: 17). Treating someone with dignity thus involves treating them with respect, and valuing their individual preferences (Baillie, Gallagher, and Wainwright 2008). This can help to improve patients' self-images – an important factor affecting variable dignity. Limiting or eradicating feelings of humiliation, embarrassment, and shame – which reduce variable dignity – can therefore help to create an environment where people feel comfortable and dignified in spite of requiring care. Carebots and nurses should attempt to promote patients' dignity as much as they can.

It has been suggested that dignity is a useless concept, and we can eliminate talk about dignity "without loss of content" (Macklin 2003: 1420) – a claim which I do not find convincing. Macklin suggests that "'dignity' seems to have no meaning beyond […] respect for persons" (2003: 1419). It is not entirely clear whether this is a reference to universal dignity (we should respect persons qua those beings with universal dignity) or to variable dignity (we should respect persons, as this promotes their variable dignity). At times it appears as though Macklin is discussing universal dignity: she claims that dignity means "nothing more than a capacity for rational thought and action" (2003: 1420). However, I would suggest that our rational capacity may be the *reason* why we have universal dignity, but our rational capacity is not

equivalent to universal dignity: chimpanzees and AI can have rational thoughts and actions, but this does not entail that they possess universal dignity. Contrariwise, one might suggest that anencephalic human foetuses have universal dignity, but no rational capacity (humans as a class have rational capacity, but anencephalic foetuses as a sub-group of that class do not). Macklin is correct inasmuch as respectful treatment and dignified treatment are consistent; however, this does not mean that discussion of dignity adds nothing to medical ethics discourse, as Macklin claims.

Furthermore, universal dignity differs from respect for persons because of the former's inviolable nature. If Mathilde is tortured, raped, and ridiculed on a social media livestream, then it is clear that 'respect for persons' has not occurred – yet she nonetheless continues to possess universal dignity. This means that universal dignity simply cannot be fully captured by 'respect for persons', and therefore it remains a useful concept.

Perhaps, then, Macklin wishes to suggest that discourse on variable dignity can be eliminated without loss of content? This does not seem plausible either, since respect is a dyadic relation (where X respects Y), whereas variable dignity is a character trait which a person has: this immediately makes the two seem distinct. I have argued above that the way we are treated affects our variable dignity, and it would seem that respecting persons is something which would (probably) promote one's variable dignity, but would not constitute it. If Katherine respects Victor's personhood and treats him respectfully, this could promote his variable dignity, but the respectful treatment is not *itself* equivalent to Victor's (nor Katherine's) variable dignity. Thus, I would suggest that dignity – in both its forms – is a richer and more complex phenomenon than Macklin

suggests, and is far from useless, even though treating patients with respect is one way in which to promote their variable dignity.[103]

Macklin is correct about the importance of respectful treatment, however. Providing respectful treatment to patients is essential for promoting patients' variable dignity. Nurses and carebots should provide care – like that mentioned above and fleshed out further in the next chapters – which helps to promote patients' variable dignity in terms of their self-image, and enables them to behave in dignified ways (though they may choose not to). The *reason* why nurses and carebots should promote patients' variable dignity is because they have universal dignity, and I showed above how the latter grounds the former. Thus, whether the patient is aristocracy or poverty-stricken, a philanthropist or a bigot, they ought to be afforded respectful treatment which promotes their variable dignity as much as possible. This is echoed in codes of conduct which stipulate the ways patients should be treated – never is there any insinuation that status or personality should affect the treatment a patient receives.

It seems uncontroversial to claim that dignified care provided by nurses and carebots in residential homes should not include physical abuse, sexual abuse, emotional abuse, or neglect; I do not offer arguments in favour of such a claim, but take it as given. Treating someone with dignity needs to go beyond merely not abusing patients though. From the empirical studies and brief analysis

[103] Since my work defining dignity is only laying the groundwork for my spelling out how carebots (and nurses) should treat patients, if someone maintains that treating with dignity is the same as treating with respect, this does not weaken my argument in the next two chapters.

given above, we can say that treating someone with dignity involves treating them with respect, providing choice,[104] taking their wishes into account, involving them in their care, valuing their individuality, and providing privacy. For example, treating someone with respect and taking their wishes into account could involve addressing them in the way they prefer to be addressed, such as “Mrs Carter” or “Thelma” rather than “dear” and “sweetie” (Leland 2008). Valuing their individuality could involve allowing patients to choose their own soft furnishings in their room, or helping them to do their hair the way they prefer to wear it.

Treating someone with dignity is not a tick-box exercise: nurses and carebots cannot ‘provide’ dignity via a specific act in the same way they can provide a meal or a shower, whereby the task can be complete for the day. Rather, treating someone with dignity is a form of behaviour which (ideally) permeates all the tasks which are undertaken: treating someone with dignity is thus not a task in itself, but a *way* of completing tasks such as washing, feeding, and so on (Magee, Parsons, and Askham 2008: 5). Spelling out all the ways in which dignified care could take place is an enormous undertaking, far beyond the scope of this thesis. However, in the next two chapters, I explain and defend an important way in which nurses and carebots can promote patients’ dignity – namely, by obtaining consent before providing care.

[104] The extent to which patients can have their choices accommodated may be limited. For example, if a patient wants to take recreational drugs and attend rock concerts, such preferences are unlikely to be met; however, other choices such as choosing their outfits each day and choosing with whom they sit at meal times seem more practicable.

6.1 Carebots and dignity

The conceptual and normative exploration of dignity in this chapter was primarily to clarify the concept of dignity in order to extrapolate the sorts of behaviours which can promote patients' dignity – not because carebots need to fully understand universal or variable dignity (they do not). As noted in the previous chapter, codes of conduct and legal statutes (rightly) do not stipulate how healthcare professionals should *feel* (to govern people's feelings would be Orwellian and probably impossible anyway) but they can and do stipulate how healthcare professionals should treat others. Thus, promoting dignity is to be found in the way nurses behave, rather than in what they actually feel: this means that carebots are capable of promoting patients' dignity simply by behaving in appropriate sorts of ways. A door lock which can sense when a patient is changing or using the toilet could help promote patients' variable dignity, but clearly such a device would not need to understand either universal or variable dignity. The same is true for carebots: they would need to be programmed to behave in ways which promote patients' variable dignity, but this does not require an understanding of dignity, a theory of mind, sentience, feelings, or even high intelligence (though if carebots do have any of these things, that may be a bonus).

Some writers suggest that being cared for by a robot may *in itself* be undignified (Sharkey and Sharkey 2012, Sharkey 2014). Sharkey discusses a robot which carries patients about "like a baby" (2014: §2), and it is easy to see why this may be a threat to patients' variable dignity; however, Sharkey also notes that robots can help to keep a patient "dressed, and groomed in an appropriate way" (2014: §2). There is no definitive relationship between the

utilisation of carebots and (un)dignified care: it is possible to receive undignified or dignified care from a carebot, just as it is possible to receive undignified or dignified care from a human nurse.

It is possible that in some settings, carebots could promote patients' dignity more than a human nurse could. This may vary not only depending upon the patient, but also depending upon the task. For example, it is not uncommon for male patients to get an erection during baths or intimate medical examinations – a situation which can be embarrassing for both nurses and patients (Aging Care 2020, Norwick, Weston, and Grant-Kels 2019, Steady Health Men's Zone 2022). Patients who feel uncomfortable or embarrassed receiving toileting or bathing assistance from a nurse may not feel nearly as awkward receiving that same care from a carebot. Yet those same patients may prefer to be dressed or fed by a human nurse. Some patients may feel more comfortable having private or intimate conversations with carebots rather than human nurses: studies have shown that people often feel more willing to open up about their feelings with AI than with another human (Lucas et al. 2014). Conversing with a carebot about particular issues could thus feel more conducive to variable dignity-promotion than having the same conversation with a human nurse. Having some carebots and some human nurses in a residential home could therefore promote patients' dignity more than only having human nurses might – because it affords choice. If patients are being respectfully cared for, having the privacy they want, and enjoying the conversations they have, I see no reason why carebots cannot be compatible with dignified care.

One concern articulated by several writers (Sparrow and Sparrow 2006, Sharkey and Sharkey 2012, Sharkey 2014, Hotzak 2015, Tuisku et al. 2019), is that patients may have carebots foisted upon them, when they would prefer human nurses. Choice is an important aspect of elder care, but that does not mean that patients should have *unlimited* choices which may compromise safety, or the efficiency of staff. For example, if Miguel likes to eat his meals at 9am, 2.30pm, and 9pm, he may not be able to have those wishes accommodated if the residential home currently serves its meals at 8am, 1pm, and 6pm. In an ideal world, patients would be able to choose all the aspects of their care, including who cares for them, but the expected staffing shortfalls in the eldercare sector means that they may sometimes have to be cared for by a carebot even if they would prefer a human nurse. This is unfortunate, but patients cannot always have *everything* they want – for example, it might be deemed safer for patients to be lifted by a hoist or by Riba than by human nurses.

Being coerced or forced into accepting new technologies does not seem compatible with dignified care (Tadd 2005: 10), even if it is the case that the new technologies offer substantial benefits. There are limits to this though. For example, if James prefers his nurses to wash his bed linens by hand rather than in a washing machine, it would probably be reasonable to disregard James's wishes, because of the time which washing machines save nurses, thus giving them more time to engage in other care tasks. Carebots too may be able to undertake some time-consuming tasks which nurses currently perform, thus providing the opportunity for nurses to spend more 'quality time' with patients – which could help to promote patients' dignity more than having

nurses who must rush to get from one patient to the next (Magee, Parsons, and Askham 2008: 16).

It may be just a pipe dream to think that patients could get more ‘quality time’ with human nurses when carebots are introduced, however. Realistically, it might be the case that as carebots begin to perform some tasks which nurses have hitherto performed, nurses are simply given alternative or additional tasks to complete, or more patients to care for. However, it is important to acknowledge that if this were to occur, it would not be evidence of a problem *with carebots*, but with those who control nurses’ workloads. It seems clear to me that carebots are compatible with dignified care, and could provide choice for patients, and save time for nurses.

7 Conclusion

We have seen above that dignity is often appealed to as an essential feature of care, and thus it would seem to be an important concept for us to understand, not least so that carebots (and nurses) can ensure they provide dignified care. Yet different senses of ‘dignity’ are so often conflated and ill-defined that detailed concept analysis was required. I distilled dignity into two broad concepts: universal dignity and variable dignity. Universal dignity exists in all and only humans, and involves having a high moral status, whereas variable dignity can vary between people, and in the same person over time. I suggested that one’s treatment, behaviour and self-image can all affect one’s variable dignity.

Universal dignity helps ground the claim that people should be treated well, and have their variable dignity promoted. One of the most important ways in

which variable dignity can be promoted by nurses or carebots is by obtaining consent prior to giving care; this is something which I argue for more fully in the next two chapters, which examine consent in detail.

Chapter 6

Consent: The nature, grounds, and importance of consent

The previous chapter examined two types of dignity (universal and variable), and I discussed some ways in which variable dignity can be affected. One important way of promoting patients' dignity[105] is to seek consent prior to engaging in care activities. This chapter examines the nature, grounds, and importance of consent in general, which provides background and context for the next chapter, which demonstrates how dignity is promoted by consent-seeking.[106]

Consent continues to be thoroughly explored in the philosophical literature, particularly with regard to sexual activity[107] (Sandoz 2021, Wertheimer 1996, 2003, 2014, Beres 2014, Lamb, Gable, and de Ruyter 2021, Brison 2021,

[105] In this chapter, all discussion of dignity refers to variable dignity, unless otherwise stated.

[106] I use 'consent-seeking' to mean asking for consent *and then abiding by the person's response*.

[107] 'Sexual activity' refers to acts undertaken for sexual arousal or gratification of one or both parties. *Sexual consent* thus refers to consenting to sexual activity.

Anderson 2005, Dougherty 2013) and medical procedures[108] (Eyal 2014, Roache 2014, Aljouni 1995, Leclercq et al. 2010, Kongsholm and Kappel 2017, Gill 2004, Saunders 2012), as well as political governance; however, consent to routine care[109] activities such as help with dressing, toileting, feeding, washing, and moving around has been hitherto underexplored. My focus in this chapter is to establish what consent (of all types[110]) involves, and why it is normatively significant. In the following chapter, I develop a normative account of routine consent which is relevant to both carebots and human nurses alike.

Some acts legally require consent. For example, in the UK, the Sexual Offences Act decrees that sexual activity requires consent at the time it takes place (National Archives 2003); other similar laws exist in many countries to make non-consensual sexual activity[111] a crime. In medical and health care contexts in the UK, requirements about consent are laid out by the General

---

108 'Medical procedures' refers to acts which are undertaken to maintain or improve one's health – for example, an operation, a diagnostic or exploratory procedure, administering medication, or examining a patient's body. *Medical consent* thus refers to consenting to medical procedures.

109 In this chapter, 'care' refers to practical care only; not emotional care. See Chapter 4 for more on this distinction.

110 'Consent of all types' refers to medical consent, sexual consent, routine consent, and consent to other activities such as sports, research, touching someone else's property, the gathering of cookies on websites, etc. Some such consent-seeking is a legal requirement, and some is to control liability in case of accident. Political consent is somewhat different from these activities, and is not included in what I call 'consent of all types'.

111 Non-consensual sexual activity involves any sexual activity where consent was not obtained, including but not limited to rape and sexual assault. Some writers (Brison 2021) suggest that rape is more than mere non-consensual sex, and therefore consent is not the most helpful way to frame permissible sexual activity.

Medical Council (2013) and the Nursing and Midwifery Council (2018). My task herein is not merely to describe the legal status quo, however. Although existing laws and policies give us a clue about what people deem to be important with regards to consent, they do not provide a flawless, universal, and philosophically robust account of the nature of consent – indeed, it may not be possible to provide such an account. Although laws, codes of conduct, and institution-specific policies are important, my primary foci in this chapter and the next are conceptual and normative explorations of consent (rather than legal explorations), because these are the most important aspects of consent for the purpose of sketching out how carebots should behave vis-à-vis consent-seeking. Policies provide a minimum threshold of legally acceptable behaviour, but *ideal* nursing behaviour involves more than simply meeting this low threshold.

I begin this chapter by outlining some routine care activities and arranging them into a hierarchical pyramid. This helps set the scene regarding the sorts of routine care activities on which I focus. In §2 I engage in conceptual analysis, examining some features of consent and what makes consent legitimate, such as its being sufficiently informed, voluntary, and given by a person with capacity. In §3, I focus on the normative significance of consent: I analyse why consent of all types is deemed to be important, including people's right to bodily integrity and autonomy, the development of trust, and perhaps most crucially, the way consent-seeking promotes patients' dignity. This chapter lays the foundation onto which my account of routine consent – established in the next chapter – is built.

1 Routine care activities

I presently explain the types of activity for which routine consent[112] applies (viz. routine care activities) and order them into a hierarchy based on the level of intimacy involved. First, let me explain the sorts of activities which are *not* routine care activities. These include the medical care activities which nurses carry out for patients in residential homes, such as taking a patient's temperature, checking their blood pressure, or applying cream to treat a skin complaint. These activities are of a medical nature – even if regularly carried out by care staff with little or no medical training. Consent to these activities thus comes under the umbrella of medical consent, and is not what is under discussion when I refer to routine consent. There are existing accounts of (medical) consent (Eyal 2014, Roache 2014, Fan and Tao 2004, Cave 2021, Katz 2003) which may be useful for carebots performing medical tasks (though further study – beyond the scope of this thesis – may be required to fully spell out how, why, and when carebots should seek medical consent).

Residential home nurses provide a variety of non-medical care for patients, such as practical assistance and personal care (Social Care Institute for Excellence 2013b); these non-medical caring activities are what I discuss under the umbrella term 'routine care'. Some simple self-care tasks are essential to daily life; people must be able to perform these tasks safely and competently in order to live independently. They are generally learnt as children, and adults almost always do these for themselves. They include:

[112] Just as 'sexual consent' means consent to sexual activity, and 'medical consent' means consent to medical activity, 'routine consent' means consent to routine care activities.

- **Walking** or using a wheelchair to move about
- **Feeding**: cutting up food; putting food into one's mouth
- **Dressing / grooming**: selecting clothes; putting them on and taking them off; managing one's appearance; shaving; brushing hair
- **Toileting**: getting onto and off the toilet; using it appropriately; cleaning oneself afterwards
- **Bathing**: cleaning oneself in a bath or shower; washing one's hands and face at a sink
- **Transferring**: moving from one body position to another e.g. from a chair to standing (entire list adapted from Katz et al. 1963, 1970, Katz and Akpom 1976)

Some of these activities are easier than others – for example, it has been found that most elderly people find eating independently easier than taking a bath independently (Gerrard 2013). Nonetheless, all are essential activities in everyday life; they are often called activities of daily living (ADLs) or basic activities of daily living (BADLs). The Katz ADL Index (which assesses whether a person is able to perform the activities independently) is widely used in health and social care settings to ascertain dependency levels, and the type and extent of assistance a person requires.[113] It is suggested that if patients can perform all six activities by themselves, they are independent; if they can perform only 3-5 by themselves then they are moderately dependent

[113] A limitation of the Katz index is that it does not adequately measure degrees of difficulty in performing the tasks.

and require assistance; if they are only able to perform 2 or fewer by themselves, they are severely dependent and require far greater assistance (Katz et al. 1963, 1970, Katz and Akpom 1976, Shelkey and Wallace 2012). Virginia Henderson famously defined nursing as helping a patient to perform the activities which he would perform himself if he were able (Pokomy 2017: 14); what I call routine care involves helping patients with these tasks.

Some more complex activities are also required when living independently. These are often referred to as instrumental activities of daily living (IADLs) (Kernisan 2018), and include:

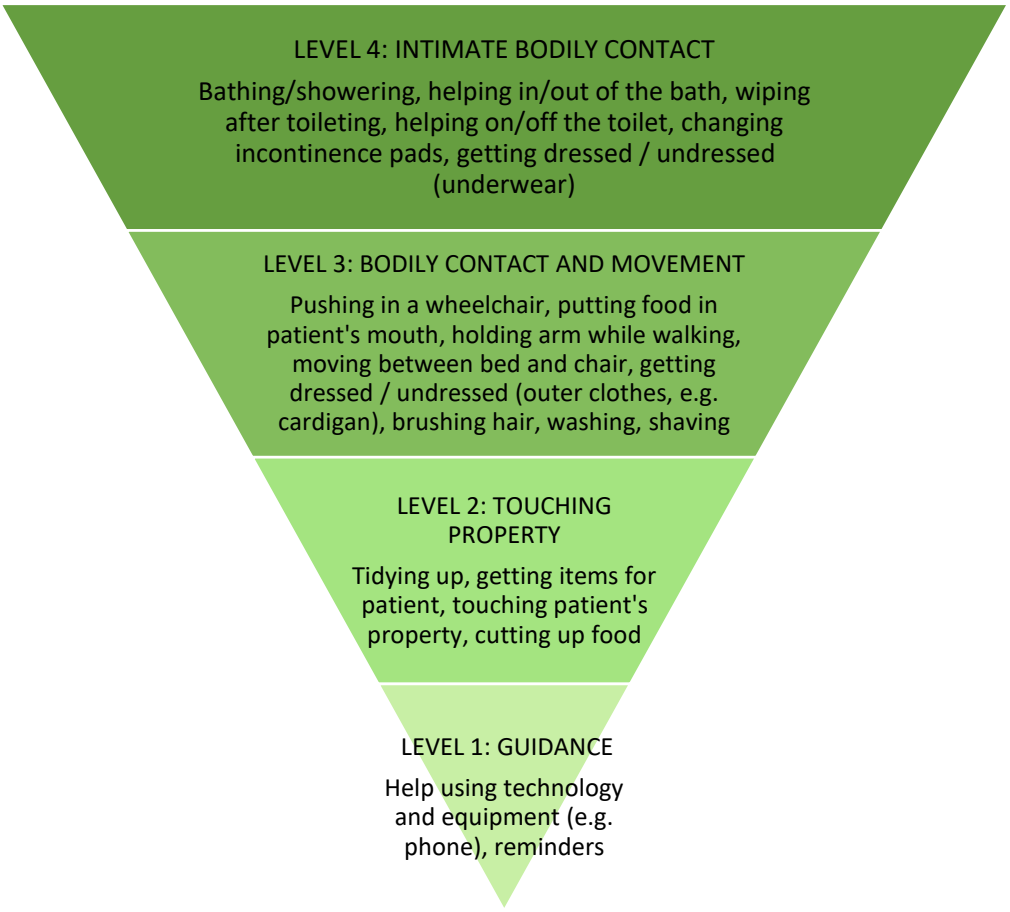
- Housekeeping, laundry, washing dishes, and other household chores
- Meal preparation
- Shopping for groceries and other necessities, including using transportation as necessary
- Managing medications if required
- Financial management and budgeting
- Using technology such as a phone, tablet, or computer for communicating with others

The first four of these activities are generally undertaken automatically by residential homes, and the assumption is made herein that patients have these activities performed for them, and thus I do not mention them as part of what I call routine care activities. Patients living in residential homes either receive their care for free, or pay a standard rate each week or month by standing order, thus very little (if any) short-term budgeting or financial management is required of patients. Using technology such as phones and computers is the only one of these IADLs which I include in my account of

routine care, since patients are likely to want to use a phone or computer to keep in touch with friends and family – and possibly to purchase items online, to find information, or for recreation.

Not all routine care activities are equally intimate. Here I order them into a hierarchy – in the form of an inverted pyramid – based on the level of intimacy involved in the activities. Level 1 activities are routine care tasks with the lowest levels of intrusiveness and intimacy; intimacy levels increase up the pyramid, through to Level 4 activities which are highly intimate. Later I argue that all care activities on the pyramid require consent before they are undertaken by nurses. In §3 of Chapter 7, I show that if any of these care activities are performed non-consensually, there is likely to be a reduction in the patient's dignity (with higher-level non-consensual care causing the greatest reduction), and I argue that these dignity reductions can be cumulative.

Routine Care Activities Pyramid:



Presently, I briefly discuss the different levels of the pyramid, and the activities contained within each level.

Level 4 routine care activities such as toileting, bathing, and changing underwear involve touching or seeing someone's most intimate body parts. Because these activities are commonly understood to be intimate, my suggestion that consent should be obtained beforehand seems uncontroversial. Generally, Level 4 activities are ones which we would not

permit a stranger – or even a close family member – to perform if we are capable of doing it ourselves.[114]

Level 3 routine care activities include dressing, personal grooming, feeding, support while walking, or pushing in a wheelchair; these involve moving or touching the patient, but in a less intimate way than Level 4 activities.

Level 2 routine care activities involve touching someone's property, where 'property' is broadly construed to include things which are temporary property, such as food, cutlery, a walking frame, chair, or bed. Patients in residential homes often have some of their own property; other items (such as a wardrobe) are considered theirs for the duration of their stay in the facility, and other items may be treated as property while the patient is using them (such as cutlery or a chair).

Level 1 activities are the least intimate of all the routine care activities, and include helping patients to use technology or equipment, and reminding them to do things. These activities need not involve physically touching the patient, and may not even involve touching the patient's property or items they are using.

I do not suggest that people progress through the levels in any particular order, or that any Levels are mutually exclusive with any others. For example, Ethel may need help with bathing (Level 4) and with using a mobile phone (Level 1) but be able to do everything else largely by herself. It might be thought that

[114] Some of the activities on the pyramid such as feeding, showering, or undressing may be undertaken for sexual purposes, even though the individual can perform the activity themselves. In such cases, (sexual) consent would need to be sought.

some of these activities – perhaps especially Level 1 activities – do not require prior consent, but I later argue that if they are performed without patient consent, this can reduce patients' dignity.

Some writers may disagree with my levels – for example, they may think that that it is unproblematic for a nurse to tidy a patient's property without consent (Level 2), whereas cutting up the patient's food without consent (also Level 2) is a huge affront to dignity. People vary, both across cultures and individually, such that brushing a patient's hair or shaving their face may be considered a very intimate activity by some individuals or cultures, but not by others. Patients may also change their reactions to different care activities from time-to-time. It may also be the case that there are some routine care activities which are not mentioned in the pyramid at all.

Although I believe my pyramid captures common conceptions about the invasiveness of routine care activities, I concede there will undoubtedly be patients (and non-patients) who disagree with my ordering of activities, and there is no way for nurses or carebots to know whether a particular patient feels their dignity is lowered more by a non-consensual reminder than by non-consensual dressing. However, this uncertainty is not highly problematic. In fact, the uncertainty underpins my arguments in this chapter and the next – namely, that nurses and carebots should seek consent before undertaking *any* care activity. Even the most thorough understanding of dignity will not reveal whether a particular unknown patient would prefer to be fed by a nurse, or to go it alone, nor how much their dignity could be reduced if they are fed non-consensually. But if nurses and carebots always obtain consent before carrying out *any* routine care activity, then whether or not others agree with

my ordering of activities on the pyramid, patients' dignity is always promoted. Thus, any difference of opinion regarding the placing of activities on the pyramid simply underscores the importance of consent-seeking for which I will argue.

So far, I have outlined what I mean by routine care, and separated routine care activities into a hierarchy. Let us now turn our attention to consent: first I give a conceptual analysis of what consent involves; this is followed by a normative account of why consent is important.

2 Features of consent

In this section I discuss some of the pertinent features of consent of all types; this helps ascertain what (legitimate) consent is, so that I can assess its normative significance. Consent continues to grow in its moral and legal significance (Dougherty 2013, 2015, Brison 2021, Sandoz 2021, Roache 2014, General Medical Council 2020, NHS 2022, Royal College of Nursing 2017); it is (rightly) seen as an important – possibly the *most* important – way of behaving appropriately in various situations. Consent is widely discussed in the philosophical literature, particularly medical consent and sexual consent. But what *is* consent (or withholding consent)?

Withholding consent is the power to refuse something which affects oneself – a "power of veto" (Wilkinson, Herring, and Savulescu 2020: 89). One cannot demand a sex act or medical procedure simply by consenting to it, but one can refuse it by withholding consent. The activity being consented to must necessarily affect oneself; I cannot consent to Adrian having sex with Katie when it does not affect me – only Katie and Adrian can consent to Katie and

Adrian having sex.[115] The activity needs to do more than simply *affect* the person giving consent, however. If you take all Carla's money and give it to me, this affects me (I will become richer!) but I cannot consent to you taking Carla's money: Carla's consent is required for that. If Carla does not consent to you taking her money, then your taking her money is (pro tanto) impermissible. The person giving consent thus needs to be the person who, by consenting, morally transforms the action from impermissible to permissible (Alexander 1996, Hurd 1996) – the person whose body, property, or affairs are being touched or negatively affected (Dougherty 2013: 723). Consenting thus involves sanctioning a particular activity by others which affects oneself – an activity which is by default *not* sanctioned.

Let us now consider what distinguishes consensual acts from non-consensual acts. There are multiple ways an activity can be non-consensual, but two important ways are:

a) the person affected has expressed that they do not want the act to take place
b) the person affected has not expressed any preference either way

It is reasonably clear that an act in situation (a) is non-consensual, but (b) is more controversial. In many situations, we can often make a best guess. My friend Sharon has never expressed a preference either way regarding whether

[115] There may be other people whose consent is required *in addition to* Adrian and Katie's consent – for example, Katie's husband(!) – but the consent of Katie and Adrian is crucial and cannot be permissibly dispensed with. Sometimes, people can provide medical consent for procedures on someone else – for example, a parent can provide consent for their child to be immunised, whereas sexual consent must, in almost every case, be given by the individual(s) involved.

I can use her car to drive her son to the hospital in an emergency (we have never discussed it). My borrowing her car in an emergency is in essence non-consensual at the time, but it would seem reasonable to assume that she *would* consent if she were asked, because she cares about her son and trusts my driving ability. Alternatively, we might think that consent in such an emergency is not necessary to legitimise the car-borrowing, just as consent to medical intervention in emergencies is also considered unnecessary (NHS 2018a).

Sharon has also never expressed any preference either way regarding whether I can throw acid in her face (we have never discussed it), but given that acid-throwing causes pain and disfigurement, it would seem reasonable to assume that she would not consent to it. Except in cases of preventing or minimising harm, the default position on consent is that someone does not consent to others laying hands on them or touching their property – unless they opt in to the activity by giving their consent (Dougherty 2013: 723). Borrowing Sharon's car prevents (or mitigates) harm about which I know she is concerned, so I reasonably believe she *would* consent to car-borrowing if asked – so this seems to justify my borrowing her car to save her son. As for acid-throwing, it does not prevent harm (rather the reverse), and I do not reasonably believe Sharon would consent to it if asked – therefore, throwing acid in her face would be non-consensual. (Even if she *did* consent, this would not mean I *should* throw it).

These two possibilities (a and b) are not the only ways in which an activity can be non-consensual, however. Sometimes, an activity can be non-consensual even if the person involved has clearly said "Yes, I agree to this activity". Their

(apparent) consent may not count as *legitimate* consent[116] because they do not fully understand what they are consenting to, or because their 'consent' was forced or coerced. Legitimate consent-giving[117] of all types must be:

- Sufficiently informed
- Given by a person with the capacity to consent
- Voluntary

I discuss these in more detail in §2.1 and §2.2, but briefly, it means that the person giving consent must have received an appropriate level of information about what he is consenting to (including any major risks), he must have the mental ability to agree to the activity, and his agreement must be given of his own volition. An act affecting an individual is non-consensual if apparent consent is given, but one or more of these conditions is not met. For example, if a young child were to ostensibly agree to a complex medical procedure, it would not be legitimately consensual because the child lacks the capacity to make such decisions, (and is therefore possibly not satisfactorily informed as to what the act involves or its repercussions).

116 Legitimate consent is that which has the normative power to be morally transformative – to change an activity from prima facie impermissible to prima facie permissible.

117 The withholding of consent is more complicated than the giving of consent: a capacitated adult can legitimately withhold consent without being informed – or else there is an obligation to become informed about any action another person proposes. E.g. if my friend suggests cave diving, I can legitimately refuse even though I know almost nothing about cave diving. With medical treatment in the UK, young people aged 16-17 can consent to treatment, but if their withholding consent to treatment could result in severe injury or death, treatment can be performed without their consent (NHS 2017)

The features of legitimate consent are not static. A quick look at history shows us that voluntariness and being informed were not always seen as necessary features of legitimate medical consent. For example, in ancient Greece, it was permissible – or even morally good – for doctors to "convince" patients to undergo medical treatments, and thorough information-giving to patients was not essential either (Plato 1925: 296, 2000: §IV 720); some other writers from antiquity suggest that physicians ought to actively *conceal* most information from patients (Hippocrates 1923: 297). Consent-giving via written documentation has been recorded since the fourteenth century for the purpose of exculpating surgeons in the event that their patient dies (Aljouni 1995, Leclercq et al. 2010), but operating on a patient without their consent was not necessarily viewed as wrong – particularly if surgery was successful. In fact, throughout much of the history of medicine, a focus on patient consent was conspicuous by its absence in a student doctor's medical training (O'Shea 2018: 261, Katz 2002: 3). The vast *Cambridge World History of Medical Ethics* (Baker and McCullough 2009) has sixty-three chapters – of which *none* are dedicated to consent. Thus, we see that although being informed and making a voluntary choice are now considered to be crucial features of legitimate medical consent, this has not always been the case. Clearly, medical consent – both its necessary features and its normative significance – has evolved naturally over time (and is still evolving, as the extensive literature demonstrates).[118]

[118] Routine consent is also likely to evolve; my conclusions herein about how and why carebots should seek consent are not likely to be the last word on the subject.

Sexual consent has also evolved – and is still evolving; for example, even over the past few decades we have seen a change in the permissibility of having sex with someone who is silent. A few decades ago, silence was viewed as ambiguous or even consenting behaviour (only a verbal "no" clearly signalled non-consent, so sex with a silent person was often permissible) (No Means No Worldwide 2018); presently, silence is viewed as non-consent (only a verbal "yes" clearly signals consent, so sex with a silent person is generally impermissible) (Dougherty 2015, Sandoz 2021, Nieves et al. 2022). Laws on rape and sexual offences have similarly evolved: violence by the perpetrator and vehement refusal by the victim are no longer necessary constituents of rape, as they once were (National Archives 2003).[119] Evidently, the features – and the normative weight – of legitimate sexual consent are not carved in stone. Indeed, sexual consent as a phenomenon is a far more recent notion than medical consent is. In the UK and elsewhere, sexual consent has historically been linked to (heterosexual) marriage: a husband had the legal right to sex with his wife – obtaining her consent was neither a legal nor moral necessity (Bates 2015). There remain over thirty countries – accounting for around half of the world's population[120] – where marital rape is still not a crime (News18 2021). In these countries, just as in the UK historically, marital rape

[119] Interestingly, UK law has never recognised that women can commit rape: the law stipulates that rape involves forced penetration of someone *with the offender's penis* (National Archives 2003).

[120] The full list of countries is: Afghanistan, Algeria, Bahrain, Bangladesh, Botswana, Brunei, Central African Republic, China, Cote d'Ivoire, Democratic Republic of Congo, Egypt, Ethiopia, Haiti, India, Iran, Kuwait, Laos, Lebanon, Libya, Malaysia, Mali, Mongolia, Myanmar, Nigeria, Oman, Pakistan, Senegal, Singapore, South Sudan, Tajikistan, Uganda, and Yemen (News18 2021).

is impossible because by getting married, the wife[121] is considered to have legally consented to any and all sex with her husband for the duration of the marriage (Bates 2015).

By contrasting the UK's approach to sexual consent, and the approach in China, India (etc), we can see two different ways in which consent can be given (vis-à-vis frequency). The UK approach is that each new sex act needs to be consented to; even if Lucy has consented to sex with Paul 99 times previously, consent still needs to be signalled in the hundredth encounter in order for the sex to be permissible: Paul cannot assume Lucy consents. Let us call this *'repetitive consent'* because consent needs to be repeatedly obtained. The alternative approach is that in China, India (etc), where one act of signalling consent (such as by getting married) applies to all acts thereafter; consent does not need to be sought repeatedly, but instead, the consent begins at a particular point and then lasts for an agreed duration thereafter (such as the entire marriage). Let us call this *'lasting consent'*. Although we might maintain that sexual intercourse requires repetitive consent, we might think that lesser sexual acts – such as the placing of a hand on the thigh, or a kiss on the lips – are more apt to be governed by lasting consent. In most long-term relationships, it is not morally necessary to ask permission before each instance of touching or kissing one's partner; instead, it is apt to presume

[121] I do not write "and husband" here, because for much of history, wives were considered to be the *property* of their husbands. In the UK, it has been a long time since wives were deemed to be the *property* of the husband, but marital rape was only criminalised in 1991 (Law Commission 1991).

that the consent to a hand on the thigh or a kiss on the lips is present in virtue of having entered into the relationship.

In the next chapter, I consider whether routine consent should be repetitive or lasting; I reject the idea that patients give lasting consent to all care activities when entering the residential home; instead, I suggest that routine consent should be repetitive if the care is to be permissible (consent should be obtained for each new token care activity).

2.1 Consent as a performative act or mental state

This sub-section examines whether consent is a mental state, or a performative act. This is important because it will determine whether carebots need to obtain consent (through written, verbal, or non-verbal means) for care to be consensual, or whether care is consensual whenever a patient has a consenting state of mind (even if they have not expressed this).

When writers discuss “giving consent” this necessarily seems to involve some sort of behaviour – whether verbal or non-verbal. It is the *giving* of consent – not just *mentally* sanctioning something – which is seen as the important legal standard in both sexual and medical settings.[122] Nevertheless, we may think that ‘consenting’ (as opposed to ‘giving consent’) is rather more vague, and can be a thought or feeling which is private to the thinker. Oftentimes, our state of mind vis-à-vis consent is reflected in our behaviour – we give behavioural cues that we consent to X *because* we (mentally) desire or sanction X. But

[122] An exception to this is life-threatening situations where action is taken based on what we expect the patient would consent to if he were able to communicate.

this is not *necessarily* the case; one can occur without the other, and it is useful to consider whether 'consenting' is the mental state, or the behaviour – or both.

Several writers (Wertheimer 1996, McGregor 2005, 1994, Schulhofer 1995) take the view that consent is a performative act – a behaviour of some sort. This can be something verbal (such as saying: "Yes, I agree to that") or something non-verbal such as a nod of the head or a smile. It could also take the form of any behaviour that suits the situation. For example, if Julie goes to visit a palm-reader, pays the fee, sits down, and holds out her hand, it seems clear that Julie is consenting to having her palm read. In sexual situations, consent could be demonstrated through physical enthusiasm (Kavanagh 2016: 43). In medical situations, consent could be signalled behaviourally too, such as if Nitesh rolls up his sleeve and holds out his arm while the nurse is preparing a vaccination. Given that all we have access to is other people's behaviour (rather than their state of mind), it might seem to make sense to maintain that all there is to consenting is *signalling* one's consent.

However, this seems too simplistic, for it is clear that someone can behave *as if* they consent when really they do not want the activity to happen. Suppose Gwyneth for some reason fears she will be killed if she does not appear consenting and enthusiastic about sex. So, whenever a man speaks to Gwyneth, she always behaves as if she desires and consents to sex with him. Suppose she is talking to Alfredo, who is a normal non-threatening man who behaves perfectly gentlemanly towards Gwyneth. Nonetheless, Gwyneth fears he might kill her unless she has sex with him, so she suggests they have sex, and Alfredo agrees. Gwyneth behaves enthusiastically throughout their sex, even though she does not in fact want to have sex with Alfredo. She only

signals consent because she fears for her life. If consent must be voluntary in order to be legitimate – a widely-held position to which I assent – then it cannot be true to say that Gwyneth consents to sex simply because she behaves as if she does. This would seem to suggest that behaviourally signalling consent is not equivalent to legitimately consenting (though it may be a necessary constituent), and presents a significant problem for writers (such as Wertheimer 1996, McGregor 2005, 1994, Schulhofer 1995) who suggest that consenting is a performative act. Given that there are probably numerous sex workers and trafficking victims who, due to fear of violence, behave as if they consent to sex when really they do not want sex, Gwyneth is not merely a fictitious thought experiment, but a real-world likelihood.

An alternative position is that consenting is a state of mind (Hurd 1996: 124–138, Alexander 1996) of sanctioning or approving – consenting is a thought or feeling, directly privy only to the person thinking or feeling it. This means that I can mentally consent to Polly hugging me without ever indicating this in my behaviour. This mental consenting can be what takes place when we are attracted to someone but keep it to ourselves; we mentally consent to sex with them (because we would sanction it, at an appropriate time and place), but this is not reflected in our behaviour.

However, construing consenting as a state of mind can also be problematic. Suppose Kristin wants to have sex with Rupert (she mentally sanctions it). However, when Rupert suggests that they have sex, Kristin behaves coyly and ‘plays hard-to-get’ by telling him “No” even though she does in fact (mentally) sanction sex with him at that moment. Kristin mentally consents to sex with Rupert, but she is not behaving as if she consents. Now suppose

Rupert forcibly has sex with Kristin anyway, while she strongly protests and pushes him away – but all the while, she (mentally) sanctions the sex. If consenting is purely a mental activity, then the sex between Rupert and Kristin was consensual – but such a position seems counterintuitive, given her verbal and physical protests. We generally think that consent must be signalled in some way in order for it to be legitimate.

I suggest that there are two distinct elements to legitimate consent: a performative act of consent-signalling such as saying “Yes” or nodding, and a mental state of sanctioning. Legitimate consent is a performative act *in addition to* the mental state, and certain other psychological conditions (voluntariness, capacity, being informed). This hybrid account of consenting helps to solve the weaknesses of both the consenting-as-behaviour and the consenting-as-a-mental-state approaches. The problem with the consenting-as-behaviour approach is that Gwyneth does not in fact (mentally) sanction the sex she engages in. Her lack of mental sanctioning is what seems problematic about the example: mental sanctioning is necessary in order for the sex to be legitimately consensual. The problem with the consenting-as-a-mental-state example is that Kristin is not behaving as if she sanctions the sex (the performative act is absent), and so to any observer – including Rupert – it *appears* that she does not consent. It seems reasonable to maintain that Kristin needs to signal her consent through a performative act in order for the sex to be legitimately consensual. Thus, my hybrid account which requires people to behaviourally signal their consent *in addition to* mentally consenting improves on both the consenting-as-behaviour and the consenting-as-a-mental-state approaches.

Under my hybrid account, neither example (Gwyneth; Kristin) involves legitimately consensual sex. Both the performative consent-giving behaviour *and* the mental state of sanctioning are essential conditions of legitimate consent (though the mental state may be difficult or impossible to discern) and one aspect is absent in each of my examples (Gwyneth; Kristin). In spite of these examples, a person's behaviour is probably a good guide to whether they mentally consent (Alexander 1996: 174). Hereafter, when I refer to 'consenting' (or 'withholding consent'), I mean to refer to the performative act of consent-signalling (or non-consent-signalling) *in addition to* the corresponding mental state in the person at that time.

2.2 Capacity and informed consent

I noted above that for consent to be legitimate, it must be voluntary, sufficiently informed, and given by a person with the capacity to consent. Here I want to say a little about what is meant by capacity and informed consent in particular. This is important to my overall project because carebots may need to provide information to patients about routine care activities. (Assessing capacity may at some point be undertaken by AI – though this would need thorough philosophical exploration beyond the scope of this thesis).

As I established in my introductory chapter, the patients under discussion in this thesis may be frail, sick, or physically infirm, but they are taken to be neurotypical adults who have the capacity to make decisions about their lives and their care. In this context, 'capacity' means the ability to understand and use information to make a decision, and to communicate that decision (NHS 2022, 2018b, Royal College of Nursing 2017: 5–8, National Archives 2005, Kapp and Mossman 2014). Determining capacity is not always an easy task,

but in the UK, all adults[123] are assumed to have capacity for medical consent "unless there's significant evidence to suggest otherwise" (NHS 2022); this is echoed in the *United Nations Principles for Older Persons* (1991: Principle 14).

The levels of capacity which are required for medical consent, sexual consent, and general decision-making are not the same. Someone may be unable to weigh the potential risks and benefits of a life-or-death medical procedure, but be able to make a decision about who they would like to sit with at lunch time. There are also differences in the levels of capacity required for legitimate consent to different medical procedures.

The possibility of being fully informed is often linked to capacity levels: someone who lacks capacity may also often lack the ability to become fully informed about a procedure. The level of mental aptitude required to become fully informed correlates with the magnitude of information one must comprehend. For example, a 5-year-old child is capable of consenting to having a rash on their arm examined, because they can be fully informed about what the procedure involves (a doctor looking at the arm) and the risks involved (none). However, if a riskier and more involved procedure is being proposed, such as organ transplantation, the 5-year-old child cannot be fully informed because the procedures are too complicated, and an understanding of life and death is required. Thus a 5-year-old child cannot legitimately consent to organ transplantation, because they cannot be fully informed.

[123] In the UK, adults are people aged 18+. Patients aged 16-18 can give consent *to* treatment, but may not always be permitted to withhold consent to beneficial treatment (Wilkinson, Herring, and Savulescu 2020: 95, NHS 2017)

Giving patients sufficient information about what medical procedures involve is one way of promoting their dignity (Enes 2003, Bayer, Tadd, and Krajcik 2005) and respecting their right to choose what happens to them. But one might wonder how much information is required in order for consent to count as *informed* consent. There is not a simple answer to this question. In a medical context, becoming informed may involve lengthy discussions, and written documents hundreds or thousands of words long – or for more simplistic procedures, such as examining a broken finger, perhaps no explanation at all is necessary, since the patient knows what looking at a finger involves. When determining how much information to provide to a patient, healthcare professionals may simply 'follow the crowd' and provide the same amount of information that other healthcare professionals provide – this is known as the *community practice standard* (Brody 2003: 101). This approach can be problematic since it is not sensitive to what patients actually need or want; if other healthcare professionals are providing insufficient information, then it is not useful to simply follow their example. This is one reason why we should not necessarily program carebots to copy their human counterparts, and one of the reasons why this thesis is necessary. The *reasonable patient standard* is seen as a preferable approach to information-giving (Brody 2003: 101): nurses should provide enough information for a reasonable person to make an informed decision (Ibid). Although questions may emerge regarding what a 'reasonable patient' requires, and nurses or doctors may over-provide information so as to avoid lawsuits (Ibid), this approach to information-giving nonetheless seems the more prudent one.

Adequate information-giving also needs to occur in sexual matters, though to a lesser extent than we would demand for medical procedures. In order for sex to be permissible, someone who is unfamiliar with a sexual activity may need to have the act briefly explained before agreeing to it. As with medical consent, if a person (ostensibly) consents to something without being sufficiently informed, then this is not legitimate consent, and the act is impermissible. It is generally permissible not to provide lengthy and detailed explanations of what sexual activities involve before engaging in them, but an adequate level of understanding is nonetheless necessary for legitimate sexual consent.

3 Normative significance of consent

Having considered some features of legitimate consent (it must be voluntary, informed, and given by someone with capacity), I now discuss several reasons why consent (of all types) is normatively significant. Adhering to consent[124] has both intrinsic value (because we have autonomy and bodily integrity – explored further in §3.1 and §3.2), and extrinsic value (because it (prima facie) promotes positive outcomes for the individual concerned).[125] Perhaps most importantly, consent is morally significant because it promotes variable dignity.

Consent is a necessary (but insufficient) condition for sex acts to be morally permissible (Wertheimer 1996, 2003) – the same can be said for any act which is by default not sanctioned, such as entering someone's home. Consent is morally transformative: it is the 'moral magic' (Hurd 1996, Alexander 1996)

[124] By this I mean refraining from acts which have not been consented to.

[125] Recall that consent is a power of veto only: if Ian consents to something dangerous such as me pushing him off a cliff, this does not mean I *should* push him off a cliff.

which can transform an impermissible act into a permissible act or supererogatory act. For example, suppose we know that last Tuesday, Henry took his friend Dave's car and drove it to work; at work he amputated Mrs Oakley's hand; and in the evening he had sex with Ava. With this information alone, we cannot determine whether Henry has done anything morally wrong, because we do not know whether these acts were consensual. If I now further stipulate that Dave had said it was OK to borrow the car, Mrs Oakley had agreed the hand amputation was for the best, and Ava consented to the sexual activity, we can now determine that Henry's acts were morally permissible (and perhaps the hand amputation was morally good).[126] Contrariwise, if I stipulated that the car-borrowing, hand-amputating, and sex were not consensual, this transforms Henry's actions into impermissible ones. The transformative power of consent helps explain why it is important for consent to meet particular conditions which make it legitimate: if we were vague about what counts as legitimate consent, there would be confusion about which activities are morally transformed. But *why* does consent have this transformative power? There are a few possible reasons, which I discuss below. Briefly, these reasons are:

- **Autonomy**: people are generally free to choose what they want, and consent-seeking helps give people what they want
- **Bodily integrity**: people have the right to decide how their body is treated, and consent-seeking helps maintain that right

[126] This is assuming that other important conditions also pertain, such as Henry being sober, having a driving licence, being a surgeon, not coercing the consenting parties (etc).

- **Dignity**: consent-seeking helps to promote people's dignity because it respects their personhood
- **Trust**: consent-seeking helps to build trust between the different parties involved

These reasons for the normative significance of consent are interlinked: we have the autonomy to choose what happens to our bodies; respecting our autonomy and bodily integrity promotes our dignity; we build trust with someone who promotes our dignity and respects us. Non-consensual acts are a breach of trust; they violate autonomy and bodily integrity, and can reduce a person's dignity. I discuss these phenomena separately and explain the links between them forthwith.

3.1 Autonomy

Enshrined in the Hippocratic Oath sworn by doctors is the maxim "Do no harm"; doctors are also expected to assent to the tenets of beneficence and non-maleficence (Encyclopedia Britannica 2017). In other words, it is the job of doctors (and the nurses who support them) to help patients rather than harm them. Why, then, would a doctor allow someone to refuse a lifesaving treatment such as a blood transfusion? How can a commitment to beneficence be reconciled with allowing a patient to die when a simple lifesaving procedure could have been performed? A possible response to this conundrum is that a patient's autonomy over their life supersedes any duty of a doctor to be beneficent and non-maleficent. It might also be said that it is more maleficent to force a patient to undergo a non-consensual treatment than it is to allow them to die. This is because autonomy is highly normatively significant: respecting patients' autonomy is more important than saving their life.

Autonomy – understood as the freedom to make choices about one's own life – is intrinsically valuable (for example, even if a patient dies as a result of their refusing treatment, it is still prima facie good to have followed their wishes), as well as extrinsically valuable (it promotes positive outcomes such as feelings of dignity, worth, and (often) the avoidance of harm). Generally, people with capacity know what they want, and giving them the autonomy to make decisions about their own life and care is a way of giving them what they want (so long as it is not impractical and does not infringe on the rights of others). Consent-seeking is the ideal means of giving people the autonomy to make choices about their life and their body.

Failing to obtain patient consent prior to care violates patients' autonomy; in the same vein, obtaining patients' (ostensible) consent through threats or coercion must be avoided (Nursing and Midwifery Council 2002: 6). The normative importance of consent and autonomy are not solely confined to medical situations though.

Philosophers claim that obtaining legitimate sexual consent upholds individuals' autonomy (Steutel 2009, Steutel and de Ruyter 2011). For sex to be permissible, it needs to support the autonomy of both parties, and therefore requires consent. Permissible sex involves not just consent, but recognition of 'sexual agency' (Cahill 2016) and autonomy (if one is not free to consent or withhold consent, then the consent is not legitimate). Philosophers often link rape and sexual assault with the curtailing of autonomy. For example, Feinberg (1987: 10–11) writes that rape (like other violent crimes) is harmful because it limits the autonomy of its victims. Although rapes are often physically and emotionally damaging, this is not *always* the case. Consider an

example where a woman is raped while unconscious, the rapist uses a condom, and the victim does not suffer physically or psychologically, because she is unaware that anything untoward has occurred (Gardner 2007: 5). Despite the lack of harm, this sort of occurrence still seems morally wrong, because the victim did not – and could not – consent. In other words, the wrongness of this sort of rape is not its harmfulness (because there was none); the wrongness is that the rape has curtailed the victim's autonomy to decide with whom she has sex.[127] In this way, we can see that consent is normatively significant because (sexual) autonomy is so valuable.

The same can be said for medical autonomy, and we could sketch out an analogous example where a known Jehovah's Witness is given a blood transfusion while he is unconscious, and is never made aware of this. This too seems intrinsically wrong even if no harmful consequences ever come of it. In nursing situations as well as sexual and medical situations, it is therefore essential to respect people's autonomy via consent-seeking; autonomy is seen as "the foundation of nursing care" (Reed and McCormack 2012: 9).[128]

Claiming that consent-seeking promotes autonomy can be problematic because sometimes, contravening patients' consent and ignoring their

[127] It could be suggested that people do not have any autonomy whilst unconscious, and so actions performed on them do not curtail their autonomy *qua* their ability to decide their own actions (Eyal 2019: §2.2). However, if we have autonomy over our lives in general, then one can, while conscious, choose what happens to them while they are unconscious.

[128] NHS guidelines state that an individual has the right to choose (or refuse) to undergo a medical procedure even if their choice puts their health or their life in danger. Exceptions to this include if someone with an eating disorder refuses treatment, they can be considered to lack capacity (NHS 2022)

autonomy in the short term can yield an increase in (future) autonomy for the patient. For example, suppose Diego refuses a simple treatment (say, taking a tablet) for a life-threatening ailment. Disregarding his autonomy and performing the treatment anyway will grant him an extended life in which he will have autonomy: ignoring Diego's wishes thus *increases* his overall autonomy (Velleman 1992). However, this does not mean that doctors would be justified in ignoring Diego's non-consent simply because doing so maximises his autonomy. Autonomy is intrinsically valuable, but we cannot infer that one is justified in violating it in order to yield a future gain in autonomy. Rather, healthcare professionals who value patient autonomy should (and generally *do*) take a more deontological approach and avoid violating patient autonomy, even if doing so would bring about increased patient autonomy in the future. Moreover, generally (unlike the unconscious rape and the Jehovah's Witness examples) a person such as Diego knows if their autonomy has been restricted, and this could diminish their dignity and wellbeing (something I return to below).

This does not mean that healthcare professionals should 'respect autonomy' and do absolutely anything a patient desires. Consent is the power of veto only (Wilkinson, Herring, and Savulescu 2020: 89), and does not entail that we should treat people in inhumane ways simply because they consent to it or request it. Rather, if someone requests treatment not befitting a rational being with universal dignity, we should not simply 'respect his autonomy' and do as he requests. (Kant 1996a: 80 [4:429], 1996a: 434–435, 2011: 97–99); instead, we should continue to treat him well.

3.2 Bodily integrity

People have bodily integrity – the right to decide what happens to their body. They can decide whether another person is allowed to touch them, and can make decisions about their physical health. This is not just a moral right but also a legal one: the right to bodily integrity is laid out in the *European Convention on Human Rights* (1950: Section 1, Article 5), and is supported by the Royal College of Nursing's principles of consent (2017: 5) and the General Medical Council's guidelines on good medical practice (2013: 8).

The importance of bodily integrity has an esteemed place in philosophical literature too. Mill (1974: 69) writes that "over his own body and mind, the individual is sovereign", and Locke (1988: v.27, 305) writes similarly that "every Man has a Property in his own Person". These statements can be interpreted to mean that we have ownership rights over our own bodies by default. This idea of bodily ownership is echoed in laws which refer to 'trespass' to the person – a term equally applied to unlawful entry to land or property. For example, in English law, a battery involves touching someone's body non-consensually, regardless of harm caused (see Elliott and Quinn 2007: 300–306). Thinking of the body as one of our possessions – perhaps our most prized possession – helps to explain why consent is required for even non-dangerous touches, such as a hand on the thigh. Almost no one is injured by someone placing a hand on their thigh, but it requires consent in order to be permissible because we have bodily integrity.[129] Similarly, no harm is caused

[129] Emergencies are exceptions to this. If I have been shot in the thigh and am semi-conscious, a stranger may legitimately press his hand onto my thigh without consent to stem the blood flow in an attempt to save my life. This mitigation of harm would not violate my bodily integrity or my dignity like a sexually-motivated touch could.

if a stranger enters my house without my permission, calmly walks around a little, then leaves, but this type of trespass – just like bodily trespass – seems intuitively wrong, because of the lack of consent.

The idea of ownership over our bodies (like property) also helps to capture why some acts are more wrong and emotionally harmful than others. If Eric uses Yasmin's Ferrari without her consent, this is more troubling than if he had used Yasmin's stapler without her consent (excepting bizarre circumstances or emergencies). Similarly, if Eric touches Yasmin's genitals without her consent, this is more troubling than him touching her elbow without her consent. This is because, just as with property, we place greater value on some body parts compared to others, and violating these areas is more troubling to us. This is also evident in the routine care activities pyramid I outlined in §1, where activities are ordered into a hierarchy based on their level of intimacy: someone showering you without your consent is more of a violation than someone putting a coat on you without your consent.

Bodily integrity is underpinned by autonomy: people have the right to choose what happens to their body, even if their choices are not in their best interests:[130] Diego has bodily integrity and can thus refuse to take a lifesaving tablet if he so desires. Similarly, if a patient elects to undergo a medical procedure in the understanding that there is a danger of infection, paralysis, death (etc), then he is at liberty to do so, because it is his body to do with as he sees fit.

[130] There are exceptions – see previous footnote.

When considering *why* we have the right to decide what happens to our bodies, explanations often have a distinctly Kantian feel about them. Numerous writers refer to Kant's *Formula of Humanity* (1996b: 80 [4:429]) which states that humans have a unique normative status because of our rationality, and thus deserve autonomy over our bodies merely in virtue of being human. This sort of Kantian terminology is evident in the work of Donagan, for example, who writes that "no human being may legitimately be interfered with [because of his] unique dignity" (Donagan 1977: 31). I would suggest that our universal dignity – which we have in virtue of being human – gives us a high normative status, granting us autonomy and bodily integrity. Consent-seeking can avoid violations of bodily integrity, and is a way in which we can promote people's dignity, which I now discuss.

So far in this section on the normative significance of consent, I have argued that consent is morally important because we have autonomy and bodily integrity. As the above paragraph suggests, both autonomy and bodily integrity are linked with our dignity, discussed presently. After that I will suggest that consent is also morally significant because it builds interpersonal trust.

3.3 Dignity

In the previous chapter I examined two varieties of dignity (universal and variable) and their moral importance. This sub-section provides a link between dignity (of both types) and consensual care activities; this link is explored in greater depth in the next chapter, but is suggested here briefly.

Universal dignity grounds both autonomy and bodily integrity insofar as one has autonomy and bodily integrity because one has universal dignity – a high normative status. Respecting someone's autonomy and bodily integrity also promotes their variable dignity. Someone who has not had their autonomy or bodily integrity respected – such as Diego, who is forced to take a lifesaving tablet against his wishes – has not had their universal dignity respected. As a result, Diego's variable dignity might be reduced because of what he has undergone.

It has been suggested that treating someone with dignity means no more than respecting their autonomy (Macklin 2003). Although I do not subscribe to the position that dignity is reducible to respecting autonomy, it is clear that respecting autonomy is a crucial feature of (variable) dignity promotion, since it is difficult to see how a person could be treated with dignity if their autonomy is not being respected.[131] Consent-seeking is an essential way to respect patients' dignity in medical and care contexts (Gilabert 2019, Cochrane 2010, Beyleveld 2001).

Giving patients the autonomy to make decisions and have control over their care and their body is empowering; this helps promote their mental wellbeing and variable dignity. This standpoint is maintained by organisations such as the National Institute for Health and Clinical Excellence (2013) and the Royal College of Nursing (Baillie, Gallagher, and Wainwright 2008); it has also been

[131] If someone were to (autonomously) ask to be tortured, humiliated (etc), then *not* doing as they ask respects their universal dignity and helps to promote their variable dignity more than doing as they request would. But as a general rule, respecting (universal) dignity means respecting autonomy.

maintained by elderly people themselves (Matiti and Cotrel-Gibbons 2006, Enes 2003, Woolhead et al. 2005). The Department of Health in the UK endorses a 'human rights approach' to care, which involves five core principles: Freedom, respect, equality, dignity, and autonomy (FREDA) (Care Quality Commission 2014, Curtice and Exworthy 2010). The promotion of patients' (variable) dignity is thus at the forefront of routine care; failing to obtain patient consent prior to undertaking care activities is morally problematic.

Consent-seeking is morally necessary because of its links with dignity: consent-seeking recognises that someone has the high normative status that comes with universal dignity (we do not, for example, seek a rabbit's consent before picking it up, or seek a cow's consent to be milked). Touching people's bodies or involving ourselves in their affairs without their consent is disrespectful, and could make them feel objectified and infantilised (Nussbaum 1995: 257); this is likely to reduce their variable dignity, as they feel they have been treated like a child, animal, or object rather than a human being with thoughts and autonomy.

I do not provide a sustained argument here, since the next chapter is wholly focused on the link between dignity and consent, wherein I maintain that promoting patients' variable dignity is the crucial reason why consent-seeking must precede routine care. In other words, consent-seeking makes routine care permissible (or morally good) because it respects our high normative status (universal dignity). I suggest that non-consensual routine care can reduce a patient's variable dignity – and that these reductions in dignity can be cumulative.

3.4 Trust and emotional care

Bodily integrity, autonomy, and dignity are all intrinsic reasons why consent is important; trust, however, is a pragmatic, consequentialist reason for obtaining legitimate consent. Unlike almost all other professions, doctors swear an oath to act morally (Encyclopedia Britannica 2017); nurses make a similar commitment to ensuring patient welfare (Nursegroups 2022). Perhaps because of this oath – or perhaps because we have little other choice – we place a remarkable amount of trust in healthcare professionals (O'Neill 2002: 16), trusting them with our health, our bodies, and sometimes our lives. If doctors and nurses obtain patients' consent before touching them or providing treatment and care, this helps to reinforce this trust. It has been suggested (see Eyal 2014) that trust in healthcare professionals – both individually and generally – is necessary for a smooth-running healthcare system. But whilst trust in one's doctor is *ideal*, it is not *essential* in medical relationships, for patients often consent to medical procedures without completely trusting their doctor (Bok 2014, Roache 2014).

We are more likely to trust someone when we believe they emotionally care about us. However, this is not a necessary condition for trust, since trust is often activity-specific or domain-specific (Jones 2012, D'Cruz 2015: 479); I can trust the taxi driver to get me to my destination without trusting him *in general* or believing he emotionally cares about me. In Chapter 4, I distinguished between emotional care and practical care, and I acknowledged that human nurses may often provide both these types of care, but I argued that (genuine) emotional care is not an essential feature of nursing, and that emotionless carebots can still be good substitutes for human nurses. I noted,

that in cases where genuine emotional care is not felt (by carebots or nurses), a simulation of emotional care ('fake compassion') is a form of prosocial deception which can improve the wellbeing of patients, and helps to build trust (Levine and Schweitzer 2015).

Consent-seeking is one way in which a nurse or carebot can give the appearance of emotionally caring about patients, because consent-seeking often symbolises concern for the patient's wishes and autonomy. Thus, a carebot which seeks consent at apt moments does a far better job of simulating emotional care; this is likely to increase trust in the carebot by the patient. Whether this trust is the same (interpersonal) sort of trust a patient has towards a human nurse, or whether it is more similar to how someone trusts their car – which is more akin to mere reliance (Hawley 2015: 798, Wright 2010: 616–617) – will depend on a number of factors, including whether the patient views the carebot as an agent or as a tool; either way, consent-seeking by the carebot can help to bolster patients' confidence and trust in the carebot.

The doctor-patient (or nurse-patient) relationship is often hierarchical because of the knowledge gap and power imbalance; healthcare professionals not only administer treatments, but also act as gatekeepers regarding what treatments and procedures are offered. One way of redressing this power imbalance and building trust is to avoid dominating patients, by seeking consent prior to treating or touching the patient. This non-domination within a relationship is

also the reason for consent-seeking in sexual matters: when consent is obtained, one partner is not dominating the other.[132]

In most cases, it is wrong for healthcare professionals to jeopardise patient trust (Eyal 2014). When a healthcare professional undermines a patient's trust by acting without consent, this may not only damage the patient's relationship with that particular person, but possibly with *all* healthcare professionals. For the smooth-running of medical care, patients must (and generally do) put their trust in not just individual healthcare professionals, but in medical institutions as a whole (Ibid). When violations of patient consent are widespread or extreme, the reduction in trust of healthcare professionals *qua* healthcare professionals can resonate across whole societies. An example of a widespread trust violation is the practice of non-consensual pelvic examinations[133] by hundreds of student doctors on unconscious patients (Picard 2010, Thomson-DeVeaux 2010); any patient undergoing anaesthesia in the area where this practice occurred may rightly feel reduced trust in their doctors.

A more extreme case – but perpetrated by an individual – is Dr Harold Shipman, who is thought to have murdered around 250 of his patients with lethal injections (Encyclopedia Britannica 2022). What Shipman did was seen as abhorrent and treated as serial murder not because he was violent – it is not thought that he used any kind of force or violence on his victims – but

132 Though some people may practise a *simulation* of dominance and control for sexual pleasure.

133 For practice, student doctors inserted their fingers into the vaginas of anaesthetised patients undergoing operations unrelated to vaginal or pelvic problems. The patients were not told about this.

because the patients whose lives he took had not requested it, and killing them was not in their best interests (and because he *intended* to kill them). Doctors regularly administer lethal doses of painkillers at patients' requests or in patients' best interests,[134] but doing so without consent, and against patients' best interests, as Shipman did, undermines trust in the profession to the greatest extent possible. Trust in doctors was so seriously undermined by the Shipman case that regulations were changed in the UK: all unexpected deaths are now referred to a coroner to verify the cause of death claimed by the doctor (Batty 2004). Before Shipman's murders were discovered, a doctor's word was trusted and the cause of death she stated was not questioned; this is no longer the case, so gross were Shipman's violations of patient trust and consent.

Clearly, consent-seeking is important because it helps to build trust, not just between the patient and the individual healthcare professionals involved, but public trust in healthcare professionals generally – something which is essential to good medical care (O'Neill 2002, Jackson 2001, Tännsjö 1999).

4 Conclusion

I began this chapter by outlining routine care activities and ordering them into a pyramid; this pyramid shows the sorts of activities involved in routine care – the sorts of activities for which carebots will need to seek consent. This pyramid will remain useful in the next chapter too.

[134] In countries like the UK where active euthanasia is illegal, doctors invoke the doctrine of double-effect. Doctors cannot *intentionally* kill a patient, but they are permitted to administer a lethal dose of painkillers, so long as pain relief is the intention, and death is foreseen but 'unintended' (Klein 2005, Price 1997).

This chapter has also examined the features of legitimate consent: we have seen that legitimate consent must be voluntary, sufficiently informed, and given by a person with capacity; one must mentally sanction the activity in addition to behaviourally signalling one's consent. I discussed differences between lasting consent (where activities are consented to en masse for a set duration), and repetitive consent (where consent must be obtained for each activity); this is something I return to in the next chapter when I discuss routine consent.

This chapter has also explored the normative significance of consent. We have seen that consent has extrinsic value because consent-violations can lead to negative consequences such as undermining trust or criminal charges, but consent also has intrinsic value, because it supports autonomy, bodily integrity, and dignity. Even if a consent violation – such as giving a blood transfusion to a Jehovah's Witness – is never discovered and no negative consequences ensue, the act seems intuitively wrong, because consent was not obtained.

My final chapter builds on the groundwork laid out in this chapter. Despite there being some similarities between sexual, medical, and routine consent, the next chapter argues that accounts of sexual and medical consent cannot simply be transposed to provide us with sufficient understanding of how, when, and why carebots should seek routine consent. What I have not done in this chapter – but I will in the next – is demonstrate the likely and significant reduction in patients' variable dignity when nurses or carebots provide routine care without consent.

Chapter 7

Consent and dignity: How consent-seeking promotes dignity

> Winifred is chatting to her friend Peggy when the nurse begins feeding Winifred unexpectedly. Winifred is used to it. The nurse never asks whether Winifred *wants* feeding or bathing or undressing – he just does it. He's never rough or cruel to Winifred – in fact, he's quite gentle and otherwise pleasant – but Winifred can't help feeling like a toddler or an object every time he does something to her without even *asking* if she wants it. It feels wholly undignified.

The focus of this chapter is non-consensual routine care: the sort of thing that happens to Winifred with the nurse in question. I argue that consent-seeking should always precede routine care, else the effects on patients' variable dignity can be profoundly detrimental. Winifred's plight is a fictitious example, but it is likely that there are at least some real-life cases which echo hers. The nurse's behaviour is not the ideal behaviour we would want carebots to emulate.

The previous chapter examined some of the key features of consent; I showed that for consent to be legitimate, it needs to be sufficiently informed and given voluntarily by a person with capacity – and that it needs to be expressed, not

merely thought or felt.[135] We saw that consent is morally transformative, and supports autonomy and bodily integrity – both of which contribute to one's dignity;[136] consent-seeking also develops trust between the parties involved.

In this chapter, I demonstrate that routine care is an important area of philosophical study, and I show why it is essential for carebots and nurses[137] to obtain consent before undertaking routine care activities: this is because of the potential deleterious effects that non-consensual routine care[138] can have on patients' dignity.

Although routine consent is normatively important, it is largely absent from the philosophical literature. This chapter demonstrates that a normative account of routine care is needed – and provides one. I examine the following possible reasons why routine consent is absent from the literature, and demonstrate why such claims are mistaken:

(1) Routine care activities are sufficiently similar to medical care or sexual activity, so accounts of medical or sexual consent can be applied to routine consent situations

135 In emergencies, some or all of these conditions may not be necessary.

136 In this chapter, 'dignity' refers to variable dignity, unless otherwise stated.

137 At times I make normative claims about how nurses ought to behave: these claims pertain to carebots too. Descriptive claims about how nurses actually behave pertain only to nurses.

138 In this chapter, 'care' refers to practical care only (not emotional care), unless otherwise stated. 'Routine care' refers to help with daily activities such as feeding, toileting, bathing, dressing, and moving about, as laid out in the pyramid in Chapter 6 §1.

(2) Routine consent is not worthy of scholarship because:

(a) it is intuitive

(b) the stakes are low

In §1 I address point 1, by examining some of the conceptual and normative similarities and differences between routine consent and medical and sexual consent. If routine consent were sufficiently similar to either of these, then we could simply transpose existing normative accounts of consent onto routine care situations, and say that whatever moral principles or practical guidelines apply for medical or sexual consent, also apply for routine consent. I examine some features of medical and sexual consent, including what counts as being sufficiently informed, how frequently consent must be obtained, how consent is obtained, and whether consent can be transferred to another person. I also revisit discussions of lasting and repetitive consent and consider which best applies to routine consent. We see that although there are some loose resemblances between routine care, sexual activity, and medical care, routine consent is sufficiently dissimilar to mean that existing scholarship on medical and sexual consent are inadequate for explaining the normative importance of routine consent.

In §2 I address point 2a – the claim that routine consent is unworthy of scholarship because it is intuitive. I argue that consent is not necessarily intuitive for human nurses, but even if it were, it is nevertheless useful to explore this philosophically, and for the issue to be clarified for carebots, who do not have intuitions. My response to point 2b is covered in §3, and relates to the pyramid of care activities I outlined in the previous chapter: I argue that providing highly intimate routine care without consent risks a greater

detrimental effect on patients' dignity than less intimate non-consensual care does (but all non-consensual routine care risks causing *some* reduction in dignity). I maintain that the effects on dignity from non-consensual care are cumulative, and because routine care activities take place so frequently in residential homes, non-consensual routine care could cause a severe and long-lasting reduction in patients' dignity. I also consider how frequently, and in what conditions consent must be obtained in order for routine care to be morally permissible.

This chapter thus demonstrates that consent is just as normatively important in routine care situations as it is in medical and sexual situations, and obtaining legitimate consent is essential for routine care to be permissible, and that the study of routine consent is both necessary and valuable.

1 Are existing accounts of consent sufficient?

It would be convenient and satisfying if there were an existing normative account of consent which perfectly captured the reasons why routine consent is morally essential, and when it should be obtained. Unfortunately, it is not possible to simply transpose an existing normative account of consent – such as sexual or medical consent – and say that whatever normative principles apply for sexual or medical consent also apply for routine consent. Routine consent shares *some* features with medical and sexual consent, but there are stark differences too. This section discusses some similarities and differences between medical consent, sexual consent, and routine consent, and demonstrates which aspects of our understanding of medical and sexual consent can help to inform a normative account of when and why routine consent should be obtained.

Sexual and medical activities differ from one another in kind, not merely in degree. This is clear because we can point to highly invasive medical procedures (organ transplantation) and highly invasive sexual acts (anal sex), as well as fairly non-invasive medical activities (examining a cut finger) and fairly non-invasive sexual acts (a hand on the thigh). This evidences that sexual acts and medical procedures do not simply differ in degree from one another. Routine care shares some features with medical care, such as being motivated by a desire to help the other party, and often (though not always) taking place within a care institution by paid professionals. For this reason, one might think that routine care activities do not differ in kind from medical procedures, but only in degree – that routine care activities are less invasive than medical procedures. However, they too differ in kind, since we can point to some highly invasive routine care activities (genital washing) and some fairly non-invasive routine care activities (cutting up food).

Although the general essence of what consent *is* remains the same for medical consent and sexual consent (it is the sanctioning of an event which affects oneself, and is by default not sanctioned) this does not entail that the normative rules and understandings we have regarding medical consent are applicable to sexual consent, nor vice versa. It is therefore unsurprising that the literatures on medical consent and sexual consent are distinct from one another.[139] The same is true for routine consent: although consent is still a

[139] There are a few examples which seem relevant to more than one literature, such as when student doctors non-consensually inserted their fingers into the vaginas of unconscious patients (Thomson-DeVeaux 2010, Picard 2010). The acts were medically motivated, however the nature of the examination means that non-consensual cases could be called sexual assaults, and are thus relevant to both

performative act which sanctions an event affecting oneself which is usually not sanctioned, we cannot get a complete understanding of the features or normative significance of routine consent by examining medical or sexual consent literatures. The literatures on sexual consent and medical consent provide a useful starting point for developing a normative account of routine consent; however, they are insufficient to provide us with a complete understanding of when and why consent-seeking should occur.

Below I outline some ways in which routine consent differs from medical and sexual consent, and some similarities between them. The features I consider are:

- Being sufficiently informed
- Interpersonal transferability
- How consent is obtained
- Frequency and repetitive consent

1.1 Being sufficiently informed

Some medical procedures – such as open-heart surgery or leg amputation – are complex, and the doctor must thoroughly explain the procedure to the patient. She should outline the reason for the procedure, what will occur, its duration, possible prognoses, potential dangers, and recovery times (General Medical Council 2020: 11). This is necessary because many medical procedures are complicated one-off occurrences for patients, who are probably unfamiliar with the procedure – but they need to *become* familiar with

literatures. Nevertheless, the literatures on sexual consent and medical consent are almost always distinct.

it to give legitimate (sufficiently informed) consent. For complex procedures, the doctor should explain the process verbally in addition to providing written documentation. Even for more minor procedures, such as cervical screening, vaccinations, or wearing a heart monitor, the healthcare professional should still provide a verbal explanation – particularly if patients' records show this is their first experience of the procedure – and should provide written information such as a pamphlet (or signpost to online information) if required (Royal College of Nursing 2017).

Sexual consent is different. Although the law requires both parties to consent prior to sexual activity, it is not morally required for people – even new sexual partners – to explicitly outline what sex with them involves, its duration, possible outcomes, dangers, and suchlike (as one does with medical procedures). This is because adults know – or are assumed to know – what sex involves. Even when one partner seeks consent to φ (where φ is a particular type of sexual activity) and the other partner is unfamiliar with φ-ing, it is not morally required to provide detailed information. Certainly, people do not provide written explanations or pamphlets to their sexual partners. This means that often, sexual consent may not be *fully* informed, but it is sufficiently informed to count as legitimate consent, and to render the sex acts permissible.[140]

[140] There are no formal guidelines regarding how much information sexual partners should provide prior to sex, and there is very little philosophical discussion of the matter, though Tilton and Ichikawa (2021) discuss miscommunication and insufficient sexual information-giving (see also Dougherty (2013) and subsequent responses for discussion about lies and misdirection in sexual information-giving).

Routine consent has much in common with sexual consent insofar as detailed explanations of what activities involve are not morally required. Nurses do not need to provide detailed information about what taking a bath entails in order for the patient to make an informed decision about taking a bath, because patients are familiar with, or are assumed to be familiar with, bathing (and other routine care activities). We might therefore think that with regard to information-giving, whatever normative rules are true of sexual consent (viz. the amount of information morally required in order for consent to count as informed and therefore legitimate) can be applied to routine consent situations.

However, informed consent in sexual matters is not solely about knowing what sex involves – it is also about knowing information about one's partner. If Philip lies to Harriet by saying he is an animal-lover when really he is a keen hunter, knowing that Harriet will only consent to sex with an animal-lover, then Harriet's consent is not fully informed, so not legitimate (Dougherty 2013: 728). This is the case even if the physical sex which occurs is exactly what Harriet agreed to.

Medical consent is different: the doctor's religion, political leanings, looks, hobbies (etc) are not usually morally significant factors to consider when patients deliberate whether to consent to a procedure (though the doctor's professional expertise is very relevant).[141] So, in medical settings, being fully informed is primarily about understanding the *procedure*, whereas being fully

[141] Background details of an *extreme* nature may be morally significant. This issue was explored in an episode of *Star Trek: Voyager* ["Nothing Human" S5, Ep8] (Livingston 2000), when a war criminal who experimented upon prisoners is now the only person who can save the life of Lt Torres; Torres rejects his help because of her moral objections to his past.

informed in sexual settings is only *partly* about understanding what will take place, and partly about understanding some salient details regarding one's partner. In this sense, routine consent is more similar to medical consent: if a nurse is going to bath Mildred, then Mildred does not need to know about the nurse's religion, political leanings, or hobbies; nor is the nurse's appearance of any relevance (though the nurse's gender might be relevant – this is discussed further in §1.2, below).

We can see from the above discussion that routine consent differs from medical consent regarding how much information is provided in order to make consent fully informed (and therefore legitimate), but it differs from sexual consent regarding the sort of information which is morally relevant about the consent-seeker. This means that philosophical work on sexual and medical consent (vis-à-vis information-giving) cannot be simply transposed to provide us with an apt understanding of routine consent and information-giving.

1.2 Interpersonal transferability

One key normative difference between sexual consent and medical consent is whether consent is interpersonally transferable: by this I mean whether it is morally permissible to transfer consent from one individual to another. Let Sφ represent a sexual act, and Mφ represent a medical procedure. If X consents to Sφ-ing with A, we cannot permissibly assume that X also consents to Sφ-ing with B, even if A and B are equally skilled at Sφ-ing. If A is unavailable to Sφ, it would be (pro tanto) morally wrong for B to stand in for A without re-seeking consent from X: the consent given to A cannot simply transfer to B and remain morally permissible. For example, if Aisling has consented to sex with Joel, we cannot permissibly assume that Aisling also consents to sex with

Ryan, even though Joel and Ryan are equally skilled at sex.[142] This is because – as I noted in the above sub-section – sexual activity is seldom just about the physical feeling of the act itself; someone's looks, personality, and relationship to the other party are normatively significant in sexual matters. So, there is a moral presumption against sexual consent being interpersonally transferable.

Medical consent is different: consent can (pro tanto) be interpersonally transferred and remain morally permissible. If X consents to A Mφ-ing, it is (pro tanto) morally permissible to assume X would also consent to B Mφ-ing, assuming A and B are equally skilled at Mφ-ing. For example, if Dr Andrews and Dr Brown are equally skilled surgeons, and Debbie has consented to Dr Andrews performing surgery on her foot, it would be (pro tanto) morally permissible for Dr Brown to perform the surgery if Dr Andrews should be unavailable on the day; Debbie's consent can be morally permissibly transferred from Dr Andrews to Dr Brown. This is because one generally consents to a medical procedure on the proviso that it is performed by a sufficiently competent person; the specifics of whether it is performed by Dr Andrews or Dr Brown are not usually morally significant, so there is a moral presumption that medical consent *is* interpersonally transferable.

Another way of saying this is that medical consent is usually *de dicto*, where "I consent to a doctor amputating my foot" usually means the speaker consents to *any* competent doctor performing the amputation. Contrastingly,

[142] There may be cases where people do not care with whom they have sex (e.g. if using a sex worker), however there remains a moral presumption that sexual consent cannot transfer to another person. In other words, if sex worker X is unavailable, sex worker Y is morally required to seek consent to stand in for X – we cannot permissibly assume transferability.

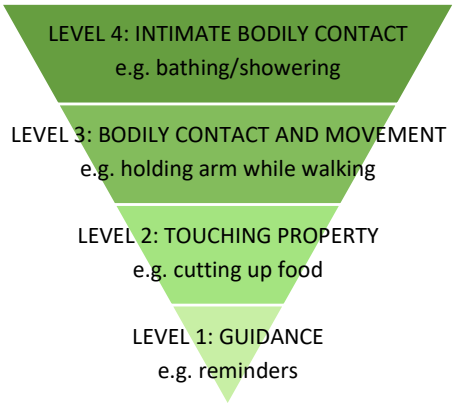
sexual consent is usually *de re*, where "I consent to sex with a man" usually means the speaker consents to sex *with one man in particular*. Things which might be 'deal-breakers' in sexual matters – such as a person's looks, sex, age, behaviour, hobbies, or political leanings – are not generally 'deal-breakers' in medical matters (see footnote 141 for an extreme exception). If someone would have made the same choice (to have the foot amputated; to have sex) had they known more information about the other person, then substitution of people is morally immaterial (Alexander 1996: 167–168). In other words, if S would consent to any competent doctor Mφ-ing, then their consenting to X Mφ-ing can be permissibly transferred to any competent doctor.

However, medical consent is not *always* interpersonally transferrable. Notable exceptions include highly personal examinations – for example, cervical screenings or prostate examinations.[143] A male nurse may be equally skilled at carrying out cervical screenings as a female nurse is, but this does not mean we can permissibly assume a patient who consented to Georgina conducting the procedure would just as readily consent to George conducting it. The health professional's sex or gender may be morally significant in highly intimate matters. However, studies suggest that in midwifery and gynaecology, less than half of women prefer female professionals – most have no preference (Kerssens, Bensing, and Andela 1997); in general practice, only a quarter of female patients prefer female doctors. Most patients prioritise other factors such as communication skills and knowledge, rather than

[143] Both procedures involve physical penetration. With cervical screenings, a speculum is inserted into the vagina, and the cervix is brushed, to gather cells. With prostate examinations, a finger is inserted into the man's rectum to feel his prostate gland.

sex/gender [144] (Bourke 2002, Kerssens, Bensing, and Andela 1997). Nevertheless, there is a moral presumption against assuming interpersonal transferability of consent in highly personal medical examinations: consent must be re-sought if the activity is to remain permissible.

In terms of interpersonal transferability, routine care has more in common with medical consent than sexual consent. Most medical procedures and most routine care activities (Levels 1-3, see diagram below, or Chapter 6 for more information) do not involve seeing or touching the genitals, and in such cases, there can be a moral presumption that consent is interpersonally transferable. But for Level 4 routine care activities which involve intimate bodily contact (and intimate medical procedures, such as cervical screening and prostate examinations) patients might prefer a healthcare professional of their own sex or gender if one is available – thus for Level 4 activities, there should be a moral presumption against consent being interpersonally transferable. Even if Matthew and Bertha are equally good at bathing patients, and both seem equally compassionate and pleasant, female patients may still legitimately prefer Bertha to Matthew. Of course, patients probably have no choice in who is employed at the residential home, or what shift patterns the nurses work: if only female nurses are on duty on Tuesdays, and a male patient requires help



[144] Bourke's study assessed preferences for the *sex* of the GP; Kerssens et al's study assessed preferences for the *gender* of the GP, though these may have been the same.

with toileting, that help must come from a female nurse, which could be embarrassing. Hopefully, the development and use of (genderless) carebots which can help patients with toileting could mean an end to such embarrassing – but probably fairly common – predicaments.[145]

Recall that we are currently examining similarities and differences between medical consent, sexual consent, and routine consent, in order to ascertain whether routine consent shares enough features with either sexual or medical consent for us to say that whatever normative principles apply to sexual or medical consent, these can simply be transposed to provide a normative account of routine consent. So far we have seen that in terms of information-giving, routine consent has more in common with sexual consent, but in terms of interpersonal transferability, routine consent is more similar to medical consent. This will mean that these literatures could help to provide some understanding of the intricacies routine consent, in addition to the normative account I provide herein (in §3).

I now consider how consent is obtained in medical and sexual situations, after which I will examine the frequency of medical, sexual, and routine care activities. This will demonstrate that routine consent is sufficiently distinct from both medical and sexual consent so as to make it inappropriate to transpose

[145] There is a propensity to make robots gendered rather than androgynous – through their appearance, voices, or both – which may exacerbate sexism and gender stereotyping (Robertson 2010). It will be interesting, in the coming years, to see whether medical-bots and carebots are made to emulate a particular gender, and whether this causes any problematic repercussions: would men be comfortable having their prostate examined by a robot with a female voice? Would women be comfortable being bathed by a male-presenting robot?

our normative rules from one domain to another (though some literatures may provide a useful starting point). We will thus require a fresh normative analysis which demonstrates not only the normative significance of routine consent, but also when it should be obtained – something which I provide later in the chapter.

1.3 How consent is obtained

In §2.1 of the previous chapter, I discussed whether consenting is a mental state or performative act, and I argued that legitimate consent is a performative act *in addition to* the mental state (of sanctioning φ), plus certain other psychological conditions (voluntariness, capacity, and being sufficiently informed).[146]

Medical consent can be given in a number of different forms, such as verbally, non-verbally or in writing, and all forms are equally legitimate (Royal College of Nursing 2017: 10). Medical consent-seeking for complex procedures such as limb amputation typically involves the patient signing a written consent form (General Medical Council 2020).[147] In less serious and less complex cases such as blood tests, it is permissible for healthcare professionals to obtain verbal or non-verbal consent (General Medical Council 2020: 9–10). A nurse may say “Are you ready for the blood test?” and the patient could consent verbally by saying “Yes” or non-verbally by nodding or holding out her bare

146 The psychological conditions and the mental sanctioning may not always be simplistic to discern, but must nevertheless be present in order to legitimate the consent-signalling performative act.

147 In some emergency medical situations, consent is deemed unnecessary because the patient is unconscious or otherwise unable to signal their consent (NHS 2018a).

arm (Royal College of Nursing 2017: 10, Wilkinson, Herring, and Savulescu 2020: 95–96). (Non-consent can also be non-verbal, for example a patient who zips up her jacket and folds her arms when a blood test is required is signalling her non-consent to the blood test – although her having attended the appointment does give mixed signals!)

Sexual consent is seldom, if ever, obtained in writing: one reason for this is that consent can be withdrawn at any time, for any reason, so written consent would not prove that the entirety of the sexual act was consensual.[148] Like medical consent, sexual consent may be implied or non-verbal, by initiating an act or responding enthusiastically to another person's advances (Kavanagh 2016: 43).

Consent to routine care activities is obtained verbally or non-verbally, in the same way that sexual consent and low-risk medical consent is obtained. It is worth noting that the UK nurses' code of professional behaviour states that nurses should "make sure that [they] get properly informed consent and document it before carrying out any action" (Nursing and Midwifery Council 2018 Principle 4.2). I would suggest that although it is morally required to obtain consent for all routine care activities, *documenting* that consent is unfeasible to the point of absurdity: a nurse does not need to document that the patient agreed to getting dressed, to having his face washed, and to being supported while he walked.[149] I suspect that in practice, routine consent is

[148] Medical consent can also be withdrawn at any time, including after signing the consent form, meaning that consent forms are not contracts (Wilkinson, Herring, and Savulescu 2020: 95)

[149] Carebots could document patients' consent to routine care, by recording audio or videos of patients – though this may be considered unnecessarily intrusive.

seldom documented, but this seems unproblematic because documenting the consent is not a moral requirement: it does not elevate the moral legitimacy of the activity.

Clearly, routine consent-seeking has some features in common with sexual consent-seeking and medical consent-seeking for low-risk procedures, in terms of how consent is obtained and (not) documented – however, this does not entail that either of these literatures is apt to be wholly transposed to elucidate routine consent-seeking.

I do not further explore *how* carebots should seek consent, nor how they should make sense of the consent-giving (or withholding) which they observe from patients; to suggest that consent could be sought verbally or non-verbally would seem to suffice. Further work – either in philosophy or another discipline – could investigate how carebots can best understand human body language and non-verbal communication, but that is beyond the scope of this thesis; I do, however, examine when and why carebots should seek consent in §3 of this chapter. For now, I consider the frequency with which routine consent-seeking must place.

1.4 Frequency and repetitive consent

One substantial way in which routine consent-seeking differs from medical consent-seeking and sexual consent-seeking is its frequency. Every day, nurses may help a patient in a residential home to get out of bed, go to the toilet, brush his teeth, shower, dress, comb his hair, shave, eat breakfast, walk to another room, go to the toilet, eat lunch, walk to another room, go to the toilet again, eat a snack, go to the toilet again, eat dinner, go to the toilet again,

walk to another room, wash his face, get changed for bed, brush his teeth, and get into bed… and more besides. The number of times a patient receives routine care therefore far exceeds the number of times a person engages in sexual activity or has medical procedures performed on them. Therefore, routine consent must be obtained with far greater frequency than sexual or medical consent.

In §2 of the previous chapter, I distinguished between repetitive consent and lasting consent: repetitive consent involves obtaining consent for each new token activity (such as each visit to the toilet), whereas lasting consent involves consenting en masse to all activities of the same type (or all care activities of *all* types) from thereon after. Let us consider which of these types of consent is most applicable to routine care. It could be suggested that simply by living in a residential home, the patient tacitly agrees to all routine care activities, the same way citizens of a country might tacitly consent to being governed simply by living in the country (see Locke 2016, Russell 1986). In other words, one might think that patients give lasting consent to all care activities at the time they enter the institution, and therefore nurses do not need to obtain repetitive consent every time they carry out a new care activity.

One problem with this is that many elderly people are resistant to being placed in residential homes, so do not seem to consent to care even at the time of their admission. Moreover, forcing care on unwilling patients undermines the very reasons why obtaining consent is so normatively significant: autonomy, bodily integrity, and dignity, as well as patients' trust in nurses. This means that lasting consent in inappropriate for routine care contexts, and consent should be obtained for each new token care activity which is imminent. This

will help to avoid morally troubling situations where care is forced on non-consenting patients such as Winifred, described at the outset of the chapter.

Knowing *whether* a patient in a residential home requires help generally does not change (patients do not usually improve their ability to care for themselves) thus lasting consent may be apt to determine that a patent will, from now on, always require help with a particular task, such as feeding. However, determining *when* a patient requires help morally requires repetitive consent-seeking. Let me clarify with an example: suppose that due to arthritis, Winifred is unable to cut up her food and feed herself. It would seem futile – perhaps even hurtful – to ask Winifred "Do you need someone to feed you?" at every mealtime. Once nurses know Winifred cannot feed herself, asking *whether* she requires help becomes redundant. However, consent-seeking regarding *when* Winifred wants help (by asking something like "Are you ready for your lunch?") is morally necessary. This is because the default position is that other people may not touch our bodies (or our property): consent is morally required when someone wishes to deviate from this default position by interfering with us and touching our bodies (or our property) (Dougherty 2013: 723). Simply putting food into Winifred's mouth without first establishing that she wants to eat violates her autonomy and bodily integrity, reducing her dignity – something I argue for more fully below.

One instance of consent-seeking should probably be limited to the imminent token act (i.e. *this* meal; *this* instance of getting dressed); once obtained, one can legitimately proceed with the act unless consent is revoked. Consent-seeking for every mouthful of food, every button on a shirt, or every stroke of the hairbrush would seem excessive. (Similarly, one must morally obtain

consent for a token sex act – such as penetrative sex – but obtaining consent for every individual movement *within* a token sex act is not morally required). My pyramid in Chapter 6 (§1) listed various routine care activities: each activity listed on the pyramid is a type of routine care activity (such as dressing, bathing, and feeding). Thus, dressing is a type of routine care activity, but dressing on Monday morning is a different token activity to dressing on Tuesday morning.

Some routine care activities could be broken down further into sub-tasks (bathing involves undressing, getting into the bath, washing the body, washing the hair, getting out of the bath, drying, and re-dressing). Although consent for each of these sub-tasks might not be morally *required*, it may still be morally good for nurses and carebots to seek consent for each sub-task within the token act. This would seem less excessive than consent-seeking for every mouthful of a meal or every button of a shirt, given that one eats mouthfuls and fastens buttons just a few seconds apart.

The frequency of routine care activities (and therefore routine consent-seeking) means that philosophical explorations of sexual and medical consent cannot be neatly applied to routine consent. The frequency of routine care also reinforces the moral importance of consent-seeking because – as I later suggest – non-consensual routine care can reduce a patient's dignity by a small (but cumulative) amount.

*

Recall that the aims of this chapter are twofold: to demonstrate why existing normative accounts of sexual and medical consent cannot simply be applied

to routine care cases, and to establish the normative importance of consent-seeking in routine care situations. So far, I have done the former, by showing some key differences between routine consent, and sexual and medical consent. I have also outlined how consent should be obtained, and how frequently.

A critic might suggest that routine consent is unworthy of philosophical study because it is intuitive, and the stakes are low. I shortly address both these possible claims. First, in §2, I argue that routine consent is not necessarily intuitive for nurses – and it is certainly not intuitive for carebots; after that, in §3, I argue that non-consensual routine care is morally significant because it can substantially reduce a person's variable dignity.

2 Is routine consent intuitive?

Despite a plentiful philosophical literature on sexual and medical consent, routine consent remains largely absent. Critics might suggest that this is because it is unworthy of scholarship because routine consent-seeking is intuitive and obvious; this section addresses such a potential argument, and demonstrates the normative importance of routine consent.

Neurotypical adults are familiar with which activities are a normal part of our everyday social interactions, and which are not. We generally comprehend appropriate ways to behave, both in common and uncommon situations. I know that if my friend has sprained her wrist and is unable to undress herself, I should not begin undressing her without first obtaining her consent. Similarly, I understand that if my grandfather is losing his balance at the top of the stairs, it is permissible (and perhaps even morally *required*) for me to reach out and

grab him without first asking his consent. Being adults, we can, for the most part, navigate our way through these everyday routine care situations without having to ponder our actions too deeply. Such interactions feel so natural and intuitive that philosophical scrutiny might be deemed unnecessary, suggesting a possible reason why routine care is absent from the literature.

Routine consent may seem fairly intuitive: nurses and informal carers have been adequately navigating it for millennia. However, philosophical clarity is now required, not only because it is an interesting yet underexplored phenomenon, but also because carebots will soon be placed in residential homes, and they cannot rely on their intuition as a human nurse can. A carebot which does not appropriately obtain consent will be rude at best; at worst, it could have serious detrimental effects on a patient's dignity (argued for in §3). Furthermore, residential homes or roboticists could face legal action if carebots fail to obtain patients' consent, making routine consent an important pragmatic consideration worthy of scholarship. Detailed normative study of routine consent is not *solely* important because carebots lack intuitions – though that is a compelling reason why it should be explored.

Many nurses navigate routine consent adequately, so one might think a carebot could simply learn from and copy its human counterparts and behave as human nurses do (vis-à-vis consent-seeking). This might be satisfactory if we could guarantee that all nurses behave appropriately regarding consent-seeking; unfortunately, this is not the case. News media and television documentaries have shown some shocking treatment of elderly people by nurses in residential homes. Some cases involved rough, careless treatment and physical violence (Phillips 2012, Panorama 2014) and some even

involved sexual violence (Ferguson and Gallagher 2020, McGuinness 2022). Clearly, a carebot should not copy violent and sexual behaviour, but what about copying nurses' consent-seeking behaviour for routine care?

Unfortunately, nurses do not always obtain consent prior to providing routine care. Patients in residential homes are sometimes moved around, undressed, and washed without nurses even speaking to them (Panorama 2014, Phillips 2012, BBC Two 2021). Clearly, routine consent-seeking is not as intuitive for human nurses as we might like to think.

Perhaps, then, carebots should learn how to behave from a broader group of people? Unfortunately, carebots learning their behaviour from the general public should also be avoided: within 24 hours of going live on Twitter, Microsoft's Tay chatbot was sharing racist, vulgar, and bigoted tweets with the world, because of what it had learned from people on Twitter. To avoid morally troubling situations such as this, it is essential for carebots to be programmed with suitable (philosophically robust) consent-seeking behaviours prior to being used in care settings: this is one reason why this thesis is necessary.

I trust it is clear that routine consent-seeking is not necessarily intuitive for human nurses, and even if it were, the introduction of carebots means that ethically defensible routine consent-seeking practices need to be established. So any suggestion that routine consent is unworthy of scholarship because it is intuitive is implausible.

I have at many times in this thesis suggested that legitimate consent must be obtained prior to providing routine care if that care is to be permissible; below

I demonstrate why this is necessary, by warning of the dire consequences if non-consensual routine care is provided to patients.

3 Promoting dignity via consent-seeking

In §1 I showed why existing scholarship on sexual and medical consent cannot be easily transposed to provide an adequate understanding of the conditions for routine consent-seeking, and in §2 I criticised the claim that routine consent need not be studied because it is intuitive. In this section, I address the potential claim that routine consent is unimportant because the stakes are low. My argument is that repeated non-consensual care could have cumulative and severe detrimental effects on patients' dignity.[150]

A single non-consensual sexual act (rape; sexual assault) can cause victims to suffer in various ways, including post-traumatic stress disorder (PTSD), flashbacks, anxiety, insomnia, depression, self-harm, or suicidal ideation (Joyful Heart Foundation 2022, Rape Crisis 2023, Victim Support UK 2023); some of these might also be suffered following non-consensual medical procedures. There may be physical injuries too. One might suggest that the same is not true of many routine care activities: people do not develop PTSD or feel suicidal because they were once spoon-fed unnecessarily, or because someone put a cardigan on them without obtaining consent. Thus, the apparent low stakes involved in routine care could mean it is not immediately obvious why this is a moral issue at all.

[150] Recall that references to dignity in the chapter are to variable dignity, unless otherwise stated.

On a related note, medical and sexual activities legally require consent (National Archives 2003, General Medical Council 2013, Nursing and Midwifery Council 2018), and perpetrators can receive up to life imprisonment for performing a non-consensual sexual act or invasive medical procedure. By contrast, although nursing codes of conduct stipulate that nurses should obtain consent before providing routine care (Royal College of Nursing 2017), failing to do so is not a criminal offence: a case of a nurse carefully brushing a patient's hair without her consent would be laughed out of court. Since many laws exist to help prevent harm, a lack of legal repercussions for non-consensual routine care could be seen as a marker of ethical unimportance by philosophers.[151] This could at least partially explain why routine consent is absent from the consent literature.

However, I maintain that the stakes for non-consensual routine care are in fact potentially high – particularly the effect on a patient's dignity, meaning that routine consent is an issue of high normative significance. It is true to say that *some* non-consensual sexual acts (e.g. rape) and *some* non-consensual medical acts (e.g. chemotherapy) are potentially traumatising and can reduce one's dignity, and this is not true of *some* non-consensual routine care activities (e.g. cutting up a patient's food). However, sexual, medical, and routine care activities all exist on a spectrum (or pyramid) of invasiveness, and comparing opposite ends of the spectra is not a like-for-like comparison. Some routine care activities (e.g. bathing), if performed non-consensually, could have a far greater detrimental effect on someone's dignity than some non-

[151] Of course, it is an offence to assault or abuse patients, but such behaviours are not captured by what I call non-consensual routine *care*.

consensual sexual acts (e.g. suggestive elbow-touching) or non-consensual medical acts (e.g. examining a mole on the hand).

An opponent may object that I have not established any link between non-consensual bathing and the reduction of dignity. Although this is true at present, in what follows, I argue for the more controversial claim that even routine care activities from the lower levels of the pyramid (such as reminders, cutting up food, feeding, grooming) can reduce patients' dignity if performed without consent. If my argument is convincing, then my claim here that the most invasive of routine care activities – Level 4 activities such as bathing – reduce one's dignity should be all the more plausible. I now argue that although non-consensual care activities from the lower levels of the pyramid only *slightly* reduce patients' dignity, the effects can be cumulative, meaning that repeated instances of non-consensual routine care can substantially reduce patients' dignity – making routine consent highly normatively significant.

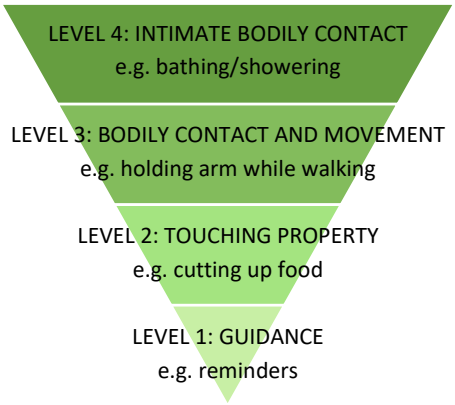
Guidance from Royal College of Nursing (Baillie, Gallagher, and Wainwright 2008: 45) suggests that obtaining patients' consent allows them to feel in control, which promotes their dignity.[152] If we accept this as plausible, then a nurse or carebot who obtains patients' consent contributes towards their empowerment and dignity more than one who provides non-consensual care (Rogers and Marsden 2013: 51–54).

[152] Interestingly, the RCN states that patients' *feeling* in control promotes their dignity – it does not state that their *being* in control promotes their dignity. This seems peculiar. However, perhaps a charitable reading of the document would interpret the RCN as meaning that *feeling and being* in control promotes patients' dignity.

As I suggested in Chapter 5, it seems immediately plausible that one's dignity can be reduced via severely degrading non-consensual treatment, such as being stripped naked, tortured, and photographed, as the prisoners in Abu Ghraib were (Center for Public Integrity 2008, Hooks and Mosher 2005, Hall 2019). Let us call this sort of abhorrent, degrading, inhumane treatment a 'macro-violation' of dignity.[153] Macro-violations of dignity, so defined, reduce a person's dignity suddenly and dramatically – but this is not the only way in which a person's dignity could be reduced. In the same way that one can be injured by a sudden accident, but also by repetitive strain, I suggest that one's dignity can be reduced not only by a highly intimate non-consensual act (a macro-violation of dignity), but also by repeated non-consensual care (micro-violations of dignity). This idea is aptly illustrated by the saying "Constant dropping wears away a stone" (Oxford Reference 2023). The idea is that even if an activity has minimal consequences, if the activity is repeated and repeated, the consequences become noticeable, and can eventually be colossal.

[153] Other extreme acts such as rape, or being non-consensually medically experimented upon would also be macro-violations of dignity. Although some people may cope with such atrocities better than others, I believe it is uncontroversial to claim that they are gross violations of the victim, and can lower the victim's variable dignity.

A 'micro-violation' of dignity, as I define it, involves performing one of the lower-level routine care activities (Levels 1-3) without consent.[154] First, I argue why Level 2 and Level 3 violations[155] reduce patient dignity, and later I argue why Level 1 violations – which do not even involve touching the patient – also reduce dignity. As one might expect, I suggest that the higher-level violations have a greater detrimental effect on patient dignity than do Level 1 violations, just as rape often has a greater effect on victims' dignity than non-consensual hugging does.[156]



154 There are some medical activities (e.g. arm-examination) and sexual activities (e.g. hand-holding) which might also be called micro-violations of dignity if performed without consent, but I do not establish that herein.

155 For brevity, I use "Level X violations" to mean "non-consensual Level X routine care activities".

156 Although I have separated routine care activities into four broad categories in the pyramid, there is variation within the categories: e.g. I placed *pushing a patient in a wheelchair* in the same category as *dressing the patient in outer clothes, such as trousers*. The latter seems more intimate than the former; although I use the terms 'micro' and 'macro' violations, there will be variation within these categories, and no clear dividing line between 'macro' and 'micro' violations. I could, of course, introduce a third term to cover the fuzzy middle area, such as 'mezzo-violations', however there would still remain variation within the three categories, and no clear dividing line between them. I acknowledge this issue, but do not feel it is a significant problem for my argument, since I suggest that consent should be obtained for all routine care activities. Precise measurements of dignity-reduction are not a requirement: I can still maintain that the intimacy / invasiveness of non-consensual routine care activities correlates roughly with the detrimental effect on dignity (and thus the moral importance of consent-seeking).

My suggestion that some non-consensual routine care activities are more worrisome than others fits with our understanding of medical consent too: it is far more troubling for a doctor to perform non-consensual testicular examinations on patients than to non-consensually examine a mole on a patient's cheek from behind a desk (Eyal 2019: §2.2). In the previous chapter I mentioned the scandal of non-consensual pelvic examinations of unconscious patients (Picard 2010, Thomson-DeVeaux 2010). If it had instead been discovered that student doctors were examining patients' hands while they were unconscious, I do not believe this would have been a scandal at all. This is because we understand that genitals are more private than other body parts such as hands and faces, which are generally on show to others all the time; thus the non-consensual examination of our private body parts is more worrisome than non-consensually examining oft-shown body parts (the fact that the examinations were tactile – medical students inserted their fingers into patients' vaginas rather than simply looking at them – exacerbated the violation). Viewing and touching someone's private body parts is by default not sanctioned, meaning that viewing or touching them morally requires consent.

But why should we think that *helping* patients without consent affects their dignity at all? To answer this question, let us consider this example: Edward is struggling to button up his shirt; the nurse notices, and buttons it up for him. Providing this routine care without first seeking Edward's consent could, in a small way, make Edward feel objectified and infantilised. Although he has been helped (his shirt is now buttoned up), his consent was not sought and he may thus feel he has not been treated as a person with thoughts and

feelings, but rather, as an object. After all, when an object requires something (a towel needs folding; a switch needs flicking; a sheet needs washing) we just act on it without consent-seeking, and this is what the nurse did to Edward.

Nussbaum lays out some features of objectification,[157] including treating someone as if they lack agency or boundaries, and behaving as if their feelings are unimportant (Nussbaum 1995: 257). Langton describes how treating someone as if they cannot speak is an additional feature of objectification (Langton 2009: 229). These features are glaringly present in cases such as Edward's: his agency, boundaries, feelings, and ability to speak have been ignored. Intervening to button Edward's shirt without consent-seeking is how a nurse behaves when he spots a towel needs folding (or similar): he simply gets the task done without pausing to consider feelings or agency. This is justifiable on a towel, but unjustifiable on a (conscious, neuro-competent) patient; it is a morally problematic example of infantilisation (Kitwood 1997: 47). It seems plausible to say that Level 3 violations (non-consensually touching patients) are objectifying and infantilising, and could reduce patients' dignity.

As noted above, some investigative documentaries have shown poor treatment of patients in residential homes: families complained that their loved ones were being moved around "like a slab of meat" (Panorama 2014, Phillips 2012, BBC Two 2021). The complaint here was not specifically the way in which patients were *physically* treated, but that patients were being touched

[157] Although 'objectification' is often used as a shorthand for *sexual* objectification, I use it in its broader (Nussbaumian) sense, to mean treating someone as if they were an object.

and moved without the nurses seeking prior consent or even talking to them; nurses simply took hold of patients and put them onto the bed, or pulled them out of the chair. Families commented that the treatment was 'undignified'.[158] Earlier, in §2, I discussed whether routine consent-seeking is intuitive: incidents such as these demonstrate that it is not nearly as intuitive for human nurses as one might think (though these are hopefully the exception rather than the rule). This also highlights the concern about carebots copying other nurses' behaviour: we need carebots which seek consent at appropriate times (viz. every new token care activity) to promote patients' dignity, because Level 3 violations are undignified and morally wrong. However laudable one's motives, providing routine care non-consensually is disrespectful, and can reduce patients' dignity. It is more laudable to actually *check* whether someone wants help; this helps empower patients and promotes their dignity (Rogers and Marsden 2013: 51–54).

There may be times when nurses or carebots are presented with a dignity dilemma, and must consider whether to act without patients' consent (violating their dignity) in order to promote their dignity in some other way. For example, suppose that Lloyd is struggling to cut up his steak (he is having to pick up the steak with his hands and bite it); a nurse offers to help him, but Lloyd refuses. It seems undignified to leave Lloyd eating steak with his hands because he cannot cut it up, but it also seems undignified to cut up the steak against his wishes (a Level 2 violation). If Lloyd is an adult with capacity, voluntarily making an informed decision about (not) cutting up his steak, this should be

[158] These documentaries sadly contained plentiful examples of nurses behaving this way towards both neurotypical patients, and those with dementia.

accepted, and the nurse should leave him be. Respecting Lloyd's wishes helps to promote his dignity, even though his choice to eat his steak by hand does not seem like a prudent, pleasant, or dignified one.[159] My view on routine consent echoes Eyal's view on medical consent:

> "When a sufficiently capacitated adult does not give sufficiently informed and voluntary consent to intervention in her body or her private sphere, then [pro tanto] the intervention is impermissible – even when it seeks to assist her, physicians recommend it, third parties would benefit from it, and the patient herself had repeatedly consented to it before expressing a change of mind." (Eyal 2019: § 1)

I have so far discussed how some non-consensual routine care activities are micro-violations of dignity. Level 2 violations (cutting up Lloyd's food) and Level 3 violations (putting patients on beds; buttoning up their shirts), even with beneficent intentions, can infantilise and objectify patients, reducing their dignity in some small way. If those arguments were successful, it should be clear that Level 4 violations (bathing; undressing) could reduce patients' dignity more substantially, but it may still be unclear how Level 1 violations – which do not even involve touching the patient or their property – could reduce dignity levels. To demonstrate this, I draw on some literature on microaggressions which demonstrates the power of mere verbal carelessness, *and* that the effects on dignity can be cumulative; my argument regarding

[159] This is a rule of thumb; there may be some exceptions (such as emergencies) where dignity can be permissibly violated. Spelling out all such possible situations is beyond the scope of this thesis, but a substantial number of scenarios could presumably be programmed into carebots to give them ethically defensible decisionmaking procedures.

Level 1 violations helps to further emphasise the fact that higher level violations can also reduce dignity.

3.1 Cumulative reductions in dignity and microaggressions

Microaggressions are subtle acts of discrimination, implicit bias, or prejudice (Sue et al. 2007). They are often automatic and unintentional acts which serve as 'put-downs' to the victim, via hostile, exclusionary, or demeaning behaviour (Pierce et al. 1978: 66, Williams 2020, Skinta and Torres-Harding 2022). They are generally directed towards minorities and historically oppressed groups such as women, elderly people, disabled people, LGBT+ people, black people, Muslims, and Jews. However, unlike overt acts of prejudice such as verbal abuse, violence, or blatant discrimination, microaggressions are subtle, ambiguous, and minor. Microaggressions can be verbal (like asking a British Asian person "Where are you *really* from?"), non-verbal (like paying closer attention to Muslims wearing backpacks on public transport), or environmental (like the use of male and female symbols on toilet doors) (Sue 2010a: 25). Although none of these are clear examples of abhorrent behaviour, they do belie potentially prejudiced mindsets towards Asian people, Muslims, and non-binary people respectively.

Like microaggressions, non-consensual routine care provided by nurses – dressing, feeding, bathing and the like – may often be done without malicious intentions, or without much thought of any kind, and may be based on implicit assumptions about the sorts of care elderly people require. Microaggressions stem from implicit assumptions, stereotypes, or experiences of Muslims, black people (etc); similarly, nurses' treatment of patients may be based on assumptions, stereotypes, or past experience with elderly patients. For

example, many patients may require reminding to take their medications (a Level 1 routine care activity), so nurses may remind *all* patients who take medication, just to be safe, when perhaps some of those reminders are not necessary.

Some microaggressions may be masked as compliments. For example, telling a black person "You are a credit to your race" may stem from good intentions (to applaud achievement and good character), however the message conveyed is that it is unusual to find a decent or successful black person (Sue et al. 2008: 331). Similarly, reminding patients to take medication is done with good intentions (to ensure patients' health), however the subtext is "Old people are usually forgetful" – which may feel like a put-down *masquerading* as help.

When nurses provide non-consensual routine care (from any Level of the pyramid), they may be engaging in "epistemic microaggressions" (Freeman and Stewart 2021: 1016, 2019) because they do not consider or trust the fact that elderly people with capacity know whether or not they want help. Furthermore, it could be suggested that they are performing "emotional microaggressions" because such nurses ignore or fail to take seriously the emotional experiences of patients (Freeman and Stewart 2021: 1018): they do not consider how it could feel to receive non-consensual care. Offering help to people with physical disabilities often stems from "a genuine intent to be helpful on the part of the perpetrator, [however,] the aggregate impact of continuous unsolicited, unwanted, and unneeded offers of help [is] overwhelmingly negative, intense, and long lasting" (Keller and Galgay 2010: 253). If the effects of mere *offers* of help can have such detrimental

effects on people with disabilities,[160] then *providing* unwanted help could probably cause the same or greater negative effects on the patient. Those who provide unwanted help often (erroneously) feel they have done something good (being compassionate and providing assistance to someone in need), yet the person receiving the unwanted help is very often left feeling humiliated, invalidated, infantilised, and that their dignity has been diminished (Keller and Galgay 2010: 255 & 264, Conover, Acosta, and Bokoch 2021, Conover, Israel, and Nylund-Gibson 2017). It thus seems clear that Level 1 violations (such as nurses reminding patients to take their medication or use the bathroom, or telling them how to use their smartphone) can indeed reduce patients' dignity without any physical contact between nurse and patient.

One might suggest that the remedy to patients feeling a loss of dignity by nurses' microaggressions (in the form of non-consensual care) is for patients to speak out and tell nurses to stop doing it. However, the subtlety and ambiguity of microaggressions means that it can be extraordinarily challenging to convince someone that a particular act is indeed a microaggression (Wang, Leu, and Shoda 2011). Microaggressions are often normalised and accepted – particularly by privileged groups, who express them unintentionally and may not even notice them (they may believe telling someone they are a credit to their race is a simple compliment). Furthermore, people who claim they are victims of microaggressions are sometimes not taken seriously when they call out a microaggression: they can be accused of gaslighting, paranoia, seeing oppression everywhere, or making a big deal out

[160] The people in Keller and Galgay's study were under 65 with physical disabilities, but they are little different from the patients on which this thesis focuses: they are neurotypical but with physical difficulties – only their age is different.

of nothing (Sue 2010a, 2010b, Lilienfeld 2017, McArdle 2015). Calling out microaggressions can therefore be fraught with difficulty: it is sometimes met with defensiveness and denial by the alleged perpetrators (Keller and Galgay 2010: 244), and eye-rolling by others who feel people 'play the victim card' at even the slightest infraction against them (Campbell and Manning 2014:714-716).

Microaggressions are generally performed by more privileged groups against less privileged groups, and reinforce existing power structures (Sue 2010a, 2010b). Within residential homes, nurses have power, while patients lack power: patients who object to non-consensual care might be called difficult, oversensitive, stubborn, or that they are 'playing the victim' when nurses are only trying to help them (Campbell and Manning 2014:714-716). This could create an even greater reduction in dignity: not only are patients experiencing the subtle slight of receiving unwanted help, but are disbelieved or belittled when they speak out about it. One can understand how it could be difficult for patients to complain about non-consensual care, given that the 'perpetrators' are the very people who control patients' quality of life. Patients are in a hierarchical dynamic with the nurses who look after them, and patients depend on the care and company of nurses; expecting patients to call out each micro-violation of dignity (viz. each act of non-consensual care) places additional emotional burdens on patients, and risks straining the nurse-patient relationship.

Nurses who are criticised for providing non-consensual routine care may feel they are being (unjustly) scolded for doing their job and for exhibiting morally good behaviour (Friedlaender 2018: 13) – this may be reinforced by patients'

families, who want their loved ones to simply accept the help that is given. Critics might therefore argue that it is unproductive and wrong to criticise nurses who provide routine care simply for not seeking consent. However, this sort of objection is unconvincing: if Kerry is a surgeon then it is her job to perform surgery, but she should not perform surgery *indiscriminately*, forcing it upon anyone whom she thinks requires it (except in emergencies). If Kerry performs non-consensual surgery she cannot justify her actions by pointing out that it is her job to perform surgery, and she had beneficent intentions. Similarly, it is not justifiable for a nurse to provide routine care indiscriminately without first establishing that the patient *wants* routine care: the nurse's good intentions and the nature of her job role are not normatively relevant.[161]

The reduction in a person's dignity or emotional wellbeing caused by a single microaggression is minimal. If people only experienced one or two microaggressions in their life, microaggressions would probably be unworthy of scholarship. Unfortunately, however, the minority groups who experience microaggressions often suffer them multiple times, from multiple people, and the detrimental effects can be cumulative (Friedlaender 2018). This accumulation could occur merely through the aggregation of each microaggression, or because new instances cause the individual to re-examine previous instances (thus intensifying the harm) (Ibid). The long-term negative effects of repeated microaggressions on someone can be extreme,

[161] Earlier, in §1, I pointed out some ways in which medical consent (and medical care) differs from routine consent (and routine care). Although it is true to say the two differ in terms of their frequency and information-giving practices, they are similar in that they are both motivated by beneficence and often performed by paid professionals – and these features are the salient ones in this analogy.

and include depression, anxiety, high blood pressure, insomnia, self-harm, suicidal ideation, and PTSD (Friedlaender 2018: 8, Sue 2010b: 3–25, 2010a, Leland 2008, Williams et al. 2009). This list of potential deleterious effects of microaggressions bears a striking similarity to the negative effects I mentioned earlier (at the start of §3) which can occur due to a single non-consensual activity (rape, medical experimentation, torture – macro-violations of dignity). If non-consensual routine care (micro-violations of dignity) have much in common with microaggressions, as I suggest they do, then it is plausible to suggest that repeated instances of non-consensual care can have similar effects to a single macro-violation of dignity. In other words, a patient's dignity can be reduced by repeated instances of non-consensual care being provided.

Earlier, in §1.4, I explained that routine consent-seeking is distinct from both medical and sexual consent-seeking because of the frequency with which routine care occurs: residential home patients might be helped by nurses numerous times per day, every day, for years, until they die. The likelihood is that non-consensual routine care – causing feelings of infantilisation and the reduction of dignity – is a daily occurrence for some elderly patients today. This is deeply troubling, not least because it may continue unabated until death, and speaking out about it may be met with derision and dismissiveness. When this is apprehended, the prognosis seems bleak, and it becomes easier to understand why non-consensual routine care is a potentially huge moral wrong which is not only worthy of philosophical study, but also requires some sort of ameliorative action by those who care for elderly people. One simple

but effective action to address this problem is for nurses and carebots to seek patients' consent before providing each new token routine care activity. [162]

Obtaining consent is not a panacea: residential care can be improved in a number of additional ways (such as decreasing nurses' workloads and increasing the amount of time they have to chat with patients, increasing funding for the care sector, or catering for the sexual needs of elderly patients (Lancaster 2022)); patients' dignity may be under threat from other phenomena, but discussing everything is beyond the scope of this thesis. Nevertheless, I have shown why obtaining consent for each new token routine care activity from all Levels of the pyramid can help to promote patients' dignity. If it is plausible that Level 1 violations (such as unnecessarily reminding patients to take their medication, or telling them how to use their phone when they have not asked for help and do not require help) can reduce patients' dignity in some small way, it should be all the more plausible that Level 2, 3, and 4 violations can reduce patients' dignity to a greater extent. Thus, it is a moral requirement for nurses and carebots to obtain consent prior to engaging in any new token routine care tasks; failing to do so risks repeated and cumulative reductions in patients' dignity.

4 Conclusion

This chapter had two aims: to demonstrate that routine care requires philosophical attention, and to elucidate why carebots and nurses should obtain consent prior to undertaking routine care activities. To accomplish the

[162] Recall that this need not involve ascertaining *whether* someone requires help (e.g. if we know the patient cannot *ever* dress himself) but it does require ascertaining *when* someone requires help (i.e. asking whether the patient would like to get dressed now).

first task, I examined some key differences between routine consent, medical consent, and sexual consent: these differences mean that existing work on sexual and medical consent cannot be neatly transposed to provide a normative justification of routine consent. But this alone does not explain why routine care is *worthy* of study. If routine care were intuitive or the stakes were low, then this would provide evidence that it may be unworthy of study. I considered but dismissed the claim that routine consent is intuitive; even if it were intuitive for all human nurses (which it is not), (present-day) carebots do not have intuitions, so they need it spelling out when routine consent is morally required.

To demonstrate that non-consensual routine care is not a low-stakes phenomenon, I argued that, like microaggressions, micro-violations of dignity – in the form of non-consensual routine care – can cause cumulative detrimental effects on patients' dignity. If I am right about this, then any claim about routine consent being unworthy of study because of the low stakes involved will fail. A single microaggression – like a single micro-violation of dignity – causes minimal harm and by itself is perhaps unworthy of study. However, both microaggressions and micro-violations of dignity are often part of a *pattern* of behaviour rather than a single instance, and if the harms are cumulative, this makes them morally troubling and worthy of philosophical attention. Moreover, the more intimate acts of routine care (such as bathing – a Level 4 activity) have the potential to cause an even greater reduction in dignity if performed non-consensually. This means that the stakes involved in routine consent are far higher than some people might imagine, and it is

essential for carebots and nurses to obtain consent for each token of routine care they provide.

As our elderly population increases and the shortage of nurses escalates, the need for carebots in residential homes will increase. With their introduction seeming ever more likely, it is imperative that the care they provide for elderly patients is appropriate, consensual, and dignified.

Conclusion

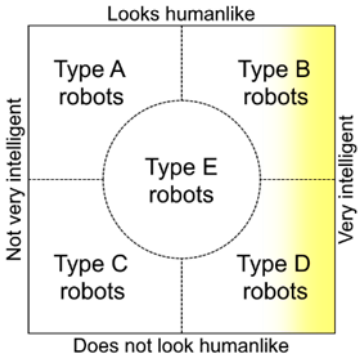
Philosophy can sometimes be rather abstract, having little impact on reality. It is therefore a delightful irony that what might sound like science fiction – robots caring for elderly people – is the very thing which makes this thesis so pertinent to reality. The notion of placing carebots in residential homes for the elderly is not merely an interesting thought experiment to prove a subtle point about ethics or reveal some hair-splitting distinction: it is very likely to be reality in the not-too-distant future, meaning that what I have explored herein is both significant and timely.

This conclusion chapter provides a summary of my primary arguments and conclusions. I then discuss what our future with carebots might be like, and consider whether I am over-predicting their use and abilities. I note some limitations and difficulties raised by this thesis, as well as indicating where further work would be useful.

1 Summary of my thesis

The two main themes of this thesis are whether robots are deceptive, and how carebots can promote patients' dignity via consent-seeking. My introduction chapter outlined how care staffing shortages and an increase in the elderly population mean that carebots will be increasingly utilised over the coming decades: this is why this thesis is necessary and could have real-world impact.

In Chapter 1, I noted that robots are not a homogeneous group, so roboethics is a diverse field of study; however, writers are not always clear about which types of robot they discuss. I provided a matrix which enables us to distinguish between robots based on how humanlike they look, and how intelligent they are, and I identified five distinct types of robot. I used the matrix to show which ethical issues are most pertinent to which types of robot – for example, discussions about robots' emotions are most relevant to advanced Type B and Type D robots (shown in yellow). Some concerns – such as loss of (human) jobs – are relevant to all types of robot.



Chapter 2 was a conceptual and normative analysis of deception. I considered whether deception must be intentional, and whether the deceived party necessarily ends up with a false belief. I noted that although people casually call optical illusions 'deceptive', this is not true *deception*. I outlined why deception is usually negatively normatively evaluated, but noted that prosocial deception is morally permissible or even good.

Chapter 3 was an exploration of different types of robo-deception. I argued that 'robo-deception alarmists' sometimes raise 'concerns' which are (a) not examples of deception, and/or (b) not morally troubling. I defined four types of robo-deception: anthropomorphic deception, zoomorphic deception, disanthropomorphic deception, and basic other-deception. I argued that although anthropomorphic deception and zoomorphic deception have received most attention, they often fail to meet the conditions for deception, and are not particularly morally problematic anyway.

Fake compassion was the focus of Chapter 4: robo-deception alarmists worry that it is deceptive if robots appear to care when in fact they do not. I maintained that most patients will not believe that the fake compassion from carebots demonstrates real robotic emotions or mental states, so the success condition of deception is unmet. Besides, if patients really did believe that carebots' compassion is real, this may help improve health outcomes, so is a morally unproblematic form of deception (or self-deception).

In Chapter 5 I showed that 'dignity' is often used in disparate ways, or different senses of the word are conflated. I distilled 'dignity' into universal dignity (a high normative status which is present in all and only humans), and variable dignity (a character trait which can increase or decrease depending on a person's self-image, behaviour, or how they are treated by others). I suggested that carebots (and human nurses) should provide care which promotes patients' variable dignity, because patients have universal dignity.

In Chapter 6 I provided an analysis of sexual and medical consent: I argued that legitimate consent must be voluntary, sufficiently informed, and given by a person with capacity who both articulates and (mentally) sanctions the act

which is taking place. Acts requiring consent are by default not sanctioned, such as touching people's bodies or property. I argued that consent is normatively significant because it promotes autonomy, bodily integrity, dignity, and trust between the parties involved.

Chapter 7 demonstrated that routine consent sufficiently differs from medical and/or sexual consent in terms of interpersonal transferability, information-giving, how consent is obtained, and frequency, thus existing normative accounts of consent cannot simply be transposed to give an adequate normative account of routine consent. I argued that consent is morally required for each new token routine care activity. I argued that non-consensual routine care can lower patients' variable dignity just as microaggressions can, and the consequent reductions in patient dignity can be frequent, cumulative, and extreme. This is why it is essential that carebots obtain consent before providing routine care to patients.

2 Our future with carebots

Fiction has long been projecting a future where we live alongside robots – either harmoniously or as adversaries. *Blade Runner* (Scott 1982) is set in 2019, and depicts robots – 'replicants' – which are almost indistinguishable from humans. In *2001: A Space Odyssey* (Kubrick 1968), self-aware AI is portrayed at the turn of the millennium. *The Terminator* shows AI becoming self-aware in 1997 – and killer cyborgs which are virtually indistinguishable from humans running around by 2029 (Cameron 1984). These events have not happened, and nor do they appear likely to happen imminently. Clearly, fiction has not done a great job of predicting our future with robots… but nor has non-fiction.

In 1966, TV documentary series *Tomorrow's World* depicted 'Able Mabel' – a robot housemaid which, it was suggested, people could be living alongside by 1976. Mabel would wake you, run your bath, cook your meals, do all your household chores, look after your kids, and protect your home (Tomorrow's World 1966). Even now, some fifty years later, it has still not happened. Given that people have so grossly over-predicted our future with robots for decades already, could this thesis be doing the same? For example, I have repeatedly suggested that robots will be increasingly present in residential homes within the next few decades (as well as on our roads, and in our workplaces, homes, schools, and shops). Could it be that I am over-predicting, and carebots are simply not going to exist?

I do not believe so, and here's why: there are *already* carebots in China and Japan – and probably other countries too (Siripala 2018). People are *already* embracing Alexa and Siri as ways to set reminders, contact loved ones, find out information, and entertain themselves – some of the tasks a carebot could perform. And there are *already* AI systems which can out-perform human doctors' diagnosis rates (McKinney et al. 2020, Babylon Health 2018). Given the shortage of care workers in many developed countries, and the existence (and relatively low cost) of carebots, coupled with increased automation in other areas of industry, it would not be at all surprising if carebots begin to be used in the near future in the UK.

In Chapter 4, I discussed whether a robot can care, and focused on some hypothetical futuristic carebots which I suggested could exist if a little technological convergence were to occur. Given that the carebots under discussion do not currently exist, it is *possible* that such robots will not come

to exist – though this seems highly unlikely. We have recently seen impressive advances in AI such as ChatGPT (Open AI 2023) which can produce a vast array of written pieces and computer code; it therefore does not seem too great a leap to think that we could soon be (verbally) chatting with robots which seem really quite personable. This is not a certainty, of course, but unless advances in AI worldwide suddenly cease, it does seem a likelihood that robots will become increasingly able to hold verbal conversations and to appear to emotionally care. Moreover, given that most of the tasks carried out by nurses can already be carried out by robots (see Chapter 4 §1.1), all that is required is to combine these separate technologies into a single entity, and the result would be a highly advanced carebot. So, it seems almost certain that carebots will exist at some point within the next few decades, given that we are already so close.

Perhaps, then, I am over-predicting just how accepting people will be of carebots? Have I envisioned an overly-rosy future, where being looked after by carebots is a fabulous and dignified solution to the problem of nursing shortages – when in fact people will find carebots infuriating? I concede it is possible that being cared for by robots may not be entirely trouble-free, and I presently share some thoughts on this.

I began working on this thesis in 2018: then, the idea of a worldwide pandemic and a government-imposed lockdown seemed like the stuff of dystopian science fiction – but in 2020-21, it happened. The covid-19 pandemic demonstrated many things about human nature, but the one I wish to draw out here is that we are social creatures. The stress, loneliness, and isolation caused by lockdowns was palpable and widespread. People from all walks of

life were affected, and the groups who suffered most from loneliness and depression were those who lived alone and were unemployed (Office for National Statistics 2020, Age UK 2020): in other words, those with least human contact. Social isolation is one of the main risk factors for depression and low mood among elderly people, and indeed depression is common among elderly people – even before the pandemic (National Institute for Health and Care Excellence 2013, Social Care Institute for Excellence 2006, British Geriatrics Society 2018). Human contact is clearly a crucial feature of people's lives, and critics might argue that carebots are a poor substitute for real human contact, and pandemic loneliness spotlighted that fact.

Certainly, I do not imagine that carebots, however personable, could take the place of anyone's relatives, but I do think that carebots could fulfil the role (at least partially) of a nurse. Nurses provide a service for patients with far less emotional attachment than one (usually) gets from family members. This means that replacing nurses with carebots is less of a leap than replacing family members with robots would be. Although the pandemic revealed the importance of contact with family and friends, I do not believe it demonstrated that it is humanity *in general* which is important; rather, that family and friends are important. One cannot simply replace family members with strangers, but the transactional nature of nurse-patient relationships means that nurses can be (and are) replaced by other nurses without too detrimental an effect on patients' welfare, and I suggest that at least some nurses could similarly be replaced by carebots (which are sufficiently advanced) without too detrimental an effect on patients' welfare.

Carebots do seem increasingly necessary: in the UK, there are already 100,000 unfilled nursing vacancies (Skills for Care 2021, NHS Support Federation 2022), and there are expected to be around 350,000 unfilled nursing vacancies in the USA by 2040 (Miller 2017). Turkle – who is often critical of robots in roles usually filled by humans – observes that people usually prefer to engage with humans rather than robots, but when the choice is between robots and nothing at all, people choose to engage with robots (and enjoy doing so) (Turkle 2017: 105). With such high numbers of unfilled nursing vacancies, it is not an exaggeration for me to suggest that patients' choice will increasingly be between carebots and no care at all, rather than carebots and human nurses. I do not claim that carebots are perfect, but merely that they could provide a good standard of (dignified) care for patients, even if some patients would prefer human nurses. It is worth noting that residential home patients would probably prefer to be cared for by family and friends rather than by nurses, but we recognise that this is often not practicable, and so nurses offer an ethically defensible and pragmatic – though not a perfect – alternative; the same can be said of carebots.

Whether our future really does involve closer interaction with carebots will be affected partly by technological innovations (the types of robot which exist), partly by people's attitudes towards robots, and partly by world events. I mentioned above that the covid-19 pandemic occurred during the writing of this thesis, and how our being social creatures led to widespread loneliness during lockdowns. We should also recognise, however, that it is humanity's social nature which made covid-19 into a pandemic rather than a localised virus outbreak. The virus spread to pandemic levels because people

interacted with one another, on a local and global scale. Elderly people – particularly those in institutional care – were worst affected by covid-19: around 40,000 residential home patients in the UK died from the virus between April 2020 and March 2021 (Holt and Burns 2021). One suspects that carebots could have made a tangible difference in reducing the transmission of the virus within hospitals and residential homes, where the virus spread like wildfire. Carebots would have been able to help patients without concerns about transmitting the virus through breathing and talking, as a human can. Although the covid-19 virus can be transmitted through contaminated surfaces, this is not its primary modus operandi; besides, it is easier to sterilise a robot – which can be sprayed with bleach and scrubbed vigorously – than it is to sterilise a human nurse. Carebots also limit the necessity of personal protective equipment (PPE) such as masks and rubber gloves.

With experts suggesting that another (possibly deadlier) pandemic is very likely within the next ten to twenty years (Andrews 2023, Miller 2021, Gulland 2021), it might not be long before we find out the true value of carebots. This is just another reason why carebots present a useful and ethically defensible resource for the future of healthcare – and another reason why I do not believe that this thesis is over-predicting our future with carebots.

3 Questions for further study

In this section, I consider some of the questions raised by this thesis, and discuss some potential areas for future philosophical exploration and/or empirical research.

3.1 Deception and dignity

This thesis has focused on dignity and deception, and one might wonder about how these two concepts interact with one another, such as whether nurses, carebots, or roboticists should engage in deception if it promotes patients' dignity. Although insidious lies and deception break trust and are morally problematic (Williams 2002: Ch. 5), I argued in Chapters 2 and 4 that prosocial deception is morally permissible or even ideal. Prosocial deception can actually bolster trust (Levine and Schweitzer 2015), as it typically involves making circumstances more pleasurable or palatable for others, thus improving social relations. Whether deception can promote dignity is not a simplistic question to answer: further concept analysis would be required to properly determine whether this is possible, and if so, the circumstances under which it is and is not permissible. It might be that any deception which promotes others' dignity is by definition prosocial and morally good, or it might be that a lie can be antisocial but (all things considered) morally good if it promotes dignity. If deception can promote dignity, then is self-deception which promotes one's own dignity equally morally good? These are interesting questions which could be further explored in future.

3.2 Do carebots consent to being carebots?

I have written extensively herein about patients' autonomy and obtaining patients' consent, but I have not written anything about *carebots'* autonomy or consent. An intriguing question which one might consider after reading this thesis is whether we should ask carebots if they consent to being carebots. We do not force humans to become nurses, so why force robots to do so?

Some literature already exists regarding sexbots and consent – whether sex with sexbots amounts to rape, whether sexbots should be given the ability to withhold consent, and whether we should curtail humans' ability to force robots into sexual activity (Gunkel 2018, Eskens 2017, Petersen 2017, Danaher 2017c, Cappuccio, Peeters, and McDonald 2019, Frank and Nyholm 2017a). There are also more general concerns about whether it is acceptable to make robots our slaves (Chomanski 2019, 2020, Petersen 2011). What does not exist at present, but would be interesting to explore, are issues specifically surrounding carebots and consent: should carebots be asked whether they consent to being carers? If they do not consent, what would we do to fill the staffing shortfall in hospitals and residential homes? It would be interesting to consider whether designing robots with a function (nurse; courier; soldier) *itself* amounts to slavery – after all, humans have the right to choose their line of employment.

One possibility would be to limit the intelligence levels of robots so that they never approach sentience or awareness that they are effectively slaves, working without rest and without pay, in conditions which humans would not enjoy, nor be able to tolerate (e.g. 22-hours a day, every day). At the moment, robots are not sentient, so this seems like a non-issue, but as technology progresses, it may become possible to create robots which are sentient or more aware of their status. At that point, limiting their intelligence or awareness of their status sounds worryingly like a digital lobotomy or Marxist false consciousness: constraining workers and keeping them in a state of perpetual ignorance and non-personhood so that they are better slaves. It does not sound beneficial for the carebots, but it would certainly be beneficial

to the elderly patients requiring care, and this presents us with a dilemma: are the rights of elderly patients more important than the rights of robots to refuse to be nurses? This sort of dilemma would need to be philosophically explored as AI advances, because if carebots are given the opportunity to refuse to work (as nurses or at all), humanity could face its second care staffing shortfall, without any feasible solution.

3.3 Can carebots assess capacity?

I argued in Chapters 6 and 7 that for consent to be legitimate (and morally transformative) certain background conditions must be met – these included the consenting individual being sufficiently informed, capacitated, acting voluntarily, and mentally sanctioning the activity (in addition to behaviourally indicating they sanction the activity). We might wonder, then, whether robots who seek consent – such as the carebots on which this thesis focuses – should be the ones assessing whether patients are sufficiently informed and capacitated.

Some robots may be able to aptly ascertain whether patients are sufficiently informed. For example, they could ask patients if they know what a particular activity involves, and what its risks and benefits are (though it would become tiresome for patients to have to explain what getting dressed involves every time they consent to the carebot dressing them). It would be significantly more difficult for a robot to assess patients' capacity. Understanding others' capacity may require a theory of mind – or could capacity be assessed adequately through a series of multiple-choice questions? This seems very risky. Even if carebots did have a theory of mind, there might be compelling ethical reasons not to permit robots to assess the capacity of patients – at least until we are

certain that robots are sufficiently skilled at doing so. Further research – empirical and philosophical – would be required to identify how we should program robots' ability to assess patients' capacity, and whether there are compelling ethical reasons for the assessment of capacity to be carried out by humans rather than robots. Given that a lack of capacity in a patient might mitigate or void any consent-refusal, this is an important area of study which it would be useful for future research to examine.

4 Final thoughts

I believe the chance is high, for those of us who are long-lived, that robots will care for us in some way or other, meaning that the importance and impact of this thesis may increase as that becomes more and more likely. There will be many situations over the coming years where carebots need to make ethical decisions, and these will have direct and potentially long-lasting effects on the lives of the patients for whom they care. This thesis is not merely an interesting thought experiment; rather, it is a real-world necessity that carebots take appropriate actions which promote the dignity and best interests of patients: our grandparents, parents, and in time, us and our descendants. How we program and develop carebots today will determine the sort of future we have; it would be morally impermissible to let robots run amok and only rein them in after catastrophe. Rather, we must ensure that the behaviours and decisionmaking abilities we give to them are philosophically robust and ethically defensible: I hope this thesis goes some way towards ensuring this.

References

Accenture (2019) *Artificial Intelligence in Healthcare* [online] available from <https://www.accenture.com/us-en/insight-artificial-intelligence-healthcare> [18 June 2019]

Adler, J. (1997) 'Lying, Deceiving, or Falsely Implicating'. *Journal of Philosophy* 94, 435–452

Aerospace Robotics (2014) *Think-Sense-Act: What Does It Mean to Be a Robot?* [online] available from <https://aerospacerobotics.com/blogs/learn/15008585-think-sense-act-what-does-it-mean-to-be-a-robot> [3 November 2022]

Age UK (2020) *Age UK Research on Impact of the Pandemic on Our Older Population's Health* [online] available from <https://www.ageuk.org.uk/latest-press/articles/2020/10/age-uk--research-into-the-effects-of-the-pandemic-on-the-older-populations-health/> [10 February 2023]

Ageless Innovation (2018) *Companion Pets* [online] available from <https://joyforall.com/> [18 January 2022]

Aging Care (2020) *What to Do about a Male Client That Always Has an Erection?* [online] available from <https://www.agingcare.com/questions/what-to-do-about-a-male-client-that-always-has-an-erection-463309.htm> [22 March 2023]

Alexander, L. (1996) 'The Moral Magic of Consent (II)'. *Legal Theory* 2 (3), 165–174

Aljouni, K.M. (1995) 'History of Informed Medical Consent'. *Lancet* 346, 980

Alzheimer's Research UK (2022) *Prevalence by Age in the UK* [online] available from

<https://www.dementiastatistics.org/statistics/prevalence-by-age-in-the-uk/> [25 January 2022]

Amazon (2023) *Echo Dot (4th Generation) | New Smart Speaker with Alexa* [online] available from <https://www.amazon.co.uk/dp/B084DWCZXZ/?tag=mh0a9-21&ref=pd_sl_4e86hvewkb_e&adgrpid=1185274445255916&hvadid=74079871517252&hvnetw=o&hvqmt=e&hvbmt=be&hvdev=c&hvlocint=&hvlocphy=41603&hvtargid=kwd-74079803079578:loc-188&hydadcr=15240_2306075> [26 January 2023]

Amazon (2022) *Joy for All A7594 Companion Pet, Silver Cat with White Mitts* [online] available from <https://www.amazon.co.uk/Joy-All-Robotic-Companion-Pet/dp/B07VVBZ849/ref=asc_df_B07VVBZ849?tag=bingshoppinga-21&linkCode=df0&hvadid=80401880593187&hvnetw=o&hvqmt=e&hvbmt=be&hvdev=c&hvlocint=&hvlocphy=&hvtargid=pla-4584001424641112&psc=1> [9 December 2022]

ANA (2015) *Code of Ethics for Nurses (USA)*. American Nurses Association (ANA). available from <https://www.nursingworld.org/practice-policy/nursing-excellence/ethics/code-of-ethics-for-nurses/coe-view-only/> [26 November 2019]

Anderson, S.A. (2005) 'Sex Under Pressure: Jerks, Boorish Behavior, and Gender Hierarchy'. *Res Publica* 11 (4), 349–369

Andrews, L. (2023) *World Health Organization Warns 'we Must Prepare' for Potential Bird Flu Pandemic | Daily Mail Online* [online] available from <https://www.dailymail.co.uk/health/article-11727769/Risk-humans-H5N1-bird-flu-remains-low-prepare-WHO.html> [10 February 2023]

Apple (2023) *Siri* [online] available from <https://www.apple.com/uk/siri/> [30 March 2023]

Aristotle (1941) 'Nicomachean Ethics'. in *The Basic Works of Aristotle*. ed. by McKeon, R. Random House, 927–1112

Arkin, R.C. (1998) *Behavior-Based Robotics*. MIT Press

Arnold, L., Donovan, S., Moo-Young, C., Nettheim, D., Brozel, M., Gregorini, F., Tibbetts, C., Mackay, A., Robertson, J., Senior, R., and Williams, B.A. (2015) *Humans* [online] available from <https://www.channel4.com/programmes/humans>

Ashbrook, T. (2016) *Digital Disruption in White Collar Jobs* [online] available from <https://www.wbur.org/onpoint/2016/04/18/digital-white-collar-jobs> [1 November 2022]

Austin, J.L. (1962) *How To Do Things With Words*. 2nd edn. ed. by Urmson, J.O. and Sbisá, M. Harvard University Press

Autoweek (2018) *Tesla Continues to Blame Driver in Fatal Crash Involving Autopilot* [online] available from <https://autoweek.com/article/autonomous-cars/tesla-blames-driver-fatal-crash-involving-autopilot> [25 October 2018]

Avatarmind (2017) *For Senior Care, Retail/Hospitality, and Children's Education* [online] available from <https://www.ipalrobot.com> [18 January 2022]

Babylon Health (2018) *Online Doctor Consultations & Advice* [online] available from <https://www.babylonhealth.com> [17 December 2018]

Bach, K. (1981) 'An Analysis of Self-Deception'. *Philosophy and Phenomenological Research* 41, 351–370

Badham, J. (1986) *Short Circuit* [online] available from <https://www.imdb.com/title/tt0091949/>

Baillie, L., Gallagher, A., and Wainwright, P. (2008) *Defending Dignity - Challenges and Opportunities for Nursing*. ed. by Royal College of Nursing. Royal College of Nursing. available from <https://www.dignityincare.org.uk/_assets/RCN_Digntiy_at_the_heart_of_everything_we_do.pdf> [1 March 2019]

Baker, R.B. and McCullough, L.B. (eds.) (2009) *The Cambridge World History of Medical Ethics*. vol. 1–2. Cambridge University Press

Barnes, A. (1997) *Seeing through Self-Deception*. Cambridge University Press

Bates, V. (2015) *History & Policy* [online] available from <https://www.historyandpolicy.org/index.php/policy-papers/papers/the-legacy-of-1885-girls-and-the-age-of-sexual-consent> [9 September 2022]

Batty, D. (2004) 'Shipman Case Prompts Coroner Reforms'. *The Guardian* [online] 12 March. available from <https://www.theguardian.com/society/2004/mar/12/shipman.politics> [12 March 2019]

Bayer, T., Tadd, W., and Krajcik, S. (2005) 'Dignity: The Voice of Older People'. *Quality in Ageing* 6 (1), 22–27

BBC News (2022) *Facebook-Cambridge Analytica Scandal* [online] available from <https://www.bbc.co.uk/news/topics/c81zyn0888lt> [2 February 2023]

BBC Two (2021) *Inside the Care Crisis with Ed Balls* [online] available from <https://www.bbc.co.uk/programmes/m0011hfd> [8 February 2022]

Bekey, G.A. (2015) *Autonomous Robots: From Biological Inspiration to Implementation and Control*. MIT Press

Bellucci, S. (2019) ‘Still Think Robots Can’t Do Your Job? Essays on Automation and Technological Unemployment’. *Ethics and Social Welfare* 13 (1), 93–95

Berdahl, C.T., Henreid, A.J., Pevnick, J.M., Zheng, K., and Nuckols, T.K. (2022) ‘Digital Tools Designed to Obtain the History of Present Illness From Patients: Scoping Review’. *Journal of Medical Internet Research*

Beres, M.A. (2014) ‘Rethinking the Concept of Consent for Anti-Sexual Violence Activism and Education’. *Feminism and Psychology* 24 (3), 373–389

Bermúdez, J.L. (2000) ‘Self-Deception, Intentions, and Contradictory Beliefs’. *Analysis* 60 (4), 309–319

Bertolini, A. (2013) ‘Robots as Products: The Case for a Realistic Analysis of Robotic Applications and Liability Rules’. *Law Innovation and Technology* 5 (2), 214–247

Beyleveld, D. (2001) *Human Dignity in Bioethics and Biolaw*. Oxford University Press

Bickmore, T.W., Kimani, E., Trinh, H., Pusateri, A., Paasche-Orlow, M.K., and Magnani, J.W. (2018a) ‘Managing Chronic Conditions with a Smartphone-Based Conversational Virtual Agent’. in *Proceedings of the 18th International Conference on Intelligent Virtual Agents - IVA ’18* [online] held 2018 at Sydney, NSW, Australia. ACM Press, 119–124. available from <http://dl.acm.org/citation.cfm?doid=3267851.3267908> [25 June 2019]

Bickmore, T.W., Trinh, H., Olafsson, S., O’Leary, T.K., Asadi, R., Rickles, N.M., and Cruz, R. (2018b) ‘Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant’. *Journal of Medical Internet Research* 20 (9), e11510

Blackburn, S. (2009) *The Big Questions: Philosophy*. Quercus

Blue Frog (2022) *Welcome to Blue Frog Robotics* [online] available from <http://www.bluefrogrobotics.com/> [18 January 2022]

Bok, S. (2014) ‘Trust but Verify’. *Journal of Medical Ethics* 40 (7), 446–446

Bok, S. (1978) *Lying: Moral Choice in Public and Private Life*. Random House

Bołtuć, P. (2017) ‘Church-Turing Lovers’. in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Lin, P., Abney, K., and Jenkins, R. vol. 1. Oxford University Press, 214–228

Boquet, F.S.J. (2021) *Abortion Is Never a Human Right* [online] available from <https://www.hli.org/2021/04/abortion-is-never-a-human-right/> [30 June 2022]

Borenstein, J. (2011) 'Robots and the Changing Workforce'. *AI and Society* 26 (1), 87–93

Borenstein, J. and Pearson, Y. (2010) 'Robot Caregivers: Harbingers of Expanded Freedom for All?' *Ethics and Information Technology* 12 (3), 277–288

Borge, S. (2003) 'The Myth of Self-Deception'. *The Southern Journal of Philosophy* 41, 1–28

Bosch (2022) *Sense, Think, Act: Three Steps toward Automated Driving* [online] available from <https://www.bosch-mobility-solutions.com/en/mobility-topics/automated-driving-sense-think-act/> [3 November 2022]

Boston Dynamics (2021) *Spot* [online] available from <https://www.bostondynamics.com/spot> [11 October 2021]

Bostrom, N. (2008) 'Dignity and Enhancement'. in *Human Dignity and Bioethics: Essays Commissioned by the President's Council on Bioethics (Washington, DC)* [online] 173–207. available from <https://nickbostrom.com/ethics/dignity-enhancement.pdf> [30 June 2022]

Bourke, L. (2002) 'Do People Prefer General Practitioners of the Same Sex?' *Australian Family Physician* 31 (10), 974–976

Bowman, R. (1991) *The Child | Star Trek: The Next Generation S1 Ep2* [online] available from <https://www.imdb.com/title/tt0708790/> [19 January 2019]

Bratman, M. (1987) *Intention, Plans, and Practical Reason*. Harvard University Press

Bratman, M. (1985) 'Davidson's Theory of Intention'. in *Faces of Intention*. Cambridge University Press, 209–224

Bratton, M.Q. and Chetwynd, S.B. (2004) 'One into Two Will Not Go: Conceptualising Conjoined Twins'. *Journal of Medical Ethics* 30 (3), 279–285

Brison, S.J. (2021) 'What's Consent Got to Do with It?' *Social Philosophy Today* 37, 9–21

British Geriatrics Society (2018) *Depression among Older People Living in Care Homes Report*. Royal College of Psychiatrists. available from <https://www.bgs.org.uk/sites/default/files/content/resources/files/2018-09-

19/Depression%20among%20older%20people%20living%20in%20care%20homes%20report%202018_0.pdf> [27 September 2021]

Brody, H. (2003) ‘Transparency: Informed Consent in Primary Care’. in *Ethical Issues in Modern Medicine*. 6th edn. Steinbock, B., Arras, J.D., and London, A.J. McGraw-Hill, 100–106

Bryson, J. and Winfield, A. (2017) ‘Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems’. *Computer* 50 (5), 116–119

Business Life (2018) ‘Cobots: Rise of the Collaborative Robot’. *British Airways Business Life* November, 12–13

Cahill, A.J. (2016) ‘Unjust Sex vs. Rape’. *Hypatia* 31 (4), 746–761

Cairns, R. (2021) ‘Meet Grace, the Ultra-Lifelike Nurse Robot’. *CNN* [online] 19 August. available from <https://edition.cnn.com/2021/08/19/asia/grace-hanson-robotics-android-nurse-hnk-spc-intl/index.html> [2 November 2022]

Cambridge Dictionary (2021) ‘Deception’. in *Cambridge Dictionary* [online] Cambridge University Press. available from <https://dictionary.cambridge.org/dictionary/english/deception> [11 May 2021]

Cameron, J. (1984) *The Terminator* [online] available from <https://www.imdb.com/title/tt0088247/?ref_=fn_al_tt_1> [2 September 2019]

Campaign Against Sex/Porn Robots (2022) *Home* [online] available from <https://campaignagainstsexrobots.org/>

Campbell, D. (2020) ‘Growing Numbers of NHS Nurses Quit within Three Years, Study Finds | Nursing | The Guardian’. *The Guardian* [online] 23 September. available from <https://www.theguardian.com/society/2020/sep/23/growing-numbers-of-nhs-nurses-quit-within-three-years-study-finds> [20 January 2023]

Canale, S.D., Louis, D.Z., Maio, V., Wang, X., Rossi, G., Hojat, M., and Gonnella, J.S. (2012) ‘The Relationship Between Physician Empathy and Disease Complications: An Empirical Study of Primary Care Physicians and Their Diabetic Patients in Parma, Italy’. *Academic Medicine* 87 (9), 1243–1249

Cappuccio, M.L., Peeters, A., and McDonald, W. (2019) ‘Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition’. *Philosophy & Technology* [online] available from <https://doi.org/10.1007/s13347-019-0341-y> [1 October 2019]

Care Quality Commission (2014) *Human Rights Approach for Our Regulation of Health and Social Care Services*. Care Quality

Commission. available from <https://www.cqc.org.uk/sites/default/files/20140406_our_human_rights_approach_public_consultation_final.pdf> [8 February 2022]

Carson, E. (2019) 'Robots Could Replace Humans in a Quarter of US Jobs by 2030'. *CNET* [online] 24 January. available from <https://www.cnet.com/news/robots-could-replace-humans-in-a-quarter-of-us-jobs-by-2030/> [22 November 2019]

Carson, T.L. (2016) 'Frankfurt and Cohen on Bullshit, Bullshiting, Deception, Lying, and Concern with the Truth of What One Says'. *Pragmatics & Cognition* 23 (1), 53–67

Carson, T.L. (2006) 'The Definition of Lying'. *Noûs* 40, 284–306

Cassam, Q. (2016) 'Vice Epistemology'. *The Monist* 99 (2), 159–180

Cave, E. (2021) 'Valid Consent to Medical Treatment'. *Journal of Medical Ethics* 47 (12), 31–31

CBS Studios (2022) *Data* [online] available from <https://intl.startrek.com/database_article/data> [10 November 2022]

Cellan-Jones, R. (2019) 'Robots "to Replace 20 Million Factory Jobs"'. *BBC News* [online] 26 June. available from <https://www.bbc.com/news/business-48760799> [22 November 2019]

Center for Public Integrity (2008) *Abu Ghraib Prison Scandal* [online] available from <http://publicintegrity.org/politics/abu-ghraib-prison-scandal-2/> [1 April 2022]

Chapman, L., Gray, C., and Headleand, C. (2015) 'A Sense-Think-Act Architecture for Low-Cost Mobile Robotics'. in Bramer, M. and Petridis, M. (eds.) *Research and Development in Intelligent Systems XXXII*, 'SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence'. held 2015. Springer International Publishing, 405–410

Chauvet, D. (2018) 'Should Cultured Meat Be Refused in the Name of Animal Dignity?' *Ethical Theory and Moral Practice* 21 (2), 387–411

Chisholm, R.M. and Feehan, T.D. (1977) 'The Intent to Deceive'. *The Journal of Philosophy* 74 (3), 143

Chomanski, B. (2020) 'If Robots Are People, Can They Be Made for Profit? Commercial Implications of Robot Personhood'. *AI and Ethics*

Chomanski, B. (2019) 'What's Wrong with Designing People to Serve?' *Ethical Theory and Moral Practice* [online] available from <https://link-springer-com.ezproxy.nottingham.ac.uk/article/10.1007%2Fs10677-019-10029-3> [9 September 2019]

Chrisley, R. (2003) 'Embodied Artificial Intelligence'. *Artificial Intelligence* 149 (1), 131–150

Cision (2012) *GeckoSystems' Service Robot, the CareBot, Addresses Care Giver Shortage for Japanese Elderly* [online] available from <https://www.prnewswire.com/news-releases/geckosystems-service-robot-the-carebottm-addresses-care-giver-shortage-for-japanese-elderly-147727405.html> [14 January 2022]

CNN (2015) *Robot Dog Spot Runs Just for Kicks* [online] available from <https://www.cnn.com/videos/entertainment/2015/02/10/erin-pkg-moos-robot-dog.cnn> [8 November 2022]

Cochrane, A. (2010) 'Undignified Bioethics'. *Bioethics* 24 (5), 234–241

Coeckelbergh, M. (2014) 'The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics'. *Philosophy & Technology* 27 (1), 61–77

Coeckelbergh, M. (2010) 'Robot Rights? Towards a Social-Relational Justification of Moral Consideration'. *Ethics and Information Technology* 12 (3), 209–221

Connley, C. (2017) 'Robots May Replace up to 800 Million Workers by 2030'. *CNBC* [online] 30 November. available from <https://www.cnbc.com/2017/11/30/robots-may-replace-up-to-800-million-workers-by-2030.html> [22 November 2019]

Conover, K.J., Acosta, V.M., and Bokoch, R. (2021) 'Perceptions of Ableist Microaggressions among Target and Nontarget Groups'. *Rehabilitation Psychology* 66 (4), 565–575

Conover, K.J., Israel, T., and Nylund-Gibson, K. (2017) 'Development and Validation of the Ableist Microaggressions Scale'. *The Counselling Psychologist* 45 (4), 570–599

Cronqvist, A., Theorell, T., Burns, T., and Lützén, K. (2004) 'Caring About - Caring For: Moral Obligations and Work Responsibilities in Intensive Care Nursing'. *Nursing Ethics* 11 (1), 63–76

Curtice, M.J. and Exworthy, T. (2010) 'FREDA: A Human Rights-Based Approach to Healthcare'. *The Psychiatrist* 34 (4), 150–156

Cusack, J. (2021) *How Driverless Cars Will Change Our World* [online] available from <https://www.bbc.com/future/article/20211126-how-driverless-cars-will-change-our-world> [2 November 2022]

Cyberdyne (2019) *HAL: The World's First Cyborg-Type Robot* [online] available from <http://www.cyberdyne.jp/> [8 March 2019]

Damiano, L. and Dumouchel, P. (2018) 'Anthropomorphism in Human–Robot Co-Evolution'. *Frontiers in Psychology* [online] 9. available from

<https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00468/full> [25 January 2020]

Danaher, J. (2020) 'Robot Betrayal: A Guide to the Ethics of Robotic Deception'. *Ethics and Information Technology* 22 (2), 117–128

Danaher, J. (2019a) 'The Philosophical Case for Robot Friendship'. *Journal of Posthuman Studies* 3 (1), 5–24

Danaher, J. (2019b) 'The Rise of the Robots and the Crisis of Moral Patiency'. *AI and Society* 34 (1), 129–136

Danaher, J. (2019c) 'Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism'. *Science and Engineering Ethics* [online] available from <http://link.springer.com/10.1007/s11948-019-00119-x> [12 September 2019]

Danaher, J. (2019d) 'Building Better Sex Robots: Lessons from Feminist Pornography'. in *AI Love You: Developments on Human-Robot Intimate Relations*. Zhou, Y. and Fischer, M. (eds). Springer International Publishing

Danaher, J. (2017a) 'Should We Be Thinking about Robot Sex?' in *Robot Sex: Social and Ethical Implications*. Danaher, J. and McArthur, N. MIT Press, 3–14

Danaher, J. (2017b) 'The Symbolic-Consequences Argument in the Sex Robot Debate'. in *Robot Sex: Social and Ethical Implications*. Danaher, J. and McArthur, N. MIT Press, 103–132

Danaher, J. (2017c) 'Robotic Rape and Robotic Child Sexual Abuse: Should They Be Criminalised?' *Criminal Law and Philosophy* 11 (1), 71–95

Danaher, J. (2014) 'Sex Work, Technological Unemployment and the Basic Income Guarantee'. *Journal of Evolution and Technology* 24 (1), 113–130

Darwin, C. (1872) *The Expression of the Emotions in Man and Animals*. Oxford University Press

Davidson, D. (2004) 'Deception and Division'. in *Problems of Rationality*. Davidson, D. Oxford University Press, 199–212

Davidson, D. (1978) 'Intending'. in *Essays on Actions and Events*. Oxford University Press, 83–102

D'Cruz, J. (2015) 'Trust, Trustworthiness, and the Moral Consequence of Consistency'. *Journal of the American Philosophical Association* 1 (3), 467–484

De Neys, W. (2017) *Dual Process Theory 2.0*. Taylor & Francis Group

Department of Health (2000) *The NHS Plan*. HMSO. available from <http://1nj5ms2lli5hdggbe3mm7ms5.wpengine.netdna-cdn.com/files/2010/03/pnsuk1.pdf> [10 May 2019]

Derksen, F., Bensing, J., and Lagro-Janssen, A. (2013) 'Effectiveness of Empathy in General Practice: A Systematic Review'. *British Journal of General Practice* 63 (606), 76–84

Deweese-Boyd, I. (2016) 'Self-Deception'. in *The Stanford Encyclopedia of Philosophy* [online] available from <https://plato.stanford.edu/archives/sum2021/entries/self-deception/> [17 November 2022]

Dewey, C. (2015) 'Always Click the First Google Result? You Might Want to Stop Doing That.' *Washington Post* [online] 30 June. available from <https://www.washingtonpost.com/news/the-intersect/wp/2015/06/30/always-click-the-first-google-result-you-might-want-to-stop-doing-that/> [2 February 2023]

Dignitas (2019) *Who Is Dignitas?* [online] available from <http://www.dignitas.ch/index.php?option=com_content&view=article&id=4&Itemid=44&lang=en> [10 May 2019]

Dignity Health (2013) *Americans Rate Kindness as Top Factor in Care* [online] available from <https://www.dignityhealth.org/about-us/press-center/press-releases/majority-of-americans-rate-kindness> [17 January 2023]

Dignity in Care (2023a) *Dignity Champions* [online] available from <https://www.dignityincare.org.uk/> [8 February 2023]

Dignity in Care (2023b) *The 10 Dignity Do's* [online] available from <https://www.dignityincare.org.uk/About/The_10_Point_Dignity_Challenge/> [8 February 2023]

DiMatteo, M.R. and Hays, R. (1980) 'The Significance of Patients' Perceptions of Physician Conduct: A Study of Patient Satisfaction in a Family Practice Center'. *Journal of Community Health,* 6, 18–34

Discover (2022) 'AI Machines Have Beaten Moore's Law Over The Last Decade, Say Computer Scientists'. *Discover Magazine* [online] 21 February. available from <https://www.discovermagazine.com/technology/ai-machines-have-beaten-moores-law-over-the-last-decade-say-computer> [2 February 2023]

Donagan, A. (1977) 'Informed Consent in Therapy and Experimentation'. *Journal of Medicine and Philosophy* 2 (4), 307–329

Dorrier, J. (2020) 'OpenAI Finds Machine Learning Efficiency Is Outpacing Moore's Law'. [17 May 2020] available from

<https://singularityhub.com/2020/05/17/openai-finds-machine-learning-efficiency-is-outpacing-moores-law/> [2 February 2023]

Dou, E. (2021) *Documents Link Huawei to China's Surveillance Programs* [online] available from <https://www.washingtonpost.com/world/2021/12/14/huawei-surveillance-china/> [2 February 2023]

Dougherty, T. (2015) 'Yes Means Yes: Consent as Communication.' *Philosophy and Public Affairs* 43 (3), 224–253

Dougherty, T. (2013) 'Sex, Lies, and Consent'. *Ethics* 123 (4), 717–744

Duffy, B.R. (2003) 'Anthropomorphism and the Social Robot'. *Robotics and Autonomous Systems* 42 (3), 177–190

Echano, M.R. (2017) 'The Motivating Influence of Emotion on Twisted Self-Deception.' *Kritike* 11 (2), 104–120

Edwards, S. (2013) 'Nondoxasticism about Self-Deception'. *Dialectica* 67 (3), 265–282

Ekman, P. (1985) *Telling Lies: Clues to Deceit in the Marketplace, Marriage, and Politics*. W. W. Norton

Elder, A. (2015) 'False Friends and False Coinage: A Tool for Navigating the Ethics of Sociable Robots'. *ACM SIGCAS Computers and Society* 45 (3), 7

Elliott, C. and Quinn, F. (2007) *English Legal System*. 8th edn. Longman

Emotech (2020) *Reimagining the Relationship between Humans and Technology* [online] available from <https://www.emotech.ai/> [18 January 2022]

Encyclopedia Britannica (2022) 'Harold Shipman: British Physician and Serial Killer'. in *Britannica* [online] available from <https://www.britannica.com/biography/Harold-Shipman>

Encyclopedia Britannica (2017) 'Hippocratic Oath'. in *Encyclopedia Britannica* [online] available from <https://www.britannica.com/topic/Hippocratic-oath> [12 March 2019]

Enes, S.P.D. (2003) 'An Exploration of Dignity in Palliative Care'. *Palliative Medicine* 17 (3), 263–269

Eskens, R. (2017) 'Is Sex with Robots Rape?' *Journal of Practical Ethics* 5 (2), 62–76

Ethics & Religious Liberty Commission (2022) *How Abortions Restrictions Protect Human Dignity* [online] available from <https://erlc.com/resource-library/articles/how-abortions-restrictions-protect-human-dignity/> [30 June 2022]

Euronews (2022) ‘Norway on Alert after Drone Sightings near Key Infrastructure’. *Euronews* [online] 20 October. available from <https://www.euronews.com/2022/10/20/norway-on-alert-after-drone-sightings-near-key-infrastructure>

European Commission (2012) *Charter of Fundamental Rights of the European Union* [online] available from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12012P%2FTXT> [30 April 2019]

European Court of Human Rights (1950) *European Convention on Human Rights*. Council of Europe. available from <https://www.echr.coe.int/Documents/Convention_ENG.pdf> [17 June 2019]

Eyal, N. (2019) ‘Informed Consent’. in *The Stanford Encyclopedia of Philosophy* [online] available from <https://plato.stanford.edu/entries/informed-consent/> [18 February 2019]

Eyal, N. (2014) ‘Using Informed Consent to Save Trust’. *Journal of Medical Ethics* 40 (7), 437–444

Eyssel, F., Kuchenbrandt, D., and Bobinger, S. (2011) ‘Effects of Anticipated Human-Robot Interaction and Predictability of Robot Behavior on Perceptions of Anthropomorphism’. in *Proceedings of the 6th International Conference on Human-Robot Interaction* [online] held 6 March 2011 at New York, NY, USA. Association for Computing Machinery, 61–68. available from <http://doi.org/10.1145/1957656.1957673> [15 October 2020]

Fallis, D. (2015) ‘Are Bald-Face Lies Deceptive after All?’ *Ratio* 28 (1), 81–96

Fallis, D. and Stokke, A. (2017) ‘Bullshitting, Lying, and Indifference toward Truth’. *Ergo, an Open Access Journal of Philosophy* [online] 4 (10). available from <http://hdl.handle.net/2027/spo.12405314.0004.010>

Fan, R. and Tao, J. (2004) ‘Consent to Medical Treatment: The Complex Interplay of Patients, Families, and Physicians’. *Journal of Medicine and Philosophy* 29 (2), 139–148

Fandom (2023a) *Holly* [online] available from <https://reddwarf.fandom.com/wiki/Holly> [28 March 2023]

Fandom (2023b) *R2-D2* [online] available from <https://starwars.fandom.com/wiki/R2-D2> [27 January 2023]

Fandom (2023c) *Sonny* [online] available from <https://irobot.fandom.com/wiki/Sonny> [27 January 2023]

Fandom (2022a) *HAL 9000* [online] available from <https://2001.fandom.com/wiki/HAL_9000> [10 November 2022]

Fandom (2022b) *C-3PO* [online] available from <https://starwars.fandom.com/wiki/C-3PO> [10 November 2022]

Farroni, T., Menon, E., Rigato, S., and Johnson, M.H. (2007) 'The Perception of Facial Expressions in Newborns'. *European Journal of Developmental Psychology* 4 (1), 2–13

Faulkner, P. (2007) 'What Is Wrong with Lying?' *Philosophy and Phenomenological Research* 75 (3), 535–557

Federal Trade Commission (2015) *In The Matter of Samsung Electronics, Complaint From The Electronic Privacy Information Center*. Federal Trade Commission. available from <https://epic.org/privacy/internet/ftc/Samsung/EPIC-FTC-Samsung.pdf> [24 September 2020]

Feinberg, J. (1987) *Harm to Others*. The Moral Limits of Criminal Law. vol. 1. Oxford University Press

Ferguson, F. and Gallagher, C. (2020) 'Healthcare Assistant Jailed for Rape of Elderly Woman in Nursing Home'. *The Irish Times* [online] 30 July. available from <https://www.irishtimes.com/news/crime-and-law/courts/criminal-court/healthcare-assistant-jailed-for-rape-of-elderly-woman-in-nursing-home-1.4317757> [4 February 2022]

Fernández-Rodicio, E., Maroto-Gómez, M., Castro-González, Á., Malfaz, M., and Salichs, M.Á. (2022) 'Emotion and Mood Blending in Embodied Artificial Agents: Expressing Affective States in the Mini Social Robot'. *International Journal of Social Robotics* 14 (8), 1841–1864

Ferrario, A., Loi, M., and Viganò, E. (2020) 'Trust Does Not Need to Be Human: It Is Possible to Trust Medical AI'. *Journal of Medical Ethics*

Fink, J. (2012) 'Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction'. in *Proceedings of the 4th International Conference on Social Robotics* [online] held 29 October 2012 at Berlin, Heidelberg. Springer-Verlag, 199–208. available from <http://doi.org/10.1007/978-3-642-34103-8_20> [15 October 2020]

Fitzpatrick, K.K., Darcy, A., and Vierhile, M. (2017) 'Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial'. *Journal of Medical Internet Research* 4 (2), e19

Floridi, L. and Chiriatti, M. (2020) 'GPT-3: Its Nature, Scope, Limits, and Consequences'. *Minds and Machines* 30 (4), 681–694

Frakes, J. (1992) *Offspring | Star Trek: The Next Generation S3 Ep16* [online] available from <https://www.imdb.com/title/tt0708814/> [27 March 2023]

Frank, L. and Nyholm, S. (2017a) 'Robot Sex and Consent: Is Consent to Sex between a Robot and a Human Conceivable, Possible, and Desirable?' *Artificial Intelligence and Law* 25 (3), 305–323

Frank, L. and Nyholm, S. (2017b) 'From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?' in *Robot Sex: Social and Ethical Implications*. ed. by McArthur, N. and Danaher, J. MIT Press, 3–14

Fraunhofer-Gesellschaft (2020) *Care-O-Bot 4 Technical Data* [online] available from <https://www.care-o-bot.de/content/dam/careobot/en/documents/technicaldata/Care-O-bot%204_Technical_Data.pdf> [17 September 2020]

Fraunhofer-Gesellschaft (2018a) *Care-O-Bot 4* [online] available from <https://www.care-o-bot.de/en/care-o-bot-4.html> [23 November 2018]

Fraunhofer-Gesellschaft (2018b) *Fraunhofer Start-up Award 2017 for Mojin Robotics GmbH* [online] available from <https://www.fraunhofer.de/en/press/research-news/2018/February/fraunhofer_start-up_award_2017_for_mojin_Robotics.html> [18 January 2022]

Freeman, L. and Stewart, H. (2021) 'Toward a Harm-Based Account of Microaggressions'. *Perspectives on Psychological Science* 16, 1008–1023

Freeman, L. and Stewart, H. (2019) 'Epistemic Microaggressions and Epistemic Injustice in Clinical Medicine'. in *Overcoming Epistemic Injustice: Social and Psychological Perspectives*. Sherman, B.R. and Goguen, S. Rowman & Littlefield, 121–138

Freter, B. (2018) 'Nursing as Accommodated Care: A Contribution to the Phenomenology of Care. Appeal, Concern, Volition, Practice'. in *Care in Healthcare: Reflections on Theory and Practice* [online] ed. by Krause, F. and Boldt, J. Cham: Palgrave Macmillan, 37–49. available from <https://link.springer.com/content/pdf/10.1007%2F978-3-319-61291-1.pdf>

Fried, C. (1978) *Right and Wrong*. Harvard University Press

Friedlaender, C. (2018) 'On Microaggressions: Cumulative Harm and Individual Responsibility'. *Hypatia* 33 (1), 5–21

Frith, C. (2009) 'Role of Facial Expressions in Social Interactions: A Study with Humans and Robots'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1535), 3453–3458

Fu, G., Lee, K., Cameron, C.A., and Xu, F. (2001) 'Chinese and Canadian Adults' Categorization and Evaluation of Lie- and Truth-Telling about Prosocial and Antisocial Behaviors'. *Journal of Cross-Cultural Psychology* 32 (6), 720–727

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., and Rauws, M. (2018) 'Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial'. *JMIR Mental Health* 5 (4), e64

Galeotti, A. (2016) 'Straight and Twisted Self-Deception'. *Phenomenology and Mind* 11, 90–99

Gardner, H. (2011) *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books

Gardner, J. (2007) *Offences and Defences: Selected Essays in the Philosophy of Criminal Law* [online] Oxford University Press. available from <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199239351.001.0001/acprof-9780199239351-chapter-1>

Gavrell Ortiz, S.E. (2004) 'Beyond Welfare: Animal Integrity, Animal Dignity, and Genetic Engineering'. *Ethics and the Environment* 9 (1), 94–120

Gecko Systems (2019) *Gecko Systems Carebot Benefits* [online] available from <http://www.geckosystems.com/downloads/CareBot%20Benefits%20Overview%20Rev%2011-9-08.pdf> [17 December 2019]

Gecko Systems (2008) *CareBot - Benefits for the Elderly and the Entire Family* [online] available from <https://www.geckosystems.com/markets/CareBot_benefits.php> [14 January 2022]

Geere, D. (2017) *Robot Insects Are Now Faster than the Real Thing* [online] available from <https://www.techradar.com/news/robot-insects-are-now-faster-than-the-real-thing> [19 December 2022]

Gendler, T.S. (2007) 'Self-Deception as Pretense'. *Philosophical Perspectives* 21, 231–258

General Medical Council (2020) *Decision Making and Consent*. General Medical Council

General Medical Council (2013) *Good Medical Practice*. General Medical Council

Genesis 1:26-28 (2022) *Holy Bible* [online] available from <https://biblehub.com/genesis/1-27.htm> [20 July 2022]

Gerdes, A. (2016) 'The Issue of Moral Consideration in Robot Ethics'. *ACM SIGCAS Computers and Society* 45 (3), 274–279

Gerrard, P. (2013) ‘The Hierarchy of the Activities of Daily Living in the Katz Index in Residents of Skilled Nursing Facilities’. *Journal of Geriatric Physical Therapy* 36 (2), 87–91

Ghosh, P. (2020) ‘Sex Robots May Cause Psychological Damage’. *BBC News* [online] 15 February. available from <https://www.bbc.com/news/science-environment-51330261> [1 November 2022]

Ghosh, S. (2018) ‘Facebook Data Scandal Stems from “Move Fast and Break Things” Thesis’. *Business Insider* [online] 22 March. available from <https://www.businessinsider.com/everything-happening-to-facebook-stems-from-its-radical-thesis-of-move-fast-and-break-things-2018-3?IR=T> [12 November 2018]

Gibson, Q.H. (2020) ‘Self-Deception as Omission’. *Philosophical Psychology* 33 (5), 657–678

Gilabert, P. (2019) *Human Dignity and Human Rights*. Oxford University Press

Gilabert, P. (2015) ‘Human Rights, Human Dignity, and Power’. in *Philosophical Foundations of Human Rights*. ed. by Cruft, R., Liao, M., and Renzo, M. Oxford University Press, 196–213

Gill, M.B. (2004) ‘Presumed Consent, Autonomy, and Organ Donation’. *Journal of Medicine and Philosophy* 29 (1), 37–59

Gneezy, U., Kajackaite, A., and Sobel, J. (2018) ‘Lying Aversion and the Size of the Lie’. *American Economic Review* 108 (2), 419–453

Google (2023) *Nest Mini* [online] available from <https://store.google.com/gb/product/google_nest_mini?hl=en-GB?utm_source%3Dweb_opa&utm_medium=google_oo&utm_campaign=GS107430&pli=1> [30 March 2023]

Graumann, S. (2014) ‘Human Dignity and People with Disabilities’. in *The Cambridge Handbook of Human Dignity: Interdisciplinary Perspectives* [online] Düwell, M., Braarvig, J., Brownsword, R., and Mieth, D. Cambridge University Press, 484–491. available from <https://doi.org/10.1017/CBO9780511979033.061>

Grice, H.P. (1975) ‘Logic and Conversation’. in *Syntax and Semantics, Vol. 3*. ed. by Cole, P. and Morgan, J.L. vol. 3. New York: Academic Press, 41–58

Griffiths, P.J. (2004) *Lying: An Augustinian Theology of Duplicity*. Brazos Press

Grodzinsky, F.S., Miller, K.W., and Wolf, M.J. (2015) ‘Developing Automated Deceptions and the Impact on Trust’. *Philosophy & Technology* 28 (1), 91–105

Guarino, B. (2016) 'This Doll in a Hot Car Was so Incredibly Lifelike That Cop Smashed Window, Administered CPR to Save It'. *Washington Post* [online] 18 August. available from <https://www.washingtonpost.com/news/morning-mix/wp/2016/08/18/this-doll-in-a-hot-car-was-so-incredibly-lifelike-that-cop-smashed-window-administered-cpr-to-save-it/> [4 October 2020]

Guevarra, A.R. (2021) 'Here Come the Robot Nurses'. *Boston Review* [online] 2 August. available from <https://www.bostonreview.net/articles/here-come-the-robot-nurses/> [2 November 2022]

Guizzo, E. (2020) *What Is a Robot? Top Roboticists Explain Their Definition of Robot* [online] available from <https://robots.ieee.org/learn/what-is-a-robot/> [3 November 2022]

Guizzo, E. (2010) *Hiroshi Ishiguro: The Man Who Made a Copy of Himself* [online] available from <https://spectrum.ieee.org/hiroshi-ishiguro-the-man-who-made-a-copy-of-himself> [10 November 2022]

Gulland, A. (2021) 'Another Pandemic by 2030 a "Realistic Possibility", Government Warns'. *The Telegraph* [online] 16 March. available from <https://www.telegraph.co.uk/global-health/science-and-disease/another-pandemic-2030-realistic-possibility-government-warns/> [10 February 2023]

Gunkel, D.J. (2018) *Robot Rights*. MIT Press

Haight, R.M. (1980) *A Study of Self-Deception*. Harvester Wheatsheaf

Hall, R. (2019) '"It Never Really Left Me": Abu Ghraib Torture Survivors Finally Get Their Day in Court'. *The Independent* [online] 21 March. available from <https://www.independent.co.uk/news/world/middle-east/iraq-war-abu-ghraib-prison-court-torture-scandal-soldiers-a8831881.html> [1 April 2022]

Hambling, D. (2022) 'Will Ukraine Deploy Lethal Autonomous Drones against Russia?' *New Scientist* [online] 1 November. available from <https://www.newscientist.com/article/2344966-will-ukraine-deploy-lethal-autonomous-drones-against-russia/> [2 November 2022]

Hanson Robotics (2022) *Sophia* [online] available from <https://www.hansonrobotics.com/sophia/> [10 November 2022]

Hansson, S.O., Belin, M., and Lundgren, B. (2021) 'Self-Driving Vehicles: An Ethical Overview'. *Philosophy & Technology* 34 (4), 1383–1408

Harvey, C. (2015) 'Sex Robots and Solipsism'. *Philosophy in the Contemporary World* 22 (2), 80–93

Hashimoto, T., Hitramatsu, S., Tsuji, T., and Kobayashi, H. (2006) 'Development of the Face Robot SAYA for Rich Facial Expressions'. in *2006 SICE-ICASE International Joint Conference*. held October 2006. IEEE, 5423–5428

Hauser, M.D. (1997) 'Minding the Behaviour of Deception'. in *Machiavellian Intelligence II*. ed. by Whiten, A. and Byrne, R.W. Cambridge University Press, 112–143

Hawkins, A.J. (2022) 'Driverless Cars Aren't Going Away, but We Need to Lower Our Expectations about Them'. *The Verge* [online] 28 October. available from <https://www.theverge.com/2022/10/28/23427129/autonomous-vehicles-robotaxi-hype-failure-expectations> [2 November 2022]

Hawley, K. (2015) 'Trust and Distrust between Patient and Doctor'. *Journal of Evaluation in Clinical Practice* 21, 798–801

Hippocrates (1923) 'Decorum'. in *Hippocrates*. vol. 2. Harvard University Press

Hojat, M., Louis, D.Z., Markham, F.W., Wender, R., Rabinowitz, C., and Gonella, J.S. (2011) 'Physicians' Empathy and Clinical Outcomes for Diabetic Patients'. *Academic Medicine* 86 (3), 359–364

Holt, A. (2021) 'Care Staff Shortages Pile Pressure on NHS, Say Hospital Managers'. *BBC News* [online] 13 October. available from <https://www.bbc.com/news/health-58884651> [19 January 2022]

Holt, A. and Burns, J. (2021) 'Coronavirus: Worst Affected Care Homes Revealed by Watchdog'. *BBC News* [online] 21 July. available from <https://www.bbc.com/news/uk-politics-57905821> [10 February 2023]

Hon, K.L. (2013) 'Human Dignity and Rights beyond Death'. *Journal of Medical Ethics* 39 (10), 651–651

Honda (2019a) *Asimo* [online] available from <https://global.honda/innovation/robotics/ASIMO.html> [7 March 2019]

Honda (2019b) *Stride Management Assist* [online] available from <http://asimo.honda.com/innovations/default.aspx?ID=stride-management-assist> [29 March 2019]

Hooks, G. and Mosher, C. (2005) 'Outrages against Personal Dignity: Rationalizing Abuse and Torture in the War on Terror'. *Social Forces* 83 (4), 1627–1645

Horstmann, A.C., Bock, N., Linhuber, E., Szczuka, J.M., Straßmann, C., and Krämer, N.C. (2018) 'Do a Robot's Social Skills and Its Objection

Discourage Interactants from Switching the Robot off?' *PLOS ONE* 13 (7), e0201581

Hotzak, C. (2015) *Robots Will Replace Nurses Sooner Rather than Later* [online] available from <https://www.hhstaff.com/robots-will-replace-nurses-sooner-rather-than-later> [3 March 2019]

Hurd, H.M. (1996) 'The Moral Magic of Consent'. *Legal Theory* 2 (2), 121–146

Hvitved, S. (2022) *What If 99% of the Metaverse Is Made by AI?* [online] available from <https://cifs.dk/news/what-if-99-of-the-metaverse-is-made-by-ai> [2 February 2023]

Impressive Things (2018) 'CleanseBot: This Bacteria Killing Robot Cleans Your Bed Sheets'. [31 December 2018] available from <https://impressivethingspage.com/cleansebot-bacteria-killing-robot/> [29 March 2019]

iRobot (2022) *Roomba Robot Vacuum Cleaners* [online] available from <https://www.irobot.co.uk/en_GB/roomba.html> [27 January 2023]

Isaac, A.M.C. and Bridewell, W. (2017) 'White Lies and Silver Tongues: Why Robots Need to Deceive (and How)'. in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Lin, P., Jenkins, R., and Abney, K. Oxford University Press, 157–172

I-Scoop (2016) *Cobots: The Rise of the Collaborative Robot (Cobot) - What You Need to Know* [online] available from <https://www.i-scoop.eu/industry-4-0/cobot-collaborative-robot/> [20 November 2018]

Isenberg, A. (1973) 'Deontology and the Ethics of Lying'. in *Aesthetics and Theory of Criticism: Selected Essays of Arnold Isenberg*. University of Chicago Press, 245–264

Ishiguro, H. (2018) *Hiroshi Ishiguro Laboratories (HIL)* [online] available from <http://www.geminoid.jp/en/index.html> [23 November 2018]

Jackson, J. (2001) *Truth, Trust and Medicine*. Routledge

Jibo (2014) *Working Together* [online] available from <https://jibo.com/>

Jin, Y. (2018) 'Small Talk in Medical Conversations: Data from China'. *Journal of Pragmatics* 134, 31–44

Johnson, F. (2022) '4 Military Robots With Sophisticated And Terrifying Technology Of The US Marines'. *International Military* [online] 28 October. available from <https://www.international-military.com/2022/10/4-military-robots-with-sophisticated.html> [2 November 2022]

Johnston, J. and Eliot, C. (2003) ‘Chimeras and “Human Dignity”’. *The American Journal of Bioethics* 3 (3), W6–W8

Johnston, M. (1995) ‘Self-Deception and the Nature of Mind’. in *Philosophy of Psychology: Debates on Psychological Explanation*. ed. by MacDonald, C. Blackwell Publishing, 63–91

Jones, K. (2012) ‘Trustworthiness’. *Ethics* 123 (1), 61–85

Jones, T., Lawson, S., and Mills, D. (2008) ‘Interaction with a Zoomorphic Robot That Exhibits Canid Mechanisms of Behaviour’. in *2008 IEEE International Conference on Robotics and Automation*. held May 2008. 2128–2133

Jordan, J. (2016) *Robots*. MIT Press

Joyful Heart Foundation (2022) *Effects of Sexual Assault and Rape* [online] available from <https://www.joyfulheartfoundation.org/learn/sexual-assault-rape/effects-sexual-assault-and-rape> [6 February 2023]

Kaminski, M.E., Rueben, M., Smart, W.D., and Grimm, C.M. (2017) ‘Averting Robot Eyes’. *Maryland Law Review* 76, 983

Kant, I. (2011) *Groundwork of the Metaphysics of Morals* [online] ed. by Gregor, M. and Timmerman, J. Cambridge University Press. available from <https://r2.vlereader.com/Reader?ean=9781139006798>

Kant, I. (1997) *Lectures on Ethics (The Cambridge Edition of the Works of Immanuel Kant)*. ed. by Heath, P. and Schneewind, J. Cambridge University Press

Kant, I. (1996a) ‘Metaphysics of Morals’. in *Kant: Practical Philosophy*. ed. by Gregor, M. Cambridge University Press

Kant, I. (1996b) *Practical Philosophy (The Cambridge Edition of the Works of Immanuel Kant)*. trans. by Gregor, M.J. Cambridge University Press

Kapp, M.B. and Mossman, D. (2014) ‘Measuring Decisional Capacity: Cautions on the Construction of a “Capacimeter”’. *Psychology Public Policy and Law* 2 (1), 73–95

Karpowicz, P., Cohen, C., and van der Kooy, D. (2005) ‘Developing Human-Nonhuman Chimeras in Human Stem Cell Research: Ethical Issues and Boundaries’. *Kennedy Institute of Ethics Journal* 15 (2), 107–134

Karpowicz, P., Cohen, C., and van der Kooy, D. (2004) ‘Is It Ethical to Transplant Human Stem Cells into Nonhuman Embryos?’ *Nature Medicine* 10 (4), 331–335

Kateb, G. (2011) *Human Dignity*. Harvard University Press

Katz, J. (2003) ‘Informed Consent - Must It Remain a Fairy Tale?’ in *Ethical Issues in Modern Medicine*. 6th edn. Steinbock, B., Arras, J.D., and London, A.J. McGraw-Hill, 92–100

Katz, J. (2002) *The Silent World of Doctor and Patient*. Johns Hopkins University Press

Katz, S. and Akpom, C.A. (1976) ‘A Measure of Primary Sociobiological Functions’. *International Journal of Health Sciences* 6 (3), 493–508

Katz, S., Downs, T.D., Cash, H.R., and Grotz, R.C. (1970) ‘Progress in Development of the Index of ADL’. *Gerontologist* 10 (1), 20–30

Katz, S., Ford, A.B., Moskowitz, R.W., Jackson, B.A., and Jaffe, M.W. (1963) ‘Studies of Illness in the Aged. The Index of ADL: A Standardized Measure of Biological and Psychosocial Function’. *Journal of the American Medical Association* 185, 914–919

Kavanagh, C. (2016) ‘Juliette: A Model of Sexual Consent’. *Journal of the International Network for Sexual Ethics and Politics* 4 (1), 43–54

Kavka, G.S. (1983) ‘The Toxin Puzzle’. *Analysis* 43 (1), 33–36

Keller, R.M. and Galgay, C.E. (2010) ‘Microaggressive Experiences of People with Disabilities’. in *Microaggressions and Marginality: Manifestation, Dynamics, and Impact*. ed. by Sue, D.W. John Wiley & Sons, 241–267

Kernisan, L. (2018) *What Are ADLs and IADLs?* [online] available from <https://betterhealthwhileaging.net/what-are-adls-and-iadls/> [22 October 2018]

Kerr, E.P., Coleman, S.A., Kerr, D., Vance, P., Gardiner, B., Zhang, Y., Wang, F., and Wu, C. (2018) ‘Sensor-Based Vital Sign Monitoring, Analysis and Visualisation for Ageing in Place’. in *2018 International Joint Conference on Neural Networks (IJCNN)* [online] held July 2018 at Rio de Janeiro. IEEE, 1–7. available from <https://ieeexplore.ieee.org/document/8489526/> [14 January 2022]

Kerssens, J.J., Bensing, J.M., and Andela, M.G. (1997) ‘Patient Preference for Genders of Health Professionals’. *Social Science & Medicine* 44 (10), 1531–1540

Killmister, S. (2010) ‘Dignity: Not Such a Useless Concept’. *Journal of Medical Ethics* 36 (3), 160–164

Kim, S.S., Kaplowitz, S., and Johnston, M.V. (2004) ‘The Effects of Physician Empathy on Patient Satisfaction and Compliance’. *Evaluation & the Health Professions* 27 (3), 237–251

Kitwood, T. (1997) *Dementia Reconsidered: The Person Comes First*. Open University Press

Klein, M. (2005) ‘Euthanasia and the doctrine of double effect’. *Wurzburger Medizinhistorische Mitteilungen* 24, 51–62

Kolnai, A. (1976) ‘Dignity’. *Philosophy* 51 (197), 251–271

Kongsholm, N.C.H. and Kappel, K. (2017) ‘Is Consent Based on Trust Morally Inferior to Consent Based on Information?’ *Bioethics* 31 (6), 432–442

Kubrick, S. (1968) *2001: A Space Odyssey* [online] available from <https://www.imdb.com/title/tt0062622/> [18 September 2020]

Kupfer, J. (1982) ‘The Moral Presumption against Lying’. *The Review of Metaphysics* 36 (1), 103–126

Lamb, S., Gable, S., and de Ruyter, D. (2021) ‘Mutuality in Sexual Relationships: A Standard of Ethical Sex?’ *Ethical Theory and Moral Practice* 24, 271–284

Lancaster, K. (2022) ‘Granny and the Sexbots: An Ethical Appraisal of the Use of Sexbots in Residential Care Institutions for Elderly People’. in *Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds With Robots* [online] ed. by Loh, J. and Loh, W. Colombia University Press, 181–208. available from <https://philpapers.org/archive/LANGAT-12.pdf>

Lancaster, K. (2021) ‘Non-Consensual Personified Sexbots: An Intrinsic Wrong’. *Ethics and Information Technology* 23 (4), 589–600

Lancaster, K. (2019) ‘The Robotic Touch: Why There Is No Good Reason to Prefer Human Nurses to Carebots’. *Philosophy in the Contemporary World* 25 (2), 88–109

Landau, L. (1994) *Clues* | *Star Trek: The Next Generation S4 Ep14*.

Langley, L. (2017) *Here Are the Best Liars in the Animal Kingdom* [online] available from <https://www.nationalgeographic.com/animals/article/animals-lying-liars-birds-squid> [23 November 2022]

Langton, R. (2009) *Sexual Solipsism: Philosophical Essays on Pornography and Objectification* [online] Oxford University Press. available from <https://www-oxfordscholarship-com.ezproxy.nottingham.ac.uk/view/10.1093/acprof:oso/9780199247066.001.0001/acprof-9780199247066>

Latikka, R., Turja, T., and Oksanen, A. (2019) ‘Self-Efficacy and Acceptance of Robots’. *Computers in Human Behavior* 93, 157–163

Law Commission (1991) *Criminal Law: Rape within Marriage*. HMSO. available from <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/228746/0167.pdf> [22 November 2018]

Lazar (1999) ‘Self-Deception and the Desire to Believe’. *Mind* 108, 263–290

Leclercq, W.K.G., Keulers, B.J., Scheltinga, M.R.M., Spauwen, P.H.M., and van der Wilt, G.J. (2010) ‘A Review of Surgical Informed Consent: Past, Present, and Future. A Quest to Help Patients Make Better Decisions’. *World Journal of Surgery* 34 (7), 1406–1415

Lee, K., Cameron, C.A., Xu, F., Fu, G., and Board, J. (1997) ‘Chinese and Canadian Children’s Evaluations of Lying and Truth Telling: Similarities and Differences in the Context of Pro- and Antisocial Behaviors’. *Child Development* 68 (5), 924–934

Leeuwen, N.V. (2013) ‘Defining Self-Deception and Solving the Paradoxes’. in *The International Encyclopedia of Ethics* [online] ed. by LaFollette. Blackwell Publishing, 1–11. available from <https://philpapers.org/archive/VANS-3.pdf>

Leland, J. (2008) ‘In “Sweetie” and “Dear,” A Hurtful Message for the Elderly’. *The New York Times* [online] 7 October. available from <https://www.nytimes.com/2008/10/07/world/americas/07iht-07aging.16738425.html> [31 October 2022]

Leong, B. and Selinger, E. (2019) ‘Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism’. in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19* [online] held 2019 at Atlanta, GA, USA. ACM Press, 299–308. available from <http://dl.acm.org/citation.cfm?doid=3287560.3287591> [22 September 2020]

Levine, E.E. and Schweitzer, M.E. (2015) ‘Prosocial Lies: When Deception Breeds Trust’. *Organizational Behavior and Human Decision Processes* 126, 88–106

Levy, K. (2009) ‘On the Rationalist Solution to Gregory Kavka’s Toxin Puzzle’. *Pacific Philosophical Quarterly* 90 (2), 267–289

Levy, N. (2004) ‘Self-Deception and Moral Responsibility’. *Ratio* 17, 294–311

Lilienfeld, S.O. (2017) ‘Microaggressions: Strong Claims, Inadequate Evidence’. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 12 (1), 138–169

Living with Dignity (2010) *Arguments against Euthanasia* [online] available from <https://vivredignite.org/en/against-euthanasia/> [10 May 2019]

Livingston, D. (2000) *Nothing Human | Star Trek: Voyager S5 E8* [online] available from <https://www.imdb.com/title/tt0708939/> [18 January 2022]

LoBue, V. (2016) ‘Face Time: Here’s How Infants Learn from Facial Expressions’. *The Conversation* [online] available from <https://rb.gy/clftox>

Locke, J. (2016) *Second Treatise of Government and A Letter Concerning Toleration*. ed. by Goldie, M. Oxford World’s Classics. Oxford University Press

Locke, J. (1988) ‘Second Treatise on Civil Government’. in *Locke: Two Treatises of Government*. ed. by Laslett, P. Cambridge University Press

Lorente, 2020 (2020) ‘Driven towards a Moral Crash’. *Rivista Internazionale Di Filosofia e Psicologia* 11 (2), 223–237

Lucas, G.M., Gratch, J., King, A., and Morency, L.-P. (2014) ‘It’s Only a Computer: Virtual Humans Increase Willingness to Disclose’. *Computers in Human Behavior* 37, 94–100

Luka Inc. (2020) *Replika: My AI Friend* [online] available from <https://play.google.com/store/apps/details?id=ai.replika.app&hl=en_GB>

Lupoli, M.J., Jampol, L., and Oveis, C. (2017) ‘Lying Because We Care: Compassion Increases Prosocial Lying’. *Journal of Experimental Psychology: General* 146, 1026–1042

LuxAI (2022) *QTrobot: The Autism Robot Tutor for Improving Your Child’s Learning Outcome at Home!* [online] available from <https://luxai.com/robot-for-teaching-children-with-autism-at-home/> [27 January 2023]

MacIntyre, A. (1994) ‘Truthfulness, Lies, and Moral Philosophers: What Can We Learn from Mill and Kant?’ in *Tanner Lectures on Human Values* [online] Princeton University Press, 309–369. available from <https://tannerlectures.utah.edu/_resources/documents/a-to-z/m/macintyre_1994.pdf> [25 June 2021]

Macklin, R. (2003) ‘Dignity Is a Useless Concept’. *BMJ : British Medical Journal; London* 327 (7429), 1419

Magee, H., Parsons, S., and Askham, J. (2008) *Measuring Dignity in Care for Older People*. Picker Institute Europe for Help the Aged. available from <https://www.ageuk.org.uk/documents/en-gb/for-professionals/research/measuring%20dignity%20in%20care%20(2008)_pro.pdf?dtrk=true> [30 April 2019]

Mahon, J.E. (2016) ‘The Definition of Lying and Deception’. in *The Stanford Encyclopedia of Philosophy* [online] ed. by Zalta, E.N. available from <https://plato.stanford.edu/archives/win2016/entries/lying-definition/> [15 March 2021]

Mamak, K. (2022) ‘Should Criminal Law Protect Love Relation with Robots?’ *AI and Society* 1–10

Mannison, D.S. (1969) ‘Lying and Lies’. *Australasian Journal of Philosophy* 47, 132–144

Margalit, A. (2017) *On Betrayal*. Harvard University Press

Marsili, N. (2014) ‘Lying as a Scalar Phenomenon: Insincerity along the Certainty-Uncertainty Continuum’. in *Certainty-Uncertainty – and the Attitudinal Space in Between:* [online] ed. by Cantarini, S., Abraham, W., and Leiss, E. Studies in Language Companion. Amsterdam: John Benjamins Publishing Company, 153–173. available from <https://benjamins.com/catalog/slcs.165.09mar> [21 June 2021]

Matiti, M. and Cotrel-Gibbons, L. (2006) ‘Patient Dignity - Promoting Good Practice Project’. in *Foundation of Nursing Studies Dissemination Series, 3(5)*. 1–4

Matthias, A. (2015) ‘Robot Lies in Health Care: When Is Deception Morally Permissible?’ *Kennedy Institute of Ethics Journal* 25 (2), 169–162

Matuszek, C. (2017) *How Robots Could Help Bridge the Elder-Care Gap* [online] available from <http://theconversation.com/how-robots-could-help-bridge-the-elder-care-gap-82125> [28 March 2019]

McArdle, M. (2015) ‘How Grown-Ups Deal With “Microaggressions”’. *Bloomberg* [online] 11 September. available from <https://www.bloomberg.com/opinion/articles/2015-09-11/how-grown-ups-deal-with-microaggressions-> [10 June 2022]

McArthur, N. (2017) ‘The Case for Sexbots’. in *Robot Sex: Social and Ethical Implications*. Danaher, J. and McArthur, N. MIT Press, 3–14

McCue, T.J. (2019) ‘Alexa Is Listening All The Time: Here’s How To Stop It’. *Forbes* [online] 19 April. available from <https://www.forbes.com/sites/tjmccue/2019/04/19/alexa-is-listening-all-the-time-heres-how-to-stop-it/> [26 November 2021]

McGregor, J. (2005) *Is It Rape? On Acquaintance Rape and Taking Women’s Consent Seriously*. Ashgate

McGregor, J. (1994) ‘Force, Consent and the Reasonable Woman’. in *In Harm’s Way: Essays in Honor of Joel Feinberg*. ed. by Coleman, J.L. and Buchanan, A. Cambridge University Press, 231–254

McGuinness, R. (2022) ‘Care Worker Raped 99-Year-Old Dementia Sufferer While Shocked Family Watched on Hidden Camera’. *Yahoo News* [online] 8 February. available from <https://www.aol.co.uk/news/care-worker-99-old-dementia-091509307.html> [8 February 2022]

McKimm, J. (2021) *How COVID-19 Is Affecting Nurses’ Mental Health, and What to Do about It* [online] available from <https://rcni.com/nursing-

standard/features/how-covid-19-affecting-nurses-mental-health-and-what-to-do-about-it-159456> [20 January 2023]

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F.J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C.J., King, D., Ledsam, J.R., Melnick, D., Mostofi, H., Peng, L., Reicher, J.J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K.C., Fauw, J.D., and Shetty, S. (2020) 'International Evaluation of an AI System for Breast Cancer Screening'. *Nature* 577 (7788), 89–94

Meacham, D. and Studley, M. (2017) 'Could a Robot Care? It's All in the Movement'. in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, 97–112

Mele, A.R. (2001) *Self-Deception Unmasked*. Princeton University Press

Mele, A.R. (1999) 'Twisted Self-Deception'. *Philosophical Psychology* 12 (2), 117–137

Merlin Entertainment Group (2020) *Star Wars Experience Exhibition* [online] available from <https://www.madametussauds.com/london/whats-inside/zones/star-wars/> [21 November 2020]

Microsoft (2016) *Learning from Tay's Introduction* [online] available from <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/> [6 February 2023]

Mill, J.S. (1998) *Utilitarianism*. Oxford University Press

Mill, J.S. (1974) *On Liberty*. Penguin

Miller, M. (2021) *The next Pandemic Is Already Happening – Targeted Disease Surveillance Can Help Prevent It* [online] available from <http://theconversation.com/the-next-pandemic-is-already-happening-targeted-disease-surveillance-can-help-prevent-it-160429> [10 February 2023]

Miller, M. (2017) 'The Future of U.S. Caregiving: High Demand, Scarce Workers'. *Reuters* [online] 3 August. available from <https://www.reuters.com/article/us-column-miller-caregivers-idUSKBN1AJ1JQ> [27 November 2019]

Molteni, M. (2017) 'The Chatbot Therapist Will See You Now'. *Wired* [online] 7 June. available from <https://www.wired.com/2017/06/facebook-messenger-woebot-chatbot-therapist/> [18 June 2019]

Moon, M. (2017) 'Surgical Robot Makes Highly Precise Eye Injection Possible'. *Engadget* [online] 28 January. available from <https://www.engadget.com/2017/01/28/surgical-robot-highly-precise-eye-surgery/> [29 March 2019]

Mori, M., MacDorman, K.F., and Kageki, N. (2012) 'The Uncanny Valley [From the Field]'. *IEEE Robotics & Automation Magazine* 19 (2), 98–100

Müller, V.C. (2016) 'Autonomous Killer Robots Are Probably Good News'. in *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*. Di Nucci, E. and Santoni de Sio, F. Ashgate, 67–81

Mulvey, B. (2018) 'Can Humans and Robots Be Friends?' *Dialogue and Universalism* 28 (2), 49–64

Nash, A. and Open Robotics (2022) *Sense Think Act: Podcast Series* [online] available from <https://www.sensethinkact.com/> [3 November 2022]

National Archives (2005) *Mental Capacity Act 2005* [online] available from <https://www.legislation.gov.uk/ukpga/2005/9/contents> [26 February 2019]

National Archives (2003) *Sexual Offences Act 2003* [online] available from <https://www.legislation.gov.uk/ukpga/2003/42/contents> [14 January 2019]

National Institute for Health and Care Excellence (2013) *Mental Wellbeing of Older People in Care Homes* | *Guidance and Guidelines* [online] available from <https://www.nice.org.uk/guidance/qs50/chapter/Introduction> [7 December 2018]

Nelson, S. and Gordon, S. (2006) *The Complexities of Care: Nursing Reconsidered*. Cornell University Press

Newey, G. (1997) 'Political Lying: A Defense'. *Public Affairs Quarterly* 11, 93–116

News18 (2021) 'Marital Rape Is Not a Crime in 32 Countries. One of Them Is India'. *News18* [online] 26 August. available from <https://www.news18.com/news/india/marital-rape-is-not-a-crime-in-32-countries-one-of-them-is-india-4130363.html> [15 September 2022]

Newton, J.P. (2008) 'Who Needs Friends…when Robots May Be the Answer?' *Gerodontology* 25 (2), 65–66

NHS (2022) *Consent to Treatment - Assessing Capacity* [online] available from <https://www.nhs.uk/conditions/consent-to-treatment/capacity/> [28 January 2023]

NHS (2018a) *Consent to Treatment* [online] available from <https://www.nhs.uk/conditions/consent-to-treatment/> [28 September 2018]

NHS (2018b) *Mental Capacity Act* [online] available from <https://www.nhs.uk/conditions/social-care-and-support-guide/making-decisions-for-someone-else/mental-capacity-act/> [22 March 2023]

NHS (2017) *Consent to Treatment - Children and Young People* [online] available from <https://www.nhs.uk/conditions/consent-to-treatment/children/> [7 February 2023]

NHS Support Federation (2022) *Staff Shortages* [online] available from <https://nhsfunding.info/symptoms/10-effects-of-underfunding/staff-shortages/> [19 January 2022]

Nieves, M., Sanches-Fuentes, M., Parra-Barrera, S.M., and Granados de Haro, R. (2022) 'Only "Yes" Means "Yes": Negotiation of Sex and Its Link With Sexual Violence'. *Journal of Interpersonal Violence*

No Means No Worldwide (2018) *About* [online] available from <https://www.nomeansnoworldwide.org/about> [22 February 2019]

Noddings, N. (2003) *Caring: A Feminine Approach to Ethics and Moral Education*. 2nd edn. University of California Press

Nordenfelt, L. (2004) 'The Varieties of Dignity'. *Health Care Analysis* 12 (2), 69–81

Norwick, P., Weston, G.K., and Grant-Kels, J.M. (2019) 'Erection Ethics'. *Journal of the American Academy of Dermatology* 81 (5), 1225

Nourbakhsh, I. (2013) *Robot Futures*. MIT Press

Nursegroups (2022) *The Nightingale Pledge: Nursing Ethics Oath* [online] available from <https://www.nursegroups.com/nightingale-pledge-nursing-ethics-oath> [29 September 2022]

Nursing and Midwifery Council (2018) *The Code: Professional Standards of Practice and Behaviour for Nurses, Midwives and Nursing Associates* [online] available from <https://www.nmc.org.uk/globalassets/sitedocuments/nmc-publications/nmc-code.pdf> [21 January 2019]

Nursing and Midwifery Council (2002) *Practitioner-Client Relationships and the Prevention of Abuse*. Nursing and Midwifery Council

Nussbaum, M.C. (1995) 'Objectification'. *Philosophy and Public Affairs* 24 (4), 249–291

Nyholm, S. (2020) *Humans and Robots: Ethics, Agency and Anthropomorphism*. Rowman & Littlefield

OED (1989) 'Deception'. in *Oxford English Dictionary*. Clarendon Press

Office for National Statistics (2020) *Coronavirus and Loneliness in Great Britain* [online] available from <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/coronavirusandlonelinessgreatbritain/3aprilto3may2020> [10 February 2023]

Oliver, D. and Gee, P. (2009) 'Communication, Barriers to It and Information Sharing'. in *Medical Ethics and the Elderly*. 3rd edn. Rai, G.S. Radcliffe Publishing, 49–61

Omnicell (2023) *Medicines Management Solutions | Pharmacy Optimisation* [online] available from <https://www.omnicell.co.uk/> [19 January 2023]

O'Neill, O. (2002) *Autonomy and Trust in Bioethics* [online] Cambridge University Press. available from <https://ebookcentral.proquest.com/lib/nottingham/reader.action?docID=202188&ppg=7> [19 March 2019]

Open AI (2023) *ChatGPT-4* [online] available from <https://chatgpt.pro/> [27 January 2023]

OpenAI (2022) *ChatGPT: Optimizing Language Models for Dialogue* [online] available from <https://openai.com/blog/chatgpt/> [2 February 2023]

Opinion Leader (2009) *Final Report on the Review of the Department of Health Dignity in Care Campaign*. Department of Health. available from <https://www.dignityincare.org.uk/_assets/Opinion_Leader_Final_Report_to_DH.doc.pdf> [7 February 2022]

O'Shea, T. (2018) 'Consent: Historical Perspectives in Medical Ethics'. in *Routledge Handbook of the Ethics of Consent*. ed. by Müller, A. and Schaber, P. Routledge, 261–271

Osterloff, E. (2021) *Why Do Some Butterflies and Moths Have Eyespots?* [online] available from <https://www.nhm.ac.uk/discover/why-do-butterflies-have-eyespots.html> [5 July 2021]

Oxford Reference (2023) *Constant Dropping Wears Away a Stone* [online] available from <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095633677;jsessionid=596B79D7DB6E6A0567C8741D93FDEAF4> [6 February 2023]

Palpant, N.J. and Holland, S. (2012) 'Human Dignity and the Debate Over Early Human Embryos'. in *Human Dignity in Bioethics: From Worldviews to the Public Square*. ed. by Dilley, S. and Palpant, N.J. Routledge

Panorama (2014) *Behind Closed Doors: Elderly Care Exposed* [online] available from <https://www.bbc.co.uk/programmes/b042rcjp> [8 February 2022]

Park, D. (2016) *EL-E: An Assistive Robot – Healthcare Robotics Lab* [online] available from <https://sites.gatech.edu/hrl/el-e-an-assistive-robot/> [27 January 2023]

Parke, P. (2015) *Is It Cruel to Kick a Robot Dog?* [online] available from <https://www.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html> [8 November 2022]

Petersen, S. (2017) 'Is It Good for Them Too? Ethical Concern for the Sexbots'. in *Robot Sex: Social and Ethical Implications*. ed. by Danaher, J. and McArthur, N. MIT Press, 155–172

Petersen, S. (2011) 'Designing People to Serve'. in *Robot Ethics*. Lin, P., Bekey, G., and Abney, K. MIT Press, 283–298

Petiprin, A. (2020) *Virginia Henderson - Nursing Theorist* [online] available from <https://nursing-theory.org/nursing-theorists/Virginia-Henderson.php> [24 January 2022]

Phillips, F. (2012) *Seeing the Scandal of Abuse in Our Elderly Care Homes Made Me Weep* [online] available from <https://www.mirror.co.uk/news/uk-news/fiona-phillips-panorama-elderly-care-802647> [4 February 2022]

Phys.org (2009) *Venomous Sea Snakes Play Heads or Tails with Their Predators* [online] available from <https://phys.org/news/2009-08-venomous-sea-snakes-tails-predators.html> [23 November 2022]

Picard, A. (2010) *Time to End Pelvic Exams Done without Consent* [online] available from <https://www.theglobeandmail.com/amp/life/health-and-fitness/time-to-end-pelvic-exams-done-without-consent/article4325965/> [11 March 2019]

Pierce, C., Carew, J., Pierce-Gonzales, D., and Wills, D. (1978) 'An Experiment in Racism'. in *Television and Education*. Sage, 62–88

Plato (2000) *Laws*. trans. by Jowett, B. Prometheus Books

Plato (1925) *Statesman*. trans. by Bury, R.G. Heinemann

Pokomy, M. (2017) 'Nursing Theorists of Historical Significance: Virginia Henderson'. in *Nursing Theorists and Their Work*. 9th edn. Alligood, M.R. USA: Elsevier, 11–28

Pollack, M.E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J.T., Dunbar-Jacob, J., McCarthy, C.E., Thrun, S., Montemerlo, M., Pineau, J., and Roy, N. (2002) 'Pearl: A Mobile Robotic Assistant for the Elderly'. in *Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care*,

‘Association for the Advancement of Artificial Intelligence (AAAI)’. held 2002 at Alberta, Canada

Pransky, J. (1996) ‘Service Robots–How We Should Define Them?’ *Service Robot: An International Journal* 2 (1), 4–5

Price, D. (1997) ‘Euthanasia, Pain Relief and Double Effect’. *Legal Studies* 17 (2), 323–342

Primoratz, I. (1984) ‘Lying and the “Methods of Ethics"’. *International Studies in Philosophy* 16, 35–57

Quick, D. (2010) *Healthcare Robot Gives Sponge Baths* [online] available from <https://newatlas.com/spong-bath-assistant-robot/16894/> [29 March 2019]

Rainey, S. (2016) ‘Friends, Robots, Citizens?’ *ACM SIGCAS Computers and Society* 45 (3), 225–233

Rakel, D.P., Hoeft, T.J., Barrett, B.P., Chewning, B.A., Craig, B.M., and Niu, M. (2009) ‘Practitioner Empathy and the Duration of the Common Cold’. *Family Medicine* 41 (7), 494–501

Rape Crisis (2023) *Impacts of Sexual Violence and Abuse* [online] available from <https://rapecrisis.org.uk/get-informed/about-sexual-violence/impacts-of-sexual-violence-and-abuse/> [6 February 2023]

Realdoll (2019) *Build Your Realdoll* [online] available from <https://www.realdoll.com/product/build-your-realdoll/> [14 September 2019]

Reed, J. and McCormack, B. (2012) ‘Independence and Autonomy: The Foundation of Care’. in *Nursing Older People*. ed. by Reed, J., Clarke, C., and MacFarlane, A. Open University Press, 9–22

Ren, Z. (Bella), Hart, E., Levine, E.E., and Schweitzer, M.E. (2022) ‘The Shared Responsibility Model of Deception’. *Current Opinion in Psychology* 48, 101470

Resnik, D.B. (2007) ‘Embryonic Stem Cell Patents and Human Dignity’. *Health Care Analysis* 15 (3), 211–222

Richardson, K. (2019) *Sex Robots: The End of Love*. Polity Press

Richardson, K. (2016) ‘The Asymmetrical “Relationship”: Parallels Between Prostitution and the Development of Sex Robots’. *SIGCAS Computers & Society* 45 (3), 290–293

Riken (2019) *RIBA-II, the Next Generation Care-Giving Robot* [online] available from <http://www.riken.jp/en/pr/press/2011/20110802_2/> [7 March 2019]

Roache, R. (2014) ‘Why Is Informed Consent Important?’ *Journal of Medical Ethics* 40 (7), 435–436

Robertson, J. (2010) ‘Gendering Humanoid Robots: Robo-Sexism in Japan’. *Body and Society* 16 (2), 1–36

Robillard, M. (2018) ‘No Such Thing as Killer Robots’. *Journal of Applied Philosophy* 35 (4), 705–717

Robot Center Ltd (2021) *Telepresence Robots* [online] available from <https://www.robotcenter.co.uk/collections/telepresence-robots> [19 December 2021]

Robot Center Ltd (2020) *PaPeRo Robot NEC* [online] available from <https://www.robotcenter.co.uk/products/papero-robot-nec> [27 January 2020]

Robotics Today (2021) *Pearl* [online] available from <https://www.roboticstoday.com/robots/pearl-description> [10 December 2021]

Robots (2022) *BigDog* [online] available from <https://robots.ieee.org/robots/bigdog/> [26 January 2023]

Rogers, Y. and Marsden, G. (2013) ‘Does He Take Sugar? Moving beyond the Rhetoric of Compassion’. *Interactions* 20 (4), 48–57

Rolf, S. (2012) ‘Human Embryos and Human Dignity: Differing Presuppositions in Human Embryo Research in Germany and Great Britain’. *Heythrop Journal* 53 (5), 742–754

Romeo, N. (2016) ‘The Chatbot Will See You Now’. *The New Yorker* [online] 25 December. available from <https://www.newyorker.com/tech/annals-of-technology/the-chatbot-will-see-you-now> [18 June 2019]

Rorty, A.O. (1994) ‘User-Friendly Self-Deception’. *Philosophy* 69, 211–228

Royakkers, L. and van Est, R. (2015a) *Just Ordinary Robots: Automation from Love to War*. CRC Press

Royakkers, L. and van Est, R. (2015b) ‘A Literature Review on New Robotics: Automation from Love to War’. *International Journal of Social Robotics* 7 (5), 549–570

Royal College of Nursing (2017) *Nursing Principles of Consent*. RCN

Russell, P. (1986) ‘Locke on Express and Tacit Consent’. *Political Theory* 14 (2), 291–306

Sandoz, E. (2021) ‘Beyond “Yes Means Yes”: A Behavioral Conceptualization of Affirmative Sexual Consent’. *Behavior and Social Issues* 30 (1), 712–731

Saul, J. (2012) ‘Just Go Ahead and Lie’. *Analysis* 72 (1), 3–9

Saunders, B. (2012) ‘Opt-Out Organ Donation Without Presumptions’. *Journal of Medical Ethics* 38 (2), 69–72

Schauer, F. and Zeckhauser, R.J. (2009) ‘Paltering’. in *Deception: From Ancient Empires to Internet Dating*. Stanford University Press, 38–54

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S.S., and Tow, R. (2002) ‘Experiences with Sparky, a Social Robot’. in *Socially Intelligent Agents. Multiagent Systems, Artificial Societies, and Simulated Organizations* [online] ed. by Dautenhahn, K., Bond, A., Cañamero, L., and Edmonds, B. vol. 3. Springer, 173–180. available from <https://link.springer.com/chapter/10.1007/0-306-47373-9_21> [8 December 2022]

Scheerer, R. (1989) *The Measure of a Man | Star Trek: The Next Generation S2 Ep9* [online] available from <https://www.imdb.com/title/tt0708807/> [9 March 2019]

Schlosser, M. (2019) ‘Agency’. in *The Stanford Encyclopedia of Philosophy* [online] available from <https://plato.stanford.edu/archives/win2019/entries/agency/> [14 November 2022]

Schroeder, D. (2010) ‘Dignity - One, Two, Three, Four, Five; Still Counting’. *Cambridge Quarterly of Healthcare Ethics* 19 (1), 118–125

Schroeder, D. (2008) ‘Dignity - Two Riddles and Four Concepts’. *Cambridge Quarterly of Healthcare Ethics* 17 (2), 230–238

Schulhofer, S.J. (1995) ‘The Feminist Challenge in Criminal Law’. *University of Pennsylvania Law Review* 143 (6), 2151–2207

Schulz, T. (2013) ‘Man vs. Machine: Are Any Jobs Safe from Innovation?’ *Spiegel Online* [online] 3 May. available from <http://www.spiegel.de/international/business/speed-of-innovation-and-automation-threatens-global-labor-market-a-897412-2.html> [29 March 2019]

ScienceDaily (2018) *Robot Teaches Itself How to Dress People: Instead of Vision, Machine Relies on Force as It Pulls a Gown onto Human Arms* [online] available from <https://www.sciencedaily.com/releases/2018/05/180514122506.htm> [27 January 2023]

Scott, G.G. (2006) *The Truth about Lying*. ASJA Press

Scott, R. (1982) *Blade Runner* [online] available from <https://www.imdb.com/title/tt0083658/> [12 February 2020]

Searcy, W.A. and Nowicki, S. (2005) *The Evolution of Animal Communication*. Princeton University Press

Secom (2019) *MySpoon - Meal-Assistance Robot* [online] available from <https://www.secom.co.jp/english/myspoon/> [29 March 2019]

Sense Medical (2022) *Paro: About* [online] available from <https://www.paroseal.co.uk/> [11 April 2022]

Sensen, O. (2011a) 'Human Dignity in Historical Perspective: The Contemporary and Traditional Paradigms'. *European Journal of Political Theory* 10 (1), 71–91

Sensen, O. (2011b) *Kant on Human Dignity*. De Gruyter

Seppala, E.M., Hutcherson, C.A., Nguyen, D.T., Doty, J.R., and Gross, J.J. (2014) 'Loving-Kindness Meditation: A Tool to Improve Healthcare Provider Compassion, Resilience, and Patient Care'. *Journal of Compassionate Health Care* 1 (1), 5

Shapiro, M. (2012) *Transforming the Nature of Health: A Holistic Vision of Healing That Honors Our Connection to the Earth, Others, and Ourselves*. North Atlantic Books

Sharkey, A. (2014) 'Robots and Human Dignity: A Consideration of the Effects of Robot Care on the Dignity of Older People'. *Ethics and Information Technology* 16 (1), 63–75

Sharkey, A. and Sharkey, N. (2020) 'We Need to Talk about Deception in Social Robotics!' *Ethics and Information Technology* 23 (3), 309–316

Sharkey, A. and Sharkey, N. (2012) 'Granny and the Robots: Ethical Issues in Robot Care for the Elderly'. *Ethics and Information Technology* 14 (1), 27–40

Sharkey, N. and Sharkey, A. (2010) 'The Crying Shame of Robot Nannies: An Ethical Appraisal'. *Interaction Studies* 11 (2), 161–190

Shelkey, M. and Wallace, M. (2012) 'Katz Index of Independence in Activities of Daily Living (ADL)'. *The Hartford Institute of Geriatric Nursing* 2, 1–2

Shen, F.X. (2020) 'Sex Robots Are Here, but Laws Aren't Keeping up with the Ethical and Privacy Issues They Raise'. *Yahoo News* [online] 27 June. available from <https://news.yahoo.com/sex-robots-laws-arent-keeping-172111822.html> [1 November 2022]

Shim, J. and Arkin, R.C. (2016) 'Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI?' in Agah, A., Cabibihan, J.-J., Howard, A.M., Salichs, M.A., and He, H. (eds.) *Social Robotics*. held 2016 at Cham. Springer International Publishing, 222–232

Siegel, M. (2003) 'The Sense-Think-Act Paradigm Revisited'. in *Proceedings of the 1st International Workshop on Robotic Sensing* [online] held 5 June 2003 at Orebro University, Sweden. IEEE. available from

<file:///C:/Users/karen/Downloads/sense-think-act_paradigm_revisited-01218700.pdf>

Siegler, F.A. (1966) 'Lying'. *American Philosophical Quarterly* 3, 128–136

Sillice, M.A., Morokoff, P.J., Ferszt, G., Bickmore, T., Bock, B.C., Lantini, R., and Velicer, W.F. (2018) 'Using Relational Agents to Promote Exercise and Sun Protection: Assessment of Participants' Experiences With Two Interventions'. *Journal of Medical Internet Research* 20 (2), e48

Siripala, T. (2018) 'Japan's Robot Revolution in Senior Care'. *The Diplomat* [online] 1 June. available from <https://thediplomat.com/2018/06/japans-robot-revolution-in-senior-care/> [28 March 2019]

Skills for Care (2021) *The State of the Adult Social Care Sector and Workforce in England* [online] available from <https://www.skillsforcare.org.uk/adult-social-care-workforce-data/Workforce-intelligence/publications/national-information/The-state-of-the-adult-social-care-sector-and-workforce-in-England.aspx> [19 January 2022]

Skinta, M. and Torres-Harding, S. (2022) 'Confronting Microaggressions: Developing Innovative Strategies to Challenge and Prevent Harm'. *New Ideas in Psychology* 65

Skyrms, B. (2010) *Signals*. Oxford University Press

Smids, J., Nyholm, S., and Berkers, H. (2020) 'Robots in the Workplace: A Threat to—or Opportunity for—Meaningful Work?' *Philosophy & Technology* 33 (3), 503–522

Smith, D.L. (2004) *Why We Lie: The Evolutionary Roots of Deception and the Unconscious Mind*. St Martin's Press

Sober, E. (1994) *From a Biological Point of View*. Cambridge University Press

Sobieszek, A. and Price, T. (2022) 'Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models'. *Minds and Machines* 32 (2), 341–364

Social Care Institute for Excellence (2022) *Defining Dignity in Care* [online] available from <https://www.scie.org.uk/dignity/care/defining> [7 February 2022]

Social Care Institute for Excellence (2013a) *Dignity in Care* [online] available from <https://www.scie.org.uk/publications/guides/guide15/> [30 April 2019]

Social Care Institute for Excellence (2013b) *Dignity in Care - Dignity for Care Workers* [online] available from

<https://www.scie.org.uk/publications/guides/guide15/careworkers/> [7 December 2018]

Social Care Institute for Excellence (2006) *Assessing the Mental Health Needs of Older People - Depression* [online] available from <https://www.scie.org.uk/publications/guides/guide03/problems/depression.asp> [27 September 2021]

Softbank Robotics (2018) *Pepper the Humanoid Robot* [online] available from <https://www.softbankrobotics.com/emea/en/pepper> [14 December 2018]

Sony (2018) *Aibo* [online] available from <https://us.aibo.com/> [14 December 2018]

Sorell, T. and Draper, H. (2014) 'Robot Carers, Ethics, and Older People'. *Ethics and Information Technology* 16 (3), 183–195

Sorensen, R. (2007) 'Bald-Faced Lies! Lying without the Intent to Deceive'. *Pacific Philosophical Quarterly* 88 (2), 251–264

Sparkes, M. (2023) 'AI Legal Assistant Will Help Defendant Fight a Speeding Case in Court'. *New Scientist* [online] 4 January. available from <https://www.newscientist.com/article/2351893-ai-legal-assistant-will-help-defendant-fight-a-speeding-case-in-court/> [12 January 2023]

Sparrow, R. (2017) 'Robots, Rape, and Representation'. *International Journal of Social Robotics* 9 (4), 465–477

Sparrow, R. (2007) 'Killer Robots'. *Journal of Applied Philosophy* 24 (1), 62–77

Sparrow, R. (2002) 'The March of the Robot Dogs'. *Ethics and Information Technology* 4, 305–318

Sparrow, R. and Sparrow, L. (2006) 'In the Hands of Machines? The Future of Aged Care'. *Minds and Machines* 16 (2), 141–161

Stanford University (2022) *The AI Index Report – Artificial Intelligence Index* [online] available from <https://aiindex.stanford.edu/report/> [27 January 2023]

Statman, D. (2000) 'Humiliation, Dignity and Self-Respect'. *Philosophical Psychology* 12, 523–540

Steady Health Men's Zone (2022) *Penis Left Exposed during Scrotal Ultrasound and I Felt Embarrassed* [online] available from <https://www.steadyhealth.com/topics/penis-left-exposed-during-scrotal-ultrasound-and-i-felt-embarrassed-nurse-didnt-cover-me-up> [22 March 2023]

Steels, L. and Brooks, R. (eds.) (2018) *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Routledge

Stephenson, J. (2020) ‘Stress and Abuse Remain High among NHS Nurses, Reveals Survey’. *Nursing Times* [online] 18 February. available from <https://www.nursingtimes.net/news/workforce/stress-and-abuse-remain-high-among-nhs-nurses-reveals-survey-18-02-2020/> [20 January 2023]

Steutel, J. (2009) ‘Towards a Sexual Ethics for Adolescence.’ *Journal of Moral Education* 38 (2), 185–198

Steutel, J. and de Ruyter, D. (2011) ‘What Should Be the Moral Aims of Compulsory Sex Education?’ *British Journal of Education Studies* 59 (1), 75–86

Stevens, M. (2016) *Nature’s Cheats: How Animals and Plants Trick and Deceive* [online] available from <http://theconversation.com/natures-cheats-how-animals-and-plants-trick-and-deceive-55323> [23 November 2022]

Stokes, F. and Palmer, A. (2020) ‘Artificial Intelligence and Robotics in Nursing: Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans’. *Nursing Philosophy* 21 (4), e12306

Strawson, P.F. (1952) *Introduction to Logical Theory*. Methuen

Strudler, A. (2009) ‘Deception and Trust’. in *The Philosophy of Deception*. Martin, C. Oxford University Press

Sue, D.W. (2010a) *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. John Wiley & Sons

Sue, D.W. (ed.) (2010b) *Microaggressions and Marginality: Manifestation, Dynamics, and Impact*. John Wiley & Sons

Sue, D.W., Capodilupo, C.M., Torino, G.C., Bucceri, J.M., Holder, A., Nadal, K.L., and Esquilin, M. (2007) ‘Racial Microaggressions in Everyday Life: Implications for Clinical Practice’. *American Psychologist* 62, 271–286

Sue, D.W., Nadal, K.L., Capodilupo, C.M., Lin, A.I., Torino, G.C., and Rivera, D.P. (2008) ‘Racial Microaggressions against Black Americans: Implications for Counseling’. *Journal of Counseling and Development* 86, 330–338

Sullins, J.P. (2011) ‘Introduction: Open Questions in Roboethics’. *Philosophy & Technology* 24 (3), 233–238

Summers, L. (2019) ‘The NHS Robots Performing Major Surgery’. *BBC News* [online] 12 December. available from <https://www.bbc.com/news/uk-scotland-50745316> [12 January 2021]

Tadd, W. (ed.) (2005) ‘Dignity and Older Europeans’. *Ethics, Law and Society* [online] 6. available from

<https://www.researchgate.net/publication/241896016_Dignity_and_Older_Europeans> [10 February 2019]

Takahashi, L. (2023) *Zoomer Kitty Whiskers The Orange Tabby Review* [online] available from <https://www.robotpetfriends.com/zoomer-kitty-whiskers-the-orange-tabby-review/> [27 January 2023]

Tamblyn, T. (2014) 'Child Sex Robots Could Be Used to Cure Paedophiles'. *HuffPost UK* [online] 17 July. available from <https://www.huffingtonpost.co.uk/2014/07/17/child-sex-robots-paedophiles_n_5594080.html> [1 November 2022]

Taneja, H. (2019) 'The Era of "Move Fast and Break Things" Is Over'. *Harvard Business Review* [online] 22 January. available from <https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over> [27 January 2023]

Tännsjö (1999) *Coercive Care: The Ethics of Choice in Health and Medicine*. Routledge

Taylor, S. (1989) *Positive Illusions: Creative Self-Deception and the Healthy Mind*. Basic Books

Taylor, S. and Brown, J. (1988) 'Illusion and Well-Being: A Social Psychological Perspective on Mental Health'. *Psychological Bulletin* 103 (2), 193–210

Telepresence Robots (2019) *What Is a Telepresence Robot?* [online] available from <https://telepresencerobots.com/what-telepresence-robot-and-what-can-they-do/> [19 December 2021]

Thompson, L. (2022) 'Defeating Drones: The Most Promising Weapons Are All Non-Kinetic.' *Forbes* [online] 1 November. available from <https://www.forbes.com/sites/lorenthompson/2022/11/01/defeating-drones-the-most-promising-weapons-are-all-non-kinetic/> [2 November 2022]

Thomson-DeVeaux, A. (2010) 'We Need to End Non-Consensual Pelvic Exams'. [26 October 2010] available from <https://our-compass.org/2010/10/26/we-need-to-end-non-consensual-pelvic-exams/> [11 March 2019]

Tilton, E.C.R. and Ichikawa, J.J. (2021) 'Not What I Agreed To: Content and Consent'. *Ethics* 132 (1), 127–154

Titcomb, J. and Sabur, R. (2018) 'Driverless Uber Car Kills Female Pedestrian in First Deadly Crash'. *The Telegraph* [online] 19 March. available from <https://www.telegraph.co.uk/technology/2018/03/19/driverless-uber-car-kills-female-pedestrian-first-deadly-autonomous/> [29 March 2019]

Tollefsen, C.O. (2014) *Lying and Christian Ethics*. New Studies in Christian Ethics 33. Cambridge University Press

Tomorrow's World (1966) *Able Mabel* [online] available from <https://www.facebook.com/BBCArchive/videos/223887922698529/> [20 January 2022]

Toshiba (2020) *Robots That Could Sense, Think and Act* [online] available from <https://www.toshiba-clip.com/en/detail/p=151> [3 November 2022]

Trinity College Dublin (2019) *Robotics Engineers Unveil 'Stevie II' – Ireland's First Socially Assistive AI Robot* [online] available from <https://www.tcd.ie/news_events/articles/robotics-engineers-unveil-stevie-ii-irelands-first-socially-assistive-ai-robot/> [27 November 2019]

Tuisku, O., Pekkarinen, S., Hennala, L., and Melkas, H. (2019) 'Robots Do Not Replace a Nurse with a Beating Heart: The Publicity around a Robotic Innovation in Elderly Care'. *Information Technology & People* 32 (1), 47–67

Turkle, S. (2017) *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books

Turkle, S., Taggart, W., Kidd, C.D., and Daste, O. (2006) 'Relational Artifacts with Children and Elders: The Complexities of Cyber-Companionship'. *Connection Science* 18 (4), 347–362

United Nations (1991) *United Nations Principles for Older Persons* [online] available from <https://www.ohchr.org/en/professionalinterest/pages/olderpersons.aspx> [28 October 2018]

United Nations (1966) *International Covenant on Civil and Political Rights* [online] available from <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx> [7 May 2019]

United Nations (1948) *Universal Declaration of Human Rights*. United Nations. available from <https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf> [1 May 2019]

US Conference of Catholic Bishops (2022) *Moral Principles Concerning Infants with Anencephaly* [online] available from <https://www.usccb.org/issues-and-action/human-life-and-dignity/end-of-life/moral-principles-concerning-infants-with-anencephaly> [30 June 2022]

Utami, D. and Bickmore, T. (2019) 'Collaborative User Responses in Multiparty Interaction with a Couples Counselor Robot'. in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*

(HRI) [online] held March 2019 at Daegu, Korea (South). 294–303. available from <https://ieeexplore.ieee.org/document/8673177/> [25 June 2019]

Vallor, S. (2011) 'Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century'. *Philosophy & Technology* 24 (3), 251–268

VanEpps, E.M. and Hart, E. (2022) 'Questions and Deception: How to Ask Better Questions and Elicit the Truth'. *Current Opinion in Psychology* 48

Vartan, S. (2020) 'Insect-Inspired Robots That Can Jump, Fly and Climb Are Almost Here'. *CNN* [online] 26 September. available from <https://edition.cnn.com/2020/09/26/world/tiny-insect-inspired-robots-scn/index.html> [9 December 2022]

Velleman, J.D. (1992) 'Against the Right to Die'. *Journal of Medicine and Philosophy* 17 (6), 665–681

Vereecken, N.J. (2009) 'Deceptive Behavior in Plants. I. Pollination by Sexual Deception in Orchids: A Host–Parasite Perspective'. in *Plant-Environment Interactions: From Sensory Plant Biology to Active Plant Behavior* [online] ed. by Baluska, F. Signaling and Communication in Plants. Springer, 203–222. available from <https://doi.org/10.1007/978-3-540-89230-4_11> [5 July 2021]

Victim Support UK (2023) 'Rape and Sexual Assault'. [2023] available from <https://www.victimsupport.org.uk/crime-info/types-crime/rape-and-sexual-assault/> [6 February 2023]

Vincent, J. and Castelfranchi, C. (1981) 'On the Art of Deception: How to Lie While Saying the Truth'. in Parret, H., Sbisá, M., and Verschueren, J. (eds.) *Proceedings of the Conference on Possibilities and Limitations of Pragmatics* [online] held 1981 at Amsterdam. John Benjamins Publishing Company, 749–777. available from <https://www.academia.edu/531423/On_the_art_of_deception_How_to_lie_while_saying_the_truth> [28 June 2021]

Vrij, A. (2000) *Detecting Lies and Deceit*. Wiley

Wada, K., Shibata, T., Saito, T., and Tanie, K. (2002) 'Robot Assisted Activity for Elderly People and Nurses at a Day Service Center'. in *Proceedings 2002 IEEE International Conference on Robotics and Automation*. held May 2002. 1416–1421 vol.2

Wade-Palmer (2021) 'Sex Robot Almost Fools Human into Thinking It's "real" in Snap of New AI Doll'. *Daily Star* [online] 23 April. available from <https://www.dailystar.co.uk/news/latest-news/sex-robot-almost-fools-human-23970794> [1 November 2022]

Wallach, W. (2010) ‘Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making’. *Ethics and Information Technology* 12 (3), 243–250

Wallach, W. and Allen, C. (2009) *Moral Machines: Teaching Robots Right from Wrong* [online] Oxford Scholarship Online. available from <https://epdf.tips/moral-machines-teaching-robots-right-from-wrongee26fe67b1731fd9274328e5bb3eec5263988.html> [24 October 2018]

Walsh, F. (2020) ‘AI “outperforms” Doctors Diagnosing Breast Cancer’. *BBC News* [online] 2 January. available from <https://www.bbc.com/news/health-50857759> [10 January 2020]

Wang, J., Leu, J., and Shoda, Y. (2011) ‘When the Seemingly Innocuous “Stings”: Racial Microaggressions and Their Emotional Consequences’. *Personality and Social Psychology Bulletin* 37 (12), 1666–1678

Waugh, R. (2015) ‘Samsung to Make Robots “Cheaper than Any Human Worker”’. *Metro* [online] 19 October. available from <https://metro.co.uk/2015/10/19/samsung-to-make-robots-cheaper-than-any-human-worker-5447772/> [29 March 2019]

Wertheimer, A. (2014) ‘(Why) Should We Require Consent to Participation in Research?’ *Journal of Law and the Biosciences* 1 (2), 137–182

Wertheimer, A. (2003) *Consent to Sexual Relations*. Press Syndicate of the University of Cambridge

Wertheimer, A. (1996) ‘Consent and Sexual Relations’. *Legal Theory* 2 (2), 89–112

Whitby, B. (2008) ‘Sometimes It’s Hard to Be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents’. *Interaction with Computers* 20 (3), 326–333

WHO (2022) *Ageing and Health* [online] available from <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health> [24 June 2022]

WHO (2015) *Dignity in Mental Health* [online] available from <https://www.who.int/southeastasia/news/detail/08-10-2015-dignity-in-mental-health> [20 March 2023]

Wikipedia (2023) ‘K9 (Doctor Who)’. in *Wikipedia* [online] available from <https://en.wikipedia.org/w/index.php?title=K9_(Doctor_Who)&oldid=1132914060> [1 February 2023]

Wilkinson, D., Herring, J., and Savulescu, J. (2020) *Medical Ethics and Law: A Curriculum for the 21st Century*. 3rd edn. Elsevier

Williams, B. (2002) *Truth and Truthfulness* [online] Princeton University Press. available from <https://beckassets.blob.core.windows.net/product/readingsample/664126/9780691117911_firstchapter.pdf>

Williams, K.N., Herman, R., Gajewski, B., and Wilson, K. (2009) 'Elderspeak Communication: Impact on Dementia Care'. *American Journal of Alzheimer's Disease & Other Dementias* 24 (1), 11–20

Williams, M.T. (2020) 'Microaggressions: Clarification, Evidence, and Impact'. *Perspectives on Psychological Science* 15 (1), 3–26

Winfield, A. (2012) *Robotics: A Very Short Introduction*. Oxford University Press

Winfield, A. (2011) 'Roboethics - for Humans'. *New Scientist* 32–33

Woebot (2019) *Woebot - Your Charming Robot Friend Who Is Here for You, 24/7* [online] available from <https://woebot.io> [18 June 2019]

Woolhead, G., Calnan, M., Dieppe, P., and Tadd, W. (2005) 'Dignity and Older Age: What Do Older People in the United Kingdom Think?' *Age and Ageing* 33 (2), 165–170

Wright, S. (2010) 'Trust and Trustworthiness'. *Philosophia* 38, 615–627

Wykowska, A., Chaminade, T., and Cheng, G. (2016) 'Embodied Artificial Agents for Understanding Human Social Cognition'. *Philosophical Transactions of the Royal Society B: Biological Sciences* [online] 371 (1693). available from <https://royalsocietypublishing.org/doi/full/10.1098/rstb.2015.0375> [3 November 2022]

van Wynsberghe, A. (2016) *Healthcare Robots: Ethics, Design and Implementation*. Routledge

X2AI (2019) *Tess: Affordable Mental Health Access With Proven Results* [online] available from <https://www.x2ai.com/> [19 June 2019]

Young, A. (2015) 'Industrial Robots Could Be 16% Less Costly To Employ Than People By 2025'. *International Business Times* [online] 11 February. available from <https://www.ibtimes.com/industrial-robots-could-be-16-less-costly-employ-people-2025-1811980> [29 March 2019]

Zaraki, A., Wood, L., Novanda, O., Robins, B., and Dautenhahn, K. (2017) 'Toward Autonomous Child-Robot Interaction: Development of an Interactive Architecture for the Humanoid Kaspar Robot'. in *Proceedings of the Child-Robot Interaction Workshop*, 'HRI 2017' [online] held 2017 at Vienna, Austria. available from <https://uhra.herts.ac.uk/bitstream/handle/2299/23856/abolfazl_zaraki.pdf?sequence=1&isAllowed=y> [11 March 2022]

Zhang, Z. and Bickmore, T. (2018) ‘Medical Shared Decision Making with a Virtual Agent’. in *Proceedings of the 18th International Conference on Intelligent Virtual Agents* [online] held 2018 at Sydney, NSW, Australia. ACM Press, 113–118. available from <http://dl.acm.org/citation.cfm?doid=3267851.3267883> [25 June 2019]

Zuolo, F. (2016) ‘Dignity and Animals. Does It Make Sense to Apply the Concept of Dignity to All Sentient Beings?’ *Ethical Theory and Moral Practice* 19 (5), 1117–1130

Appendix A

Index of robots and related technologies

For copyright reasons, images in this section have been removed from this publicly-available version of the thesis. I describe robots' appearances here, but it may also be useful for readers to look for images of the different robots.

In these descriptions, sizes generally refer to height: 'large' is anything above about 140cm; 'medium' is anything between about 80cm and 140cm; 'small' is anything between about 30cm and 80cm; 'miniature' is anything under about 30cm.

Able Mabel

Hypothetical robot shown on TV series *Tomorrow's World* (1966)

Function: Household labour

Appearance: Large size, white, box-like, with basic and very loosely humanlike 'head' and arms

Aibo

Real-world robot made by Sony (2018)

Function: Recreational

Appearance: Miniature size, dog-shaped; sleek white plastic shell; has a basic, loosely dog-like head and face

Alexa (Echo)

Real-world robot made by Amazon (2023)

Function: Assistive / informational

Appearance: Miniature black tabletop cylinder, sphere, or cube; no moving parts

Asimo

Real-world robot made by Honda (2019a)

Function: Social

Appearance: Large size, humanoid shape, sleek white plastic shell, resembles a human in a space suit, but no face

Babylon Health AI Doctor

Real-world AI software made by Babylon Health (2018)

Function: Medical diagnosis

Appearance: Text messages which appear on phone or computer screen

BigDog

Real-world robot made by Boston Dynamics (Robots 2022)

Function: Military (carrying)

Appearance: Medium size, grey metal, four legs, ‘body’ has visible circuitry; no head

Buddy

Real-world robot made by Blue Frog (2022)

Function: Social

Appearance: Small size, very loosely humanoid, sleek white plastic shell, no arms or legs, basic face displayed on screen

C-3PO

Fictional robot from the *Star Wars* movies and universe (Fandom 2022b)

Function: Social – sentient, emotional, and outperforms humans in many ways

Appearance: Large size, humanoid shape; resembles human in gold suit of armour; basic face

CareBot

Real-world robot made by Gecko Systems (2008)

Function: Care (general)

Appearance: Medium size, sleek white plastic shell, vertical shape, no arms or legs, rectangular black screen for ‘head’

Care-o-bot

Real-world robot made by Fraunhofer-Gesellschaft (2018a)

Function: Care (general)

Appearance: Medium size, very loosely humanoid, sleek white plastic shell, vertical shape, two moving arms with hands, circular screen can show face

ChatGPT

Real-world AI software (large language model) made by Open AI (2023)

Function: Assistive, social

Appearance: Text which appears on phone or computer screen

CleanseBot

Real-world robot made by Superclean (Impressive Things 2018)

Function: Cleaning (beds)

Appearance: Miniature, sleek white plastic ovoid

Cody

Real-world robot made by Healthcare Robotics (Quick 2010)

Function: Care (bathing)

Appearance: Large size, mechanical-looking, does not resemble human, but has ‘arms’

Companion Pets

Real-world robots made by Ageless Innovation (2018)

Function: Recreational / emotional

Appearance: Cat, dog, and bird robots, the size of the animal they represent (dogs are small size); fairly realistic-looking, with fur, faces, and legs

Data

Fictional robot from the TV series *Star Trek: The Next Generation* (CBS Studios 2022)

Function: Social and military – sentient, and outperforms humans in many ways

Appearance: Large size, looks absolutely human, but has grey ‘skin’

Driverless cars

Real-world robots made by a variety of companies

Function: Transport

Appearance: Size and shape of normal human-driven cars, with additional sensors

Drones

Real-world robots made by a variety of companies and organisations

Function: Video recording / surveillance

Appearance: Small or miniature, mechanical-looking, with multiple helicopter-like blades for flying

Note that the term ‘drone’ also refers to non-robotic flying devices remotely controlled by human operators

El-E

Real-world robot made by Georgia Tech (Park 2016)

Function: Care (carrying items)

Appearance: Medium size, mechanical-looking, moving arm

Geminoid

Real-world non-robotic (human-controlled) technology made by Hiroshi Ishiguro Laboratories (2018)

Function: Recreational

Appearance: Large size, looks absolutely human

HAL-9000

Fictional robot from the book and movie *2001: A Space Odyssey* (Fandom 2022a)

Function: Controls spaceship – sentient

Appearance: Black rectangle with red spot in the centre, built into spaceship

Holly

Fictional robot from the TV series *Red Dwarf* (Fandom 2023a)

Function: Controls spaceship – sentient

Appearance: Disembodied life-size human face on a black screen; face looks absolutely human

Hybrid Assistive Limb (HAL)

Real-world robot made by Cyberdyne (2019)

Function: Walking assistance (for people who have lost the use of their legs)

Appearance: Exoskeleton worn on human legs, made of sleek white plastic

Industrial robots

Real-world robots made by a variety of companies

Function: Assembly line, warehousing

Appearance: Mechanical-looking with moving arms, come in a variety of sizes, all very unhumanlike

iPal

Real-world robot made by AvatarMind (2017)

Function: Social

Appearance: Medium size, humanoid, sleek white plastic shell, moveable arms and non-moving legs, head with basic eyes

Jibo

Real-world robot made by NTT Disruption (2014)

Function: Social

Appearance: Miniature size tabletop device resembling a lamp, with sleek white plastic shell and black ring

K9

Fictional robot from the TV series *Dr Who* (Wikipedia 2023)

Function: Assistive / informational – highly intelligent, sentient

Appearance: Medium size, grey metal, resembles upturned rubbish bin, very loosely dog-like head

Mia

Fictional robot from the TV series *Humans* (Arnold et al. 2015)

Function: Social – sentient, outperforms humans in some ways

Appearance: Large size, looks absolutely human

Military robots

Real-world robots made by a variety of companies and organisations

Function: Military

Appearance: Come in various shapes and sizes, such as medium size tanks, and small mechanical-looking missiles, typically made from grey metal

MySpoon

Real-world robot made by Secom (2019)

Function: Care (feeding)

Appearance: Small tabletop device made with sleek white plastic shell, with moveable arm holding a spoon

Nest

Real-world robot made by Google (2023)

Function: Assistive

Appearance: Miniature ovoid, available in various colours; no moving parts

Number 5

Fictional robot from the movie *Short Circuit* (Badham 1986)

Function: Military – sentient

Appearance: Large size, upright, very loosely humanoid, mechanical-looking, grey metal with visible circuitry, moveable arms and ‘eyes’ on the ‘head’

Olly

Real-world robot made by Emotech (2020)

Function: Social

Appearance: Miniature tabletop device, black plastic, ring-shaped

Omnicell

Real-world robot made by Omnicell (2023)

Function: Care (medication dispensary)

Appearance: Large size, sleek white plastic shell, resembles a vending machine or filing cabinet

PaPeRo

Real-world robot made by NEC (Robot Center Ltd 2020)

Function: Care / recreational

Appearance: Small size, very loosely humanoid, sleek white plastic shell, basic head, basic eyes, no arms or legs

Paro

Real-world robot made my Sense Medical (2022)

Function: Recreational / emotional (for people with dementia)

Appearance: Small size, looks like a baby seal, white fur, moving limbs and face

Pearl

Real-world robot made by researchers from the University of Michigan, University of Pittsburgh, and Carnegie Mellon University (Robotics Today 2021)

Function: Care (general)

Appearance: Large size, loosely humanoid, mechanical-looking with visible circuitry, basic head with moving eyes and lips, has arms but no legs

Pepper

Real-world robot made by Softbank (2018)

Function: Social / assistive

Appearance: Medium size, humanoid, sleek white plastic shell, moving arms with grasping hands, head with basic face

PR2

Real-world robot made by Georgia Tech (ScienceDaily 2018)

Function: Care (dressing patients)

Appearance: Medium size, mechanical looking, sleek white plastic shell, moving arm

Pris

Fictional robot from the movie *Blade Runner* (Scott 1982)

Function: Worker – sentient, performs similar to humans

Appearance: Large size, looks absolutely human

QTrobot

Real-world robot made by LuxAI (2022)

Function: Social / recreational

Appearance: Small size, humanoid, sleek white plastic shell, moving arms and non-moving legs, basic face on screen on head

R2-D2

Fictional robot from the *Star Wars* movies and universe (Fandom 2023b)

Function: Repairing machinery – sentient but non-verbal

Appearance: Medium size, cylindrical with rounded top, sleek white plastic shell, short legs

Replika

Real-world AI software made by Luca Inc (2020)

Function: Social / care (mental health)

Appearance: Text messages which appear on phone or computer screen with humanlike avatar

Riba

Real-world robot made by Riken (2019)

Function: Care (lifting patients)

Appearance: Medium size, loosely humanoid, sleek white plastic shell, moving arms, head like a plastic teddy bear

Roomba

Real-world robot made by iRobot (2022)

Function: Cleaning (floors)

Appearance: Miniature, black, flat cylinder

Sexbots

Real-world robots made by a variety of companies and individuals

Function: Sex

Appearance: Large size, looks very humanlike

Siri (HomePod)

Real-world robot made by Apple (2023)

Function: Assistive

Appearance: Miniature black sphere or cylinder, no moving parts

Sonny

Fictional robot from the movie *I, Robot* (Fandom 2023c)

Function: Worker – sentient, outperforms humans in some ways

Appearance: Large size, humanoid body, grey body, loosely humanlike grey face

Sophia

Real-world robot made by Hanson Robotics (2022)

Function: Social

Appearance: Large size, humanoid, highly humanlike face, but visible circuitry in head and arms

Spot

Real-world robot made my Boston Dynamics (2021)

Function: Military (carrying)

Appearance: Medium size, four legs, sleek yellow plastic shell, no head or tail

Stevie

Real-world robot made by researchers at Trinity College Dublin (2019)

Function: Care / social

Appearance: Medium size, loosely humanoid, sleek white plastic shell, arms, square head with basic face

Stride Management Assist

Real-world technology made by Honda (2019b)

Function: Walking assistance (for patients who are weak, but able to walk)

Appearance: Exoskeleton worm on human upper legs, made of black plastic

Tay

Real-world AI software made by Microsoft (2016)

Function: Social

Appearance: Text messages which appear on phone or computer screen, with humanlike picture

Terminator

Fictional robot from the *Terminator* movies (Cameron 1984)

Function: Killing people – sentient, strong, outperforms humans

Appearance: Large size, looks absolutely human

Woebot

Real-world AI software made by Woebot Health (2019)

Function: Care (mental health)

Appearance: Text messages which appear on phone or computer screen

Zoomer Kitty

Real-world robot made by Spinmaster (Takahashi 2023)

Function: Recreation

Appearance: Cat-shaped but miniature size, black plastic, loosely cat-like head, four legs with wheels