Chapter 13
Social Robots: Principles of Interaction Design and User Studies

Janie Busby Grant and Damith Herath

13.1 Learning Objectives

This chapter introduces you to the basic steps in designing and conducting social robotics research. By the end of this chapter you will:

1. Be able to describe why you are conducting your research project, including your motivation for conducting the research, who the audience is for your findings, and the key research questions you will be addressing.
2. Be able to identify the key variables you want to focus on and understand how to operationalise these variables in real research environments.
3. Be able to recognise different types of research designs, know advantages and disadvantages of each, and work out which is appropriate in which situation.
4. Be aware of the concepts of validity and reliability, and be able to both identify and address issues that can emerge.
5. Understand the key principles to consider when designing and conducting ethical research.
6. Identify key factors when analysing and interpreting data.

13.2 Introduction

This chapter considers when, why and how you can conduct user-focused research when working with robots. While exciting, high-quality human–robot interaction

J. Busby Grant (✉)
Discipline of Psychology, Faculty of Health, University of Canberra, Academic Fellow, Graduate Research, Canberra, Australia
e-mail: janie.busbygrant@canberra.edu.au

D. Herath
Collaborative Robotics Lab, University of Canberra, Canberra, Australia
e-mail: Damith.Herath@Canberra.edu.au

D. Herath and D. St-Onge (eds.), *Foundations of Robotics*,
https://doi.org/10.1007/978-981-19-1983-1_13

research is carried out in laboratories and real-life settings around the world, often those working in the field of robotics feel unsure or unprepared to conduct user-focused research, as research design and analysis is typically not included in undergraduate robotics courses. Conducting well-designed research studies can allow you to identify potential issues, benefits and unexpected outcomes of real-world interactions between robots and users, and provide evidence for efficacy and impact. Being able to confidently and competently design and conduct user studies with robots will be an advantage in a myriad of workplace roles.

This chapter is designed to be an introduction and practical guide to conducting human–robot interaction research. In this chapter, we consider why you should conduct research examining human–robot interaction, how you can identify the best research design to answer your question, and select and measure appropriate variables. Throughout the chapter we provide examples of relevant real-world research projects. While we can necessarily only "scratch the surface" of each topic we discuss in this single chapter, and you are encouraged to explore deeper into the issues of relevance to you, the tools in this chapter will allow you to understand and implement well-designed research projects in human–robot interaction.

An Industry Perspective

Martin Leroux, Field Application Engineer

Kinova inc.

I was doing my bachelor's degree in engineering physics, but I didn't quite like it because I found it too abstract. At the time, I happened to find an internship for a robotics lab which took me in not so much for my then non-existent background in robotics, but rather for the advanced math skills I developed in physics. That internship was the part that really clicked for me; I finally was able to explain to my family what it is that I do in terms they could understand. I came to the field for that feeling of satisfaction, but I ended up staying because it is so wide and I get to learn new things constantly.

When I first started at Kinova, my job was to evaluate alternative control input methods to help our assistive users manipulate our robots. Once, I started working on eye-tracking and a colleague kept insisting that I let one of our users try it as soon as possible, which I found was too early, especially since I also have eyes. When I asked why he was so insistent, he told me that a while ago, their team spent multiple weeks tuning a program to help assistive users drink from a bottle - adjusting the rotation of the arm, the lift speed, and so on and so forth. Then, when they were all done and went to show it to a user, his reaction was: "Oh, I'd never bother with that. I just use a straw." Weeks of work went down the drain. Now, we always involve our end users right from the get-go.

When I first entered the field, human-robot interfaces were already very varied and fairly functional. However, their purpose was to specifically control the robot in terms that only made sense for robots. You would move individual joints or sometimes switch between translation and orientation for end-effector cartesian control. Nowadays, although the hardware hasn't evolved much, the interface itself often leverages artificial intelligence to make the entire system much smarter. Instead of asking users to think like roboticists, they can keep thinking like human beings with task-oriented commands. Where people used to need to think "I want to get my gripper there", now they can think "I want to grab this" and the robot can deduce some or all of the motion that is expected of it.

13.3 Cobots, Social Robots and Human–Robot Interaction

Considered the father of Robotics, Engineer and Entrepreneur, Joseph Engelberger, after commercialising the first industrial robot arm, stipulated in his book "Robotics in Service" that the main thrust of the industry should be towards the development of what are called service robots. He argued that technological developments in robot perception and artificial intelligence should enable the replacement of human work that is labour intensive and unpalatable with robotic devices. In fact, a considerable portion of such mundane human work is now automated using robotic technologies. In the process, robots have come to be established in increasingly human spaces. This necessitates the designers of these robotic devices to consider such elements as perceived safety, human factors and ergonomics on top of the engineering capabilities of the robots. The modern cousins of the original Unimate (like the Kinova Gen3 robots) embody such nascent technologies, providing the ability to directly interact and work alongside humans in a safe and intelligent manner. A new category

of industrial robots is emerging called *collaborative robots* or cobots. These technologies are considered to belong to the fourth industrial revolution (industry 4.0) which represents the evolution of traditional industrial manufacturing technologies and practices combining with emerging smart technologies such as the Internet of Things (IoT), cloud computing and artificial intelligence (AI).

On the other hand, the origins of the word "robotics" and as we see in popular culture, robots are meant to be a reflection of the humanity. As we discussed earlier, humans have entertained the idea of human-like machines for millennia. With recent advances in related domains in computing and hardware, there is growing interest in the engineering community to explore the technical development of socially intelligent robots. We have already seen several commercial *artificially socially intelligent* robots appearing in the market with various success. This new genre of machines called "*social robots*" are meant to interact with humans at a higher cognitive and emotional level, as compared to a typical industrial robot in a factory.

In either scenario, robots are increasingly required to interact with humans, and importantly with end users who are not technically trained to operate robots (such as is currently the case with industrial era robots). As we have already seen in Chap. 3, empathetic thinking is required when designing such robots with lay end-users in mind. And at the other end of the pipeline, before deploying these robots, it is required that we not only validate the robot's engineering functions but also its ability to interact with humans as intended. The latter requires an understanding of human psychology and associated disciplinary expertise. The study of human–robot interaction is thus an emergent field that not only encompasses many fields of engineering, but yields a broader disciplinary net towards psychology, sociology, design and the humanities. Abbreviated, HRI, the new field is gaining considerable attention from the robotics community with several key journals and conferences dedicated to the subject already, including the ACM/IEEE International Conference on Human–Robot Interaction and the International Conference on Social Robotics and their respective journals, ACM Transactions on Human–Robot Interaction and the International Journal of Social Robotics.

13.4 Why Conduct Research?

Any research project should begin with the "why?". Why do you want or need to conduct the research, what questions do you want to answer, what will the findings allow you to know or do, and who will be interested in the answers you generate?

13.4.1 Motivation for the Research

Your reasons for conducting the research may be diverse and could include:

- You want to know how people interact with the robot in order to improve it.
- You are not sure how the robot will perform in real-world settings.
- You need evidence to give to investors that your product will be successful.
- You are conducting a project in a university or research-focused setting.
- Your boss/supervisor/funding partner said so!

All of these and many more are legitimate reasons for conducting research, but how you design your project will depend on your "why", which determines your research question and your audience. For instance, you might have questions about: efficacy (how well the robot performs in controlled conditions), effectiveness (how successful the robot is in the real world), safety (both technical and perceived by the user), or perceptions of the people interacting with them (such as ratings of likeability or animacy).

Research Examples: Why and Audience

Sylax: A university-based research robotics project.

– Why: The researchers want to understand what factors play a role in converting an industrial robot into an interactive user-friendly robot.
– Audience: The researchers themselves, the wider academic field.

Tommy: A robot product designed by a start-up for commercialisation.

– Why: The researchers want to design a conversational agent for use by the general population.
– Audience: Initially the researchers, eventually the general population.

Coramand: A tech company robotics product for large-scale industry-based roll-out.

– Why: The researchers want to assess and improve perceived and operational safety of their collaborative robot.
– Audience: Initially the development company itself, then roll-out to companies that use automation at scale (individuals in those companies: executive, skilled technologists, investors).

13.4.2 Target Audience

Who is the intended target of the outcomes of the research? Remember that there may be several uses for, and audiences of your output from a particular piece of research.

Typically, the research you conduct will inherently inform your own knowledge, so the first audience is generally **yourself**. Carefully consider what you need to

know in order to move the project forward—what information will directly inform the next stage in your project? All too often researchers get carried away with their own cleverness and design complex experiments, only to find that they have tweaked their design to the point that their answers don't quite tell them what they need to know anymore! Always come back to your key research question and what answering it will allow you to conclude.

If you are a student or university-based researcher your target audience will also be **others in your academic field**. You should be familiar with the key literature, and in a fast-moving field like robotics, this also means attending the main conferences, reading new abstracts for related projects, taking part in competitions, new business ventures and start-ups. There may be typical research paradigms in your area that you are expected to use, or controversies or debates that you need to be aware of, or terminology or technological basics you should adopt to better communicate with those in your field. This applies both in terms of the design of the research itself and how you report and distribute your findings.

You may also have an audience in the form of **investors**, whether this be venture capital firms, government funding bodies or individuals. If this funding is currently supporting your research, be clear on what outcomes and reporting you have already committed to, and in what timeframes, as you may well need to set up research to directly address those requirements. If you are designing research to obtain future funding, look at the previous projects that the organisation or individual has financed, and consider what kind of evidence they provided. For instance, were they interested in projects with lab-based proof-of-concept completed, or data on projected uptake from the general public, or testing conducted in unpredictable real-world environments? Knowing the answers to these questions will help frame your own research.

Enterprise may also be an audience for your work, for example companies that use large-scale automation in industry settings. This relationship may take the form of an existing contract or potential sales. You need to be very clear on what factors will be most important to the key individuals in that organisation so that you can focus your research to demonstrate capability in those areas. You also need to identify what forms of evidence will be most convincing to that audience—are they looking for large lab safety trials, or real-world human–robot error rates, or expert review?

Consumers and/or the general public may also be the target audience for your research, particularly if you are demonstrating efficacy in a real-world environment or interest in solving a particular problem. In this case you need to think about what the target consumer would find persuasive and meaningful, and incorporate those elements into your study. For example, if safety is a concern among the general population, then part of your research should investigate the safety of the robot in general use settings so you can make an evidence-based statement about safety at the conclusion of your project.

13.4.3 Research Questions

The Scientific Method is the process used by researchers to create an accurate representation of the world. By working collaboratively, building on and sharing evidence and theories, we can assemble an understanding of how things work. The Scientific Method involves iteratively generating theories and hypotheses, gathering data and analysing that data to draw conclusions, which then feed back into our theories about the world, which then continues the process, refining our knowledge as it continues (Fig. 13.1).

In HRI research typically you begin a research project with theory about how you believe a particular interaction between a robot and a participant will play out. This may be on the basis of previous research and theory in the field, or based on observations you have personally made. A theory is a set of explanatory ideas that integrate a variety of evidence. Theories pull together facts into a general principle or set of principles—they not only help us to explain *what* is happening, but also *why* it is happening. Theories also allow us to make predictions about what will happen in a new situation. As a multidisciplinary research area, theories in HRI could be informed by many and at times conflicting ideologies. Some examples are theory of mind from psychology, perceptual control theory from cybernetics or more speculative ideas such as the Uncanny Valley effect in aesthetics.

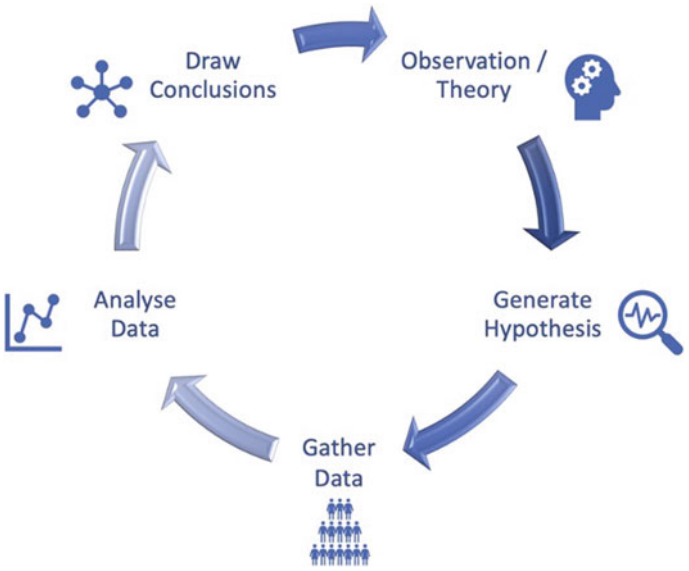

Fig. 13.1 A visual depiction of the scientific method used to conduct research

Within a particular research study, you will generate a specific hypothesis, which is typically based on a theory. A hypothesis is a specific statement about the relationship between variables in your study—it is what you expect to happen. Crucially, hypotheses are testable—that is, the findings of your study either support the hypothesis or contradict it. If a hypothesis (or a theory) is not supported, this is called falsification. The hypothesis is how you narrow down your broader theory to assess one particular effect or relationship in your research study.

Research Examples: Theories and Hypotheses

Sylax: A university-based research robotics project.

- Broad research question: What is the effect of factors like behaviour and appearance of the robot on peoples' perception of agency?
- Theory: Theory of mind (Leslie, 1987)
- Hypothesis (for one particular study): People will perceive the robot to have more agency when it shows purposeful movements rather than when it displays random movements.

Tommy: A robot product designed by a start-up for commercialisation.

- Broad research question: How can we maximise consumer satisfaction with our companion robot product?
- Theory: Attachment theory (Bretherton, 1985).
- Hypothesis (for one particular study): When given a robot to use in the home, people who have greater previous exposure to robots will report higher attachment to the product.

Coramand: A tech company robotics product for large-scale industry-based roll-out.

- Broad research question: How can perceived safety and operational safety of the robot be maximised?
- Theory: Behavioural Decision Theory (Slovic, et al., 1984)
- Hypothesis (for one particular study): People's perceived safety of the robot will be greater when the robot displays fluid movements rather than more robotic movements.

What You Should Know Going into a Research Project

I am conducting this research study because:

_____

The audience/s for my research are:

_____

The key theory or theories relevant to my study is:

_____

My main hypothesis for the study is:

_____

In summary, before you begin your study consider your motivation for conducting the study, your target audience/s and the theories that are relevant to your study. Use these to determine your hypothesis. Remember it needs to be *specific* to your study and *testable*. For example, predicting that robots with faces will be more acceptable to the general public than robots without faces is a theory—whereas making the statement that in your study the robot with a face will have higher likeability ratings than the same robot with the face obscured is an hypothesis. The hypothesis refers directly to what you are *manipulating* and *measuring* in your study, and we'll explore this in more detail below.

13.5 Deciding on Your Research Variables

13.5.1 Variables

When conducting research, one of the major tasks is to decide what you are going to manipulate and what you are going to measure. This obviously depends on your research question. For example, if you want to investigate perceived safety in the home setting, this will be measured very differently compared to a research question is focused on operational safety in an industrial setting. Assessing the likeability of a robot by young children will use a different approach from measuring the sense of agency attributed to a robot by an adult.

A variable is a characteristic which can be measured or changed. Age, object preference, experimental condition, reaction time and performance are all variables. Some of these variables are inherent to an individual and cannot be assigned (e.g. age), others change depending on the task (e.g. performance) and others can be manipulated by a researcher (e.g. exposure to different conditions). Deciding what variables you want to measure/manipulate is one of the key issues in research design and will dictate what conclusions you are able to draw.

A helpful distinction is between Independent Variables (IVs)and Dependent Variables (DVs).

- The IV is the variable you believe has an effect on the other variable/s. In some studies, the IV is manipulated (changed) by the researcher. In robotics research, this is often some aspect of the robotic system, for example your IV might be whether the robot has an anthropomorphic face or not.

- Within an IV you often have conditions (also called *levels* or *groups*)—you "do different things" to the different groups. There can be any number of conditions. Sometimes there are clear experimentaland control conditions, such that the experimental condition is the group receiving the treatment, and the control condition is the comparison group (or "usual" group). For instance, if your IV is the type of robotic face, you may have two conditions—one in which the participants view the default robot face (control condition) and one in which they view a new version of the face (experimental condition).
- The DV is the variable which is measured (observed) by the researcher. Often there are many DVs in a single study—for example you may want to measure participants' ratings of likeability, animacy and safety.
- *Note*: In some research designs, such as descriptive research, all the variables are simply measured and the researcher doesn't believe one has an effect on another. In this case, all can be considered DVs.

Research Examples: IVs and DVs

Sylax: A university-based research robotics project.

– IV: Behaviour
  - Condition 1: Random behaviour
  - Condition 2: Purposeful movement—algorithmic behaviour
– DV: Perceived agency measured by the number of interactions initiated by the participant

Tommy: A robot product designed by a start-up for commercialisation.

– IV: Reported previous exposure to robots based on a questionnaire response scale.
– DV: Attachment measured using a questionnaire response scale.

Coramand: A tech company robotics product for large-scale industry-based roll-out.

– IV: Motion type
  - Condition 1: Fluid motion
  - Condition 2: Robotic motion
– DV: Perceived safety measured as how far away people stand from the robot (in m).

13.5.2 Operationalisation

So you know what you want to investigate and why, you've identified your key theory and hypothesis, you've decided on your IV and your DV—surely you're ready to go, right? Well, not quite yet! The next key step is to operationalise your variables—which involves defining what you are manipulating and measuring, and how. For the purposes of the research project you are conducting, you are deciding very specifically how your variables will be changed or measured. This might involve a survey question or set of questions combined into a single value, or a measure of reaction time, or performance on a task.

Let's say your DV is safety—how safe people feel interacting with a particular robotic construct. There are many different ways you could operationalise this variable, for instance you could use a self-reported survey question (How safe did you feel during this interaction? 5 Very Safe to 1 Not Safe at All), or you could assess how close participants stand to the robot (with people standing closer indicating a higher level of safety), or you could measure participants' heart rate or cortisol level to indicate their level of stress during the interaction. All of these different ways of operationalising safety, and many others(!), could be appropriate depending on your "why", your research question and your audience. However, operationalisation is often driven by logistical and contextual issues as well as deeper theoretical approaches. For instance, what equipment do you have? What expertise do you and the other researchers have? What time do you have to collect the data? Are your potential participants able and willing to be measured in this way?

How you operationalise your variable will have fundamental implications for the conclusions that can be drawn from your study. Also keep in mind that how you operationalise the variable directly impacts what statistical tests you can run during analysis (see later discussion), for instance whether you use a choice task, a yes/no scale, a categorical scale or a continuous number, may require different statistical analyses.

13.5.3 Relevance-Sensitivity Trade-Off

What we're looking for in a variable is a measurement that is sensitive enough to show a difference that you're interested in (i.e. don't measure something that is unlikely to change). But you also need to be careful not to pick a variable that is so specific that it isn't relevant beyond the study to other contexts (i.e. don't measure something that isn't meaningful in the wider world). This is called the relevance-sensitivity trade-off. This can be a real issue in lab-based experiments, where the focus is on finding a way of measuring the variable that is easy in that environment, rather than operationalising the variable in a way that is more relevant to the real-world environment (Fig. 13.2).

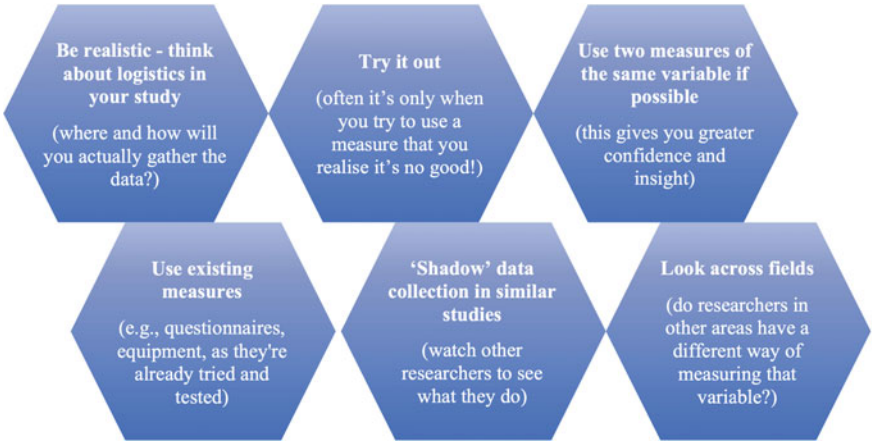

Fig. 13.2 Approaches that can help you decide how to operationalise variables

Examples of Relevance-sensitivity trade-off

Take the example of *Tommy*, the companion robot aimed at the general population. In a carefully controlled experiment you could ask participants how much they like the robot on a scale from 1 to 100. Let's say you added new capabilities and likeability scores went up from an average of 50 to an average of 60. That sounds big! But does this actually mean more people will now purchase the product? Not necessarily. The attachment scale is *sensitive* enough to pick up changes based on the improvements you made, but is that change in score *relevant* if what you really want to know if whether people will buy it or not?

13.5.4 Research Designs

There are countless ways research can be conducted, and different ways these approaches can be grouped together. One common categorisation of research designs is into descriptive, correlational, experimental research, and reviews and meta-analyses (Fig. 13.3).

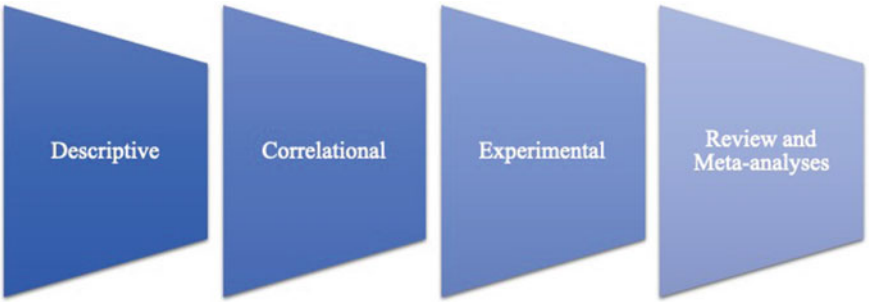

Fig. 13.3 One categorisation of the different types of research designs

13.5.5 Descriptive Research

When you conduct descriptive research you don't manipulate any variables—instead you take advantage of the natural flow of behaviour to measure what you are interested in. Some of these approaches are highly flexible, so if unexpected events happen in your research, or you develop new ideas, you can alter how you collect your data to take advantage of this. Some examples of descriptive research are observation (where you record the behaviours of your participants without interfering—e.g. examining video footage of home use of a robot), archival research (using existing data—e.g. studying web browser search history), and program evaluation (e.g. analysing outcomes of a robotic system embedded into a workplace). Focus groups are also a way of conducting descriptive research in which you sit down with small groups of participants and ask them to give their opinions and describe their behaviours. Asking descriptive information from large groups of participants is often done using surveys. In the case of the Coramand robot example used above, the company itself may want to conduct descriptive research in the form of focus groups, to get an understanding of how its employees feel about incorporating robots on the factory floor, before they commit to implementing them.

One common descriptive research method used in robotics is case studies. Case studies focus on the behaviour of a single individual or single context. They are useful in that they provide very rich information about a particular experience, so all the nuances of that person's situation, behaviour and cognition can be explored. They are also sometimes the only option, where the situation or experience or context is so unusual that other research cannot be used, such as discussed in Design Chap. 3. However, it is difficult to generalise from case studies (will this same pattern of behaviour or outcomes be seen in other contexts or individuals?) and you cannot be sure what caused any changes in the individual's behaviour—it may not have been the variable you are focused on, but could have been something else in their environment or unique to them.

13.5.6 Correlational Research

Correlational research is research which asks whether there is a relationship between two or more variables, but where the variables are not under the researcher's direct control (for logistical or ethical reasons). Often survey research falls into this category—although it can be purely descriptive (see above), sometimes surveys are used to look at whether two variables "go together", such as assessing whether age is related to perception of how dangerous robots are. Surveys are useful when you want to look at naturally occurring patterns in the world, and are relatively easy to conduct. They also allow you to measure a lot of variables at the same time. The Tommy example discussed above, in which the researchers wanted to know the relationship between previous exposure to robots and attachment to the robot product, is an example of a survey design; the researchers would conduct a correlational analysis to look at whether those two variables "go together". However, correlational research like this can only tell you about the correlation between two variables, not whether one variable causes changes in another variable. For that, you need to conduct an experiment!

13.5.7 Experimental Research

Experimental research studies the effect of an independent variable on a dependent variable. So the Sylax case discussed above, which looked at the effect of behaviour type (the IV: random or purposeful) on perceived agency (the DV) would be an experiment. Often when you manipulate an IV you measure a whole range of DVs (so for example likeability, animacy, safety). Experiments are normally conducted in highly controlled lab-based settings, but can also be conducted in natural environments—these are known as field experiments.

In an experiment, you typically want to find out if a particular manipulation (e.g. adding a face to a robot) has an effect. Let's say you want to compare a condition where you do that manipulation with one where you don't. There are several ways of doing this:

- A control condition is a group where there is no treatment or manipulation of the IV.
- A placebo condition is a group which receives what looks like the treatment/manipulation but *it is not*. It is a "look-a-like" treatment without the active component/ingredient.

13.5.8 Between-Subjects and Within-Subjects Designs

Consider the Sylax research discussed above, in which a researcher has an IV which is the type of behaviour, and measures perceived agency. In this study, the researcher has two conditions—one in which the participants are exposed to random behaviour (control condition) and one in which they are exposed to purposeful behaviour (experimental condition). The researcher then has a choice:

- They could recruit 40 participants and assign 20 to the control condition (random behaviour) and 20 to the experimental condition (purposeful behaviour), and compare the scores of the two groups.
- They could recruit 20 participants and have those participants complete both the control and experimental conditions at different times, and compare the scores of the participants on the two conditions.

In one of these cases, different people experience each of the different conditions; in the other, the same people complete both conditions. Both of these options are legitimate under particular circumstances, but they each have benefits and possible drawbacks you should be aware of when designing your study.

In a within-subjects design participants are assigned to all of the conditions of the IV, so the experimental manipulation takes place within subjects. If you had three conditions in your IV, for example, the participants complete all three conditions (they could do these sequentially in the same testing session, or take part in your study on three different days). You then compare the scores of the *same participants* across the three different conditions.

In a between-subjects design participants are only assigned to one of the conditions of the IV, so the experimental manipulation takes place between subjects. If you had three conditions in your IV, then you would have three groups of participants, and each group would complete a different condition (no individual would participate in all three conditions). You would then compare the scores of the three *different groups* of people (Fig. 13.4).

Both within- and between-subjects designs can be appropriate depending on the situation. Often, the nature of the study will determine this for you. For example, if you

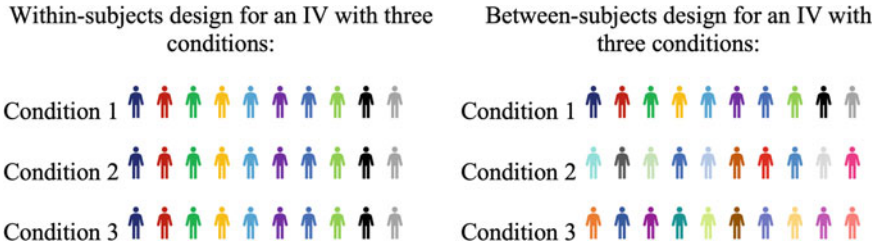

Fig. 13.4 Illustration of the differences between within-subjects and between-subjects research designs, with different individuals represented as different coloured symbols

	Pros	Cons
Within-subjects designs	• Fewer participants required. • Greater statistical power. • Less random variance.	• Exposure to one condition can affect responses to another. • Can be more difficult to set up.
Between-subjects designs	• No learning or interference across conditions. • Can be easier logistically.	• More participants required. • Lower statistical power. • More random variance.

Fig. 13.5 Pros and cons of within-subjects and between-subjects research designs

are testing whether there is a difference in learning between two interfaces, you may be unable to use a within-subjects design as learning on one will transfer to learning on another—performance on the second interface they see may be better than on the first, regardless of the interface itself. So, if exposure to one condition can potentially affect responses in other conditions, it may not be appropriate to use a within-subjects design. Similarly, if participants are only available for a single testing session, and taking part in the conditions involves a high intensity task, participants may be too fatigued to complete more than one condition. If you are using a within-subjects design one way of addressing potential ordering issues is to use counterbalancing, which involves presenting the conditions in alternating (or random) order to the participants, so that half of participants experience condition 1 first, and the other half condition 2 (and etc. if any additional conditions).

The organisation of elements such as these, and potential time constraints on participant involvement, means within-subjects designs can be more logistically difficult than between-subjects designs. However, they are statistically more "powerful" (see later discussion of power) in that this design reduces the random variance in the data collected, and means you are likely to find a significant effect if one is there than using between-subjects designs. Within-subjects designs can also be more efficient and potentially cheaper to run, as fewer individual participants are required compared with between-subjects designs (Fig. 13.5).

13.5.9 Random Assignment

If you are conducting a study in which participants are allocated to one condition of an IV only (i.e. a between-subjects design), you need to decide how to assign participants to a particular condition. That is, when a participant comes through the door, which condition are they exposed to? You could decide based on a whole range of factors—for example alternating order of arrival, surname, day of the week. Often in robotics research this is determined by technical factors—for example if

the condition takes a long time to set up, the first six weeks of data collection for a study could involve participants being assigned to Condition A, and the second six weeks to Condition B.

However, if at all possible, use random assignment to allocate participants to conditions. That is, ensure participants are randomly allocated to either condition (or all conditions, if there are more than two). This can be done by a random number generator or similar (or you can even go old-school and pick a number out of a hat!). Random assignment is important because it *rules out the possibility of systematic differences between the groups.* For example, let's say the first 10 participants who take part in your study you put in a "faceless robot" condition, and the second 10 participants who take part in your study you put in a "face robot" condition. You ask both groups of participants how safe they feel on a 7-point scale from "not safe at all" to "completely safe". The issue with this is that there may well be pre-existing differences between the groups—participants who sign up earlier for an experiment may be (for example) more enthusiastic, conscientious, have more positive views towards robots, more likely to be female. *So you may end up with significant differences between the groups based on existing differences, not on the effect of the IV you are actually testing.* **If you use random assignment to groups you mitigate systematic bias between the groups, and this means you can more confidently say that the IV causes the changes in your DV.** So if you really need to demonstrate a casual effect you need to use random assignment. If you can't use random assignment, try to "match" the participants in each condition as much as you can—so try to have a similar mix of gender, age, ethnicity etc.—but without random assignment you can't say for sure that your IV causes the change in the DV.

Random and Non-random Assignment

Coramand: Researchers are examining the effect of motion type (IV: two conditions—fluid and robotic) on perceived safety. Because of technical constraints, they expose 10 participants (people from around the office) to the fluid motion condition first, then one week later expose a different 10 participants to the robotic motion condition. They compare perceived safety as reported by participants in each group. Each participant was therefore **not randomly assigned** to the different conditions. The researchers must acknowledge that existing differences between the two groups of participants could have played a role in any differences between the groups of responses (e.g. people who said yes first to taking part could be more interested in the project and have lower safety expectations than people who said yes later).

Tommy: Researchers want to know if the robot with enhanced capability (IV: two conditions—original and enhanced) produces more attachment in participants (DV). When every participant is delivered a robot they are **randomly assigned** via an algorithm to receive either an original capability product or an enhanced capability product. When they compare the attachment ratings of

the two groups of participants, they can conclude that any differences they find are not due to existing differences between the groups.

13.5.10 Reviews and Meta-Analyses

Some research doesn't collect any new data, but instead "pulls together" all the published research on a particular topic into a single article, and summarises it—these are called reviews. Reviews are considered secondary sources, as no new raw data is collected. Review articles are incredibly useful, as they collect together the key research that has investigated a particular question. Some reviews are called systematic reviews, if they use systematic methods to search the literature—they will list what search terms they used and what databases they searched, and have pre-specified eligibility criteria about which studies to include in the review. Systematic reviews give you a rigorous assessment of findings on a participant topic, so they are the best evidence available to answer a particular research question. If you are working to sell a robot product to an organisation, for example, you could conduct a systematic review of available safety studies, to establish and communicate clearly to the potential customer the current evidence on safety in that setting. Or you could conduct a systematic review before you start a research project, as if there is enough evidence already existing, you may not need to conduct the study! However, keep in mind systematic reviews can be very narrow in focus, so they may not address the particular question you're interested in, and they sometimes don't give you the "big picture" of what's going on in that field.

Meta-analyses are typically systematic reviews in which the author also statistically analyses the data they've found in the studies they review. In this way they can generate new data that numerically summarises the findings. Meta-analyses provide an objective assessment of evidence in a field, however if the original selection of the studies is biased, this means the outcomes of the meta-analysis can be influenced.

13.5.11 Which Research Design Is Best?

A theory can be explored using a wide range of designs—one design isn't better than another, they just provide different ways of investigating the theory. Different designs will also give you different types of information and allow you to draw different conclusions! So make sure you have a clear idea of your motivation, audience and research question before you design your study.

How Different Research Designs can be used to Investigate a Research Question
Let's take the example of Sylax, a university-based research robotics project investigating the effect of factors like behaviour and appearance of the robot on peoples' perception of agency. You could use a broad range of research designs to investigate this question, depending on the particular factors you were interested in.

Descriptive Design: If you were just starting out looking at this research question and wanted to better understand human–robot interactions, you could bring in a single participant and ask them to interact with the robot for an extended period, and video record that process. You could then sit with the participant and watch the video together, examining and discussing all interactions, gaining an understanding of their thoughts and emotions during the interaction, focusing on the aspects of the interaction you're particularly interested in (appearance and agency).

Survey Design: You could recruit a large number of people to interact with the robot, then after the interaction give them a survey about what they noticed about the robot's appearance and how high they rated the perceived agency.

Experimental Design: You could run an experiment where one group of participants interact with a robot with no face, and the other group of participants interact with a robot with a humanoid face, and compare the groups' perceived ratings of agency.

Review Design: You could look at previous research which has explored this topic before. If you wanted to focus on the effect of facial appearance on perceived agency, for example, you could identify all studies published in the last 30 years which involved presenting different types of faces to participants and measuring agency, and summarise their findings to come to a conclusion about the evidence that faces affect agency.

13.6 Sampling, Reliability and Validity

13.6.1 Sampling

For purposes of generalisation, you should do your best to ensure that the people in your study—your sample—is a representative sample of the population you want to apply your findings to—the population. That is, you're getting your data from a

sample that has *the same characteristics as the population* (e.g. same gender breakdown, same age distribution, same ethnic distribution). If the sample isn't representative in this way, it can cause serious errors when you try to apply your findings to the population. For instance, if you use university students as your sample when testing a particular user interface, you might find that when you roll out your product to the general population, users who are older or younger than your sample may engage very differently with the interface.

There are two broad types of sampling approaches, known as probability samplingand non-probability sampling. Probability sampling is when you select from your intended population so that any member of the population has a specifiable probability of being sampled—for example, you could get a list of the entire population, and select every 10^{th} person on the list to contact. When you use non-probability sampling there is not an identifiable probability of each member of the population being included in the sample. One common example of non-probability sampling is convenience sampling, where you just select your sample from whoever is available around you! Purposive sampling is another type of non-probability sampling, where you deliberately recruit people who meet a certain requirement—such as interviewing elderly people if that is who the robot is aimed at, or recruiting factory workers if the robot is an industrial product.

Using non-probability sampling like convenience or purposive sampling can be fine, as long as the sample you end up with is representative of the population on the particular variables you are interested in. Generally, the bigger the sample is, the better it will reflect the population and so you'll be able to better generalise to the population. But if there is systematic bias in your sampling you'll just make incorrect inferences more confidently… For example, many robots designed in universities or start-ups are only tested using convenience sampling with students or other people involved in the business. This means often the participants are only people who are young, and already interested in and knowledgeable about robots. Findings from studies using samples like these won't necessarily apply to the general public! Also keep in mind that the size of the sample will be reduced by non-response—so people drop out of the study, or forget to enter data.

13.6.2 Reliability

Reliability is our confidence that a given finding can be reproduced again and again—that it isn't a chance finding. For example, if you find that people respond more positively to a robot with child-like features than a robot with adult features, that finding is reliable if other researchers consistently find the same pattern. You can think of reliability as similar to *consistency*. However, just because an effect or test is *reliable* doesn't mean it is *valid*.

13.6.3 Validity

Validity is our confidence that a given finding shows what we think it shows. There are four key types of validity: construct validity, external validity, internal validity and ecological validity.

Construct validity asks whether we measured what we were trying to measure. This is harder than it seems! Often people do not interpret the task or question the way you intend, or other factors affect how they respond. For example, often when asked about the usability of a robot, people's responses will actually reflect their judgements about safety instead.

External validity asks how well we can generalise what we have found in our study to other times, populations and places. Let's say you conduct a survey examining attitudes to games involving a robot, using undergraduate students from Canada as participants. Will the findings be the same if I used a sample from a nursing home? Undergraduate students from China? In 10 years?

Internal Validity asks did the outcome reflect the IV we manipulated. Did the variable we are interested in cause the result? Let's say you run a study in which you want to look at the effect of working as part of a team on performance. So you have some participants complete a difficult task on their own and others complete it with other people. Can I conclude that lower performance in the teamwork group is because teamwork per se lowers performance? Not necessarily—the effect could be due to embarrassment, personal space factors, cognitive load, for example, rather than teamwork itself.

Ecological validity asks how well the findings of the study apply to real-world settings—how well does this finding actually work in the real world? For instance, if you test how people use a new type of technology in the laboratory, will they actually use it that way at home? On the bus? At work? *Note: This is different from external validity, which is about generalising to other populations/places.*

13.6.4 Things that Can Go Wrong with Validity

As we talked about above, sampling bias can affect external validity. If only certain types of people respond to a questionnaire or take part in a study, this limits who the findings apply to. It is notoriously difficult to recruit middle-aged people into studies, for example, as they are busy with young children and full-time jobs. If this is who is going to purchase your product you need to make sure your sample includes that group. Mortality—when people drop out of a study—can also affect external validity. For example, if taking part in the study requires an hour per day, those who have less time to take part will drop out of the study, meaning the results will only apply to people like those who remained in the study. If you are running an experiment and more people in one condition drop out than the other, this can also affect internal validity. This is particularly an issue if you are conducting longitudinal

studies (research that is conducted over a long period), where you can get *differential drop out*, with participants more likely to drop out of the study if they are in one condition rather than another. Reactivity is when something about the study itself means only particular people respond or influences how people respond. So, if you advertise your study as "Come and play with robots!" you will only get people taking part who already are positively disposed to interacting with a robot, missing a large section of society.

You also need to keep in mind that social desirability can affect how people respond in a study, as people tend to behave based what they think will look good to others. For instance, people tend to over-report their vegetable intake and under-report how many hours of TV they watch! People also change their responses based on cues in research which suggest how they *should* respond—known as demand characteristics. For example, if the researcher repeatedly asks a participant how much they liked the way the robot's eyes moved to follow them, the participant will tend to provide more (and more positive!) information on this element, even if other aspects of the robot were more interesting to them. Sometimes they may even unconsciously change their responses to match what they think the researcher wants. You also need to be aware that often people change their behaviour simply because they are being watched! This is called the observer effect (e.g. if you knew you were being monitored for how much sugar you're consuming, are you likely to eat less sugar?).

Testing effects are when a previous testing situation affects the subsequent testing situation. A gap between pre- and post-test can cause practice effects (i.e. people tend to get better on the same task when they complete it for a second time) and fatigue effects (e.g. are people going to be concentrating all the way through a two-hour testing session?). Maturation effects are when changes occur just because we are measuring things over time. For example, if you are measuring a how a child with a long-term illness interacts with a robot over time, the simple effects of the child ageing are likely to influence the outcomes. Changes in society can also occur during a study, and this can result in history effects (e.g. 9–11, COVID-19, social views on technology). You need to take care that changes in the data due to these factors are not misinterpreted.

Confounds are another threat to internal validity, and reflect changes in your DV that are due to another variable, NOT your IV. Consider if you introduce a learning robot into a school to improve mathematical skills. You compare maths skills in the classrooms without the robot to those with the robot. However, when adding the robot into the classroom this involves change, excitement, new staff etc. Any improvements in "maths ability" could be due to any of those factors, rather than the robot itself.

13.6.5 Ways to Address Problems with Validity

The above section, with all the many potential problems, might make it seem like it's impossible to design a "perfect" study! While an individual study can never avoid

absolutely every potential source of bias, there are simple steps you can take that will address many of these issues.

Unobtrusive measures can be used to address reactivity, demand characteristics and the observer effect—these are measures of behaviour that are not obvious to the person being observed. Examples are using one-way mirrors, measuring factors people are usually unaware of (such as how far they stand from a robot interface, or how often they touch it) or using other methods altogether, such as archival records, which are data which was collected for a different purpose. You can help reduce demand characteristics and reactivity by hiding the real purpose of an experiment. You can use deception, where you lie to participants about the purpose of an experiment, or you can use concealment to avoid telling them the whole truth (but be careful of ethics! See below). Using blinding is also really useful in addressing a range of potential biases. Blinding is when key people involved in the study don't know information which could affect their responses. In single-blind studies, the participants don't know which treatment group (level of the IV) they are in. In double-blind studies, both the participants and the researchers don't know which treatment group participants are in.

Focus on Living Labs

- Often when you try to increase internal validity (e.g. control the study tightly) you end up decreasing external validity, so there can be a trade-off between the two.
- A recurring critique of HRI has been the lack of consideration given to the ecological validity of experiments, resulting in poorly designed robots and interactions for the intended task.
- Attempts to address this should consider use of ecologically valid approaches (real-world conditions), such as field experiments or use of "living labs".
- In-the-wild is another term used to describe such experiments. One useful approach is to compromise by situating your experiments in more accommodating venues such as technology museums and gallery spaces, where you find populations open to such experimentation but with reasonable diversity, providing a context that more closely resembles a real-world environment.

13.7 Ethics

13.7.1 Ethics and Ethics Review Boards

All throughout this chapter, we've been talking about designing and evaluating studies—recruitment, sampling, randomisation, etc.; however, it is also important to consider ethical principles in research design. Although most lab-based robotics studies are relatively benign, it is still crucial to be able to identify and address ethical issues and demonstrate to a research ethics committee or institutional review board that your study is appropriate to be conducted. This is in addition to being aware of the broader implications of ethical robot design considerations discussed in Chap. 16 and safety of robot deployment discussed in Chap. 14.

Common Ethical Issues in Robotics User Studies
So what are some of the typical ethical issues you might face when conducting human–robot interaction research?

- Some of the most important risks are around data: Who has access to the data? Where is it stored? If you are using video recordings this is particularly important, as it is difficult to ensure confidentiality when people are identifiable from their images.
- There can also be physical danger from proximity to robots that needs to be carefully considered, and this risk communicated to participants when they give consent to participate.
- Remember participants won't know most of the terminology you are used to using, so write all participant-facing information in easy-to-understand language.
- If you are asking people you know to participate (friends, relatives, other students) make sure they actually want to take part and don't feel coerced! Avoid asking people you have an unequal power relationship with, such as students you are teaching or supervising.

There are regulations that govern what research can be conducted, typically at the institutional, national and international level, so you need to check based on where you are located. Most broadly the World Medical Association Declaration of Helsinki (*Ethical Principles for Medical Research involving Human Subjects*) is applicable to any research you do in which you recruit participants.

The ethics review process is a formal procedure in which you write a statement including details about the research project and addressing any ethical concerns, which is then submitted to a research ethics board for approval. The board will approve, reject or ask for changes to the study or more information to be provided before approval. Many researchers view applying for ethical approval as purely a

logistical a stumbling block, but ethics boards will sometimes identify very real ethical concerns that the researcher has not considered. Ethics boards are typically composed of experienced researchers, legal experts and laypeople. Each of these groups can give insight from those perspectives that you may not have thought about, that necessitate changes to your project.

13.7.2 Ethical Principles in Research

There are many different ways of considering ethical principles, all of which are based on the fundamental idea that you show respect to the people who take part in your study. This means that you are considerate of their experiences and take all the steps you can to be sure that they are consenting freely to participate, and are protected from harm. Some of the key principles to keep in mind are to use informed consent, minimise risk, ensure confidentiality and provide debriefing.

13.7.2.1 Informed Consent

People should participate in your study based on free, informed consent. This means that you provide them with all relevant information about the study (including any potential risks), that they understand this information, and that they are not pressured into participating in the study. You need to ensure that your participants are able to consent. For children or people with cognitive impairments (e.g. dementia), this requires additional checks—it typically means informed consent from both the guardian and the person themselves. You also need to provide information in a way that ensures people understand what you are saying (i.e. free of technical terms or jargon). Participants should also be free to discontinue the study at any time. This means that they can withdraw from the study, without any penalty, and do not have to provide a reason for doing so.

13.7.2.2 Minimise Risk

When designing your research (and writing your ethics application), you will need to carefully clarify the benefitsand risks of your research. In terms of the benefits, be explicit—what will this particular study tell us that we don't already know? Who will benefit from what you learn from the study? Will there be any broader social value? Will the participants get any benefit out of it? Remember that this all depends on your study being well designed in the first place, so that it gives you accurate data about what you're trying to assess.

You then need to weigh these benefits against the risks. One common risk is stress. Try to minimise unintended or unnecessary stress, and remember that what might not be stressful to you, might be for participants! So consider all the ways in

which you can reduce stress for the participants. If you are using any deception—giving participants information that is false—this is a potential source of risk as it violates informed consent and can cause harm. You should only use deception if required, and if you do, you need to undertake debriefing (see below) to disclose that deception and the reason for using it to the participant. Obviously if you are doing anything which is invasive, this risky! Invasive research is any research which changes the participants, such as administering drugs, inserting a recording device into a person's body, or exposing them to a situation where they could potentially be hurt or physically impacted. You need to ensure that what you are doing is absolutely *necessary* (the study won't achieve its aims without it), that you have *minimised* any risk during the study, and *removed* any long-term negative effects.

13.7.2.3 Confidentiality

Participants often provide sensitive information during a study and it is your responsibility to keep this information confidential. Remember also that information you may personally not consider sensitive (e.g. weight, performance on a task) may be considered sensitive to others. There are various ways you can ensure confidentiality. The easiest is to ensure participants are anonymous—that the data they provide is not identifiable. Using a participant ID number rather than names is good practice, for example. If you're conducting case study research, you may choose to refer to that person by their initials or a pseudonym rather than their name. If you are using audio or video recording, you should specifically ask for the participants' permission for this on the consent form. Data storage and protection also needs to be considered—where is the data stored? No one should have access to the data that isn't part of the project. Ensure storage devices are appropriately and securely protected.

13.7.2.4 Debriefing

Debriefing is explaining to the participants who took part in your study exactly what the study was about and what occurred during the study. This will counteract any deception that took place during the study (i.e. you tell them the truth about what happened in the study and why) and hence minimise potential harm. You should also encourage them to ask questions about the study and you should answer them fully.

The Wizard-Of-Oz Paradigm

When conducting user studies, at times researchers need the participating robots to exhibit capabilities beyond their technical abilities (either because it is not possible at the current state of the technology or due to non-availability of resources). In such situations, a commonly used technique in HRI is to

augment the missing skills through the integration of a human "wizard". Let's use Tommy (see above) as an example. If we were to examine the impact on participant–robot attachment by comparing a version of Tommy which converses verbally compared with a version which uses only non-verbal cues, it might be difficult to implement a fluid Natural Language Processing system that could mimic human competency appropriately. In such a situation, a confederate (another researcher) could be placed behind a curtain to converse with the research participant through the robot, giving the illusion that the participant is conversing with the robot. The concept comes from the classic fantasy novel "The Wonderful Wizard of Oz" by L. Frank Baum. While helping researchers to circumvent technical difficulties in conducting user studies, it should be cautioned that the practice has ethical and social implications. Ethical, in that you are potentially deceiving a research participant into believing the interaction is purely with a robot. Socially, when such research is presented in the wider media, there is a risk of misrepresenting the current state of the technology, leading to false understandings about the capabilities of robots. This has implications for research funding and the formation of exaggerated expectations or fears towards robots in society.

13.7.3 Data, Analysis and Interpretation

The earlier sections of this chapter have introduced you to the key factors to consider when designing a research study. While this text does not attempt to teach you data analysis (that's several other texts just on its own!), this section walks you through some of the fundamentals of data analysis, so that you can work out what analyses you need to conduct and can use other resources to follow up how to do those analyses with whatever data analysis program you are using.

In this section, we are assuming that you have conducted your study and collected your data. You are probably looking at a data file listing a big bunch of numbers and wondering what to do now! This chapter will help you understand what you need to know to take the next steps.

13.7.3.1 Research Data

One of the first things you need to know is what type of data you have, as this will allow you to work out what analyses you can do with it.

Although there are many definitions, qualitative data is generally considered to be data that describes or characterises what it is measuring—it is typically descriptive information about attributes in the form of words, that you can't easily summarise in numbers. You often collect qualitative data if you are conducting case studies

or observational research, or using other open-ended ways of gathering data in real-world settings. There are many different ways of presenting and analysing qualitative data, including approaches like grounded theory, thematic analysis and discourse analysis. In contrast, quantitative data is data that can be represented as numbers. But although this sounds simple, not all quantitative data is the same! Overall, knowing what *type* of data you have is important and depends on how you chose to measure your variables (see *operationalisation* above).

If you're going to conduct statistical analyses on your data, you need to work out which of the following four types it is.

- Nominal variables are variables which measure what category people fall into. Examples are gender (female, male, non-binary etc.) and the condition someone is in an experiment (control, experimental, etc.).
- Ordinal variables are categorical variables that are sequenced in a certain way, such as grades in school (A, B, C etc.) or outcomes in a running race (1st, 2nd, 3rd). In other words, the categories "go" in a certain order. However, these ordered categories do not have consistent intervals between each category.
- Interval variables are variables in which responses are quantitatively related to each other, with equal intervals between them but no true zero. For example, IQ is an interval scale as the "0" is not a true absence, but just the lowest score on that measure.
- Ratio variables are variables in which the numbers are quantitatively related to each other and have a true zero. This includes variables such as weight and height.

Once you are clear what type of variables you have, you can work out what descriptive and inferential statistics you can conduct on that data.

Examples of Common Variable Type in Robotics

Many of the variables below could be several different variable types –

it depends exactly how you've chosen to measure (operationalise) them.

Reaction time (milliseconds between robot movement and participant reaction) Ratio

Perceived agency (5 levels strongly agree to strongly disagree) Ordinal.

Reported safety (combined score across 10-item questionnaire) Interval.

Distance (centimetres between robot and participant) ratio.

Experimental condition (robot without a face, robot with a humanoid face) Nominal

Previous exposure to robots (none, occasionally, frequently) Ordinal

Interactions (number of times the participant initiated conversation): Ratio

13.7.3.2 Descriptive Statistics

Descriptive statistics are numerical statements that summarise the data you've collected from your sample. You will need to report the descriptive statistics of your data when you communicate your findings to your audience. Think about it—if you collect data from 40 people, you can't just give all those "raw" numbers in your presentation or your report! You need to summarise them in some way that tells your audience what your data "looks like" in a simple overview.

How you describe your data depends on what kind of data you have. If you have categorical variables (nominal or ordinal), you typically report the number of people in each category, and/or the percentage of people in each category. For example, if you asked participants whether they trusted robots, and they were given the option of yes or no, this would be a categorical variable. You would then report as your descriptive summary the number of people (the *n*) who said yes and the number of people who said no. You could also report the percentage of participants who fell into that category. For example, "Of the participants, 10 people (25%) reported trusting robots, and the remaining 30 (75%) reported did not."). Remember also to report how many people didn't answer that question, if that occurred.

If you have numeric variables (interval or ratio), you report the "middle" of the values for that variable, and how spread out they are around that middle point, as this is more meaningful than *n* or percentages with these kinds of variables. The "middle" of the data set is usually the mean, the median or the mode. The mean (*M*) is calculated by adding all values together, and dividing by the number of values. The median is the central value when all values are ordered from smallest to largest, and the mode is the most common single value for that variable. Of these the mean is most frequently reported. The most common ways to report how spread out the data are the range and the standard deviation. The range is the difference between the smallest and largest value for that variable. The standard deviation (*SD*) is a measure of how much the values vary around the mean—a larger standard deviation means the values are more spread out, a smaller standard deviation means they are less spread out. All of these descriptive statistics can be easily calculated using available statistical software. For example, if you measured how close people stood to a robot (in *m*), you could summarise that data as "The participants stood on average 2.17 m ($SD = 0.73$) from the robot."

13.7.3.3 Inferential Statistics

While it's useful to report what the results of the study are for the participants in your sample (descriptive statistics), usually you want to make a statement that goes beyond the people in your sample, and talk about what this means for the entire population you want to apply your findings to. Inferential statistics are numerical statements that draw conclusions about the broader population based on your sample data. While teaching you inferential statistics and the associated statistical theory is (well!) beyond the scope of this chapter, the following should give you a quick

overview so that you know what kinds of questions to ask, and how to seek help, when you begin to conduct analyses.

The first thing you should do is recognise your limits! Appropriately conducting statistical tests requires a good understanding of the theory those analyses are based on, what those analyses represent and what they can tell you (and what they can't!). You should first find guidance in terms of a statistical advisor, a senior supervisor, or take part in an introductory statistics course (there are many available online) to upskill you in these factors. If you don't understand what you are doing you are likely to conduct inappropriate analyses and/or draw inaccurate conclusions.

Once you have a basic knowledge of inferential statistics in general, you then need to decide how to apply that knowledge to your particular study. The place to start is your hypotheses. You should have pinned down particular hypotheses at the beginning of the study—what exactly did you predict? These will form the basis for your analyses. You normally have several different hypotheses in the one study, and you will need to go through this process for each one to decide which analysis is appropriate for each hypothesis. For example, if you are manipulating what motion type the robot exhibits and measuring perceived safety, you might also in the same study be measuring acceptability and how far people stand from the robot. Or you might also be manipulating the colour of the robot in the same study. You would have separate hypotheses for each of these effects and so these would be different analyses.

Steps when deciding on and conducting analyses.

- What was your initial hypothesis?
- How did you operationalise these variables? Identify the IV(s) and DV(s) involved in this particular hypothesis (how exactly did you measure or manipulate them? Do have conditions, if so how many, and are they between-subjects or within-subjects?)
- What types of variables are those IVs and DVs? (i.e. nominal, ordinal, interval or ratio?)
- Work out which analysis is relevant for you to run, given your IVs and DVs (you can use a decision tree like the one provided below).
- Check the assumptions of that particular analysis (e.g. some tests require normally distributed data) and change analyses if necessary (e.g. to a test that doesn't require that assumption).
- Conduct analysis and interpret output.
- Write up the output, conveying all necessary information to your audience (Fig. 13.6).

Research Examples: Selecting and Conducting Analyses

Let's take the example of Sylax, the university-based robotics project looking at the effect of factors like behaviour and appearance of the robot on perception of

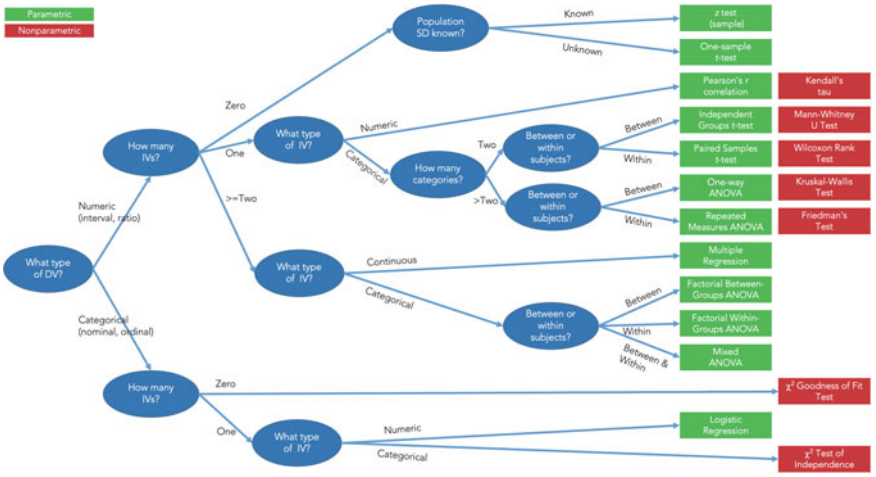

Fig. 13.6 Decision tree representing common statistical analyses

agency. The researchers have lots of questions about these different variables, but in one study they focus particularly on comparing random versus purposeful behaviour.

- Their *hypothesis* for this particular study is that people will perceive the robot to have more agency when it shows purposeful movements rather than when it displays random movements.
- They conduct a study in which they manipulate behaviour and measure perceived agency. Their *IV is behaviour*, which has *two conditions*: random and purposeful. They use the same participants in each condition, so this IV is manipulated *within-subjects*. They assess their *DV perceived agency* by measuring the number of interactions initiated by the participant.
- The IV of behaviour is *nominal* (as there are two conditions/groups). The DV of perceived agency is *ratio* (as it is the number of interactions).
- Given that the researchers have a ratio DV, have one IV which is nominal, with two conditions, manipulated within-subjects, the researchers decide they should run a *paired-samples t-test*.
- They check the *assumptions* for a paired-samples *t*-test (e.g. that their data is normally distributed) and find that it meets those requirements.
- They conduct a paired-samples *t*-test using their chosen statistical analysis program (e.g. The Jamovi Project, 2021) and conclude that although there is a difference between the means of the number of interactions between the two conditions, the analysis shows that this is not *statistically* significant.
- They write up their findings including all the relevant information and values. They conclude that their hypothesis that people's perception of

agency will differ depending on the robot's movement type is not supported, and in fact movement type does not affect perception of agency.

13.7.3.4 Presenting Your Findings

When presenting your findings, you need to tell the person reading it (or watching, if it is a presentation) all of the relevant details so that they understand what analysis you conducted, provide them with the key values so they can see for themselves what you found, and clearly communicate what this means.

Some key information you should present:

- The name of the test you conducted.
- The variables involved (and name the conditions you are contrasting if relevant).
- The outcomes of any assumption testing.
- The key descriptive statistics (e.g. the means and standard deviations of each group).
- The key values output from your analysis (typically the test value such as r, t or F, the number of people or degrees of freedom, the p value).
- Additional information such as confidence intervals and effect sizes.
- Whether or not the outcome was statistically significant.
- Additional information to aid interpretation (e.g. the effect size, confidence intervals).
- Appropriate graphs or figures to illustrate your findings.

Examples of Typical Presentations of Results of Statistical Analyses

A paired samples t-test was used to compare the perceived agency (number of interactions initiated) between 20 participants exposed to a randomly moving robot ($M = 4.75$, $SD = 1.59$) and an algorithmically-driven robot ($M = 5.55$, $SD = 2.01$). Assumption checks confirmed the data was normally distributed. There was a mean difference of 0.80, 95% CI [-1.96, 0.36] between the number of interactions generated in the two conditions, but this difference was not significant ($t(19) = 1.44$, $p = 0.166$, $d = 0.32$).

A Pearson's r correlation analysis was conducted to examine the relationship between self-reported previous exposure to robots and attachment to a robot product. The assumption of normality was supported. There was a moderate, significant, positive correlation between previous exposure and attachment scores ($r(15) = 0.54$, $p = 0.026$), such that participants with more contact with robots before the study tended to report higher attachment to the robot.

13.7.4 Common Mistakes and Pitfalls

This section introduces you to some of the common errors that both novice and experienced researchers show from time to time. By knowing what they are hopefully you can avoid them!

One key way researchers run into trouble is not planning the analyses. You should have formulated your key hypotheses before conducting your study. What are the main effects you are looking for? What are the key variables? You may have four or five key hypotheses that you want to test in a single study. Your analyses should then be pretty straightforward—you are running the analyses that test those hypotheses! This will also protect you from what is called "data fishing" or "*p*-hacking", which is when researchers run many different analyses on a single data set, but only report the significant ones. This is highly problematic statistically as it means many of those findings may actually then be false positives. It is also problematic in that it reflects a poor understanding of non-significant results—just because a finding isn't significant doesn't mean it's not interesting or useful! *Not* finding a significant difference between two conditions, for example, may tell you that you don't need to incorporate that additional capability to improve safety, or that people can't differentiate between different robotic faces.

When you first get your data from your study it is very exciting! You are keen to see what you've found and it's all too easy to rush into conducting analyses without understanding the data. If you do this, you end up with whole pile of outcomes that are a big mess! And are largely uninterpretable! Take your time to get to know the data set—what type is each variable? Should any variables be recoded to make them more useful? (e.g. turning age from a number into a category) Do you have any missing data and is this problematic? Take a look at your data using graphs and descriptive statistics. Does it "look" ok? Are there any weird values that shouldn't be there or outliers that might suggest equipment failure or data entry error? Are your numeric variables normally distributed or will you violate some assumptions? What is your plan for this?

After you have conducted your analyses there are a few common issues that can emerge. One error you see frequently is people assuming correlation equals causation. For example, just because you find that people who have previous exposure to robots also are more attached to their companion robot, this doesn't mean one causes the other—it doesn't mean that increasing people's exposure will cause their attachment to increase. Maybe it is a positive perception of robots that causes both ratings to rise? Also be careful in interpretating all your findings more generally. If your sample size is too small, you will have low power, which means that you don't (statistically) have the ability to find some effects even if they are really there. There are programs available (e.g. *G*Power*) which will enable you to calculate how many people you need in a study to have a particular level of power. This is often a problem in robotics user research where we tend to have small sample sizes. Running a power analysis before you conduct your study, to work out how many participants you need, is an important step. Finally, even if you find a significant effect, make sure you consider

effect size as well (most statistical programs will calculate this for you as well, for each analysis). Some effects can be significant, but not meaningful! For example, if you find a significant difference between the perceived safety of two robots, if the difference itself is only.5 of a point on a 1 to 20 scale, it is unlikely to be a *meaningful* difference at the end of the day.

The key thing to remember is that you have to be confident in what you're presenting. Are you sure that the data reflects your conclusions? Are there any issues the reader should know about in order to interpret your findings appropriately? Remember others will use your work and build on it, so you want it to be an accurate representation of the world!

13.8 Chapter Summary

This chapter introduced you the fundamentals in conducting research in human–robot interaction. We have highlighted the importance of carefully designing research projects so that you get the most accurate and useful information out of the research you conduct. This chapter should provide you with the necessary knowledge and confidence to design your own studies in this field.

13.9 Revision Questions

Q1: You are conducting a research study focusing on the effect of robotic faces (child-like or adult-like) on perceptions of animacy. You predict that you'll find that child-like faces have higher animacy than adult-like faces. This is an example of a:

(a) Theory
(b) Hypothesis
(c) Control condition
(d) Relevance-sensitivity trade-off.

Q2: You want to work out which of three voice options for a companion robot elicits the most positive response from the general public. You give a group of people the same robot with one of the three difference voices and after a week ask them how positively they view the robot on a scale from 1 to 7. What is the IV and what is the DV in this study?

(a) IV positive rating; DV voice type
(b) IV robot type; DV safety rating
(c) IV voice type; DV positive rating
(d) IV before and after rating; DV voice type.

Q3: You figure out the best type of articulation to use on your robotic design by reading all the previous studies conducted looking at articulation and summarising them. The research design you are using is:

(a) Descriptive
(b) Correlational
(c) Experimental
(d) Review.

Q4: To establish what factors are important in designing an industrial robot for a particular company you send out a survey to all the company employees. You particularly want to know about the relationship between how long people have worked there and how important they think particular design features are. The research design you are using is:

(a) Descriptive
(b) Correlational
(c) Experimental
(d) Review.

Q5: You are testing how people react to the new robotic interface you have designed. You ask some friends if they can drop by the lab to help you test it out. You are using what kind of sampling?

(a) Convenience
(b) Purposive
(c) Probability
(d) Random.

Q6: You are testing how people react to the new robotic interface you have designed. You ask them to do a task with the help of the old interface, then do the same task with the new interface. They report they found it easier to complete the task with the new interface. What is one explanation for this difference?

(a) Observer effects
(b) Mortality
(c) Practice effects
(d) History effects.

Q7: You place two video cameras in the corner of your laboratory to record the interactions between your participants and the robot. What ethical issues do you need to consider when using these?

(a) Informed consent provided by participants to be videoed
(b) Secure storage of the video files
(c) Protecting confidentiality of the participants in the videos
(d) All of the above.

Q8: You are conducting a pilot test installing a robot in a manufacturing setting. You measure how many times people physically touch the robot during the course of a day. This is what type of variable?

(a) Ordinal
(b) Nominal
(c) Interval
(d) Ratio.

Q9: You conduct a research project and gather some data. You present your data as the percentage of participants who responded "Strongly agree" to each question. You are using:

(a) Inferential statistics
(b) Descriptive statistics
(c) The mean and standard deviation
(d) Significance testing.

Q10: A survey of the general population finds that people who are afraid of robots are more likely to say they don't need robotic help around the house. The researchers conclude that if they make people less afraid of robots they will then want more robots to help in the household. What error are they making?

(a) They are p-hacking
(b) They have low power
(c) They are assuming correlation equals causation
(d) They have low effect size.

References

Bretherton, I. (1985). Attachment theory: Retrospect and prospect. *Monographs of the Society for Research in Child Development, 50*(1/2), 3–35. https://doi.org/10.2307/3333824.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review, 94*(4), 412.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1984). Behavioral decision theory perspectives on risk and safety. *Acta Psychologica, 56*(1–3), 183–203.

The Jamovi Project. (2021). *jamovi* (Version 1.6) [Computer Software]. Retrieved from https://www.jamovi.org.

Janie Busby Grant is a research psychologist based at the University of Canberra. She is a cognitive psychologist whose research focuses on future-oriented thought, mental health and human-robot interaction. She has experience designing and conducting collaborative cross-disciplinary research projects and engaging with stakeholders with complex relationships and data sensitivities. Janie has taught research methods at different levels across three universities and is currently an Academic Fellow mentoring Ph.D. students across all Faculties in her current university. She specialises in the use of new technologies in both research and teaching contexts.

Damith Herath is an Associate Professor in Robotics and Art at the University of Canberra. Damith is a multi-award winning entrepreneur and a roboticist with extensive experience leading multidisciplinary research teams on complex robotic integration, industrial and research projects for over two decades. He founded Australia's first collaborative robotics startup in 2011 and was named one of the most innovative young tech companies in Australia in 2014. Teams he led in 2015 and 2016 consecutively became finalists and, in 2016, a top-ten category winner in the coveted Amazon Robotics Challenge—an industry-focused competition amongst the robotics research elite. In addition, Damith has chaired several international workshops on Robots and Art and is the lead editor of the book *Robots and Art: Exploring an Unlikely Symbiosis*—the first significant work to feature leading roboticists and artists together in the field of Robotic Art.