Treating uncertainty in sociotechnical systems

Mark Andrew

Scenario, Australia

ABSTRACT

Design goals guide design efforts but complex systems can lead to designers' intentions being eclipsed. This paper's proposition is that sociotechnical systems design offers scope for improved reliability and is built on three features of current design practice. First, design teams seek cooperative cognition to work together but inadequately understood outcome scenarios can impoverish joint understanding. Second, design team collaboration is bounded by innate psychological biases, which can spoil design decisions. Third, some views of risk in design thinking suffer from a limited conception of uncertainty and its influence. These constraints in design practice are examined (referencing the reach of Artificial Intelligence as one example design domain) and how such constraints may be addressed in design practice.

KEYWORDS

Sociotechnical systems, uncertainty, artificial intelligence

Introduction

The French expression *feu de joie* (fire of joy) describes a military celebration when a regiment fire one shot after another, in close succession like a drum-roll. "Symbolically, the fire of joy is a reminder that the regiment's collective power relies on the individual, and vice versa" (James, 2020, p.1).

That individual and group performances are interwoven is both a strength and weakness of design thinking and offers fertile ground for enquiry. This paper focuses on one example of complexity in sociotechnical design, Artificial Intelligence (AI), to examine challenges and approaches to improve design practice in general for other sociotechnical systems. These sections follow: AI as an example domain for design enquiry; cooperative cognition in design teams; cognitive bias; and challenges with uncertainty. Conclusions will be described to inform design policy and values.

AI as an example for design enquiry

The notion of complex sociotechnical systems such as AI eclipsing their designers' imaginations is a reality. With the example of AlphaGo (DeepMind, 2021a), the design team harnessed uncertainty to their advantage. AlphaGo is an AI designed to play Go, a board game involving two players. Go offers an unimaginably large number of play permutations within a simple layout (black and white stones placed on a grid board). The object of the game is to occupy more territory than an opponent by bounding spaces using the stones.

In 2016 AlphaGo defeated the world champion Go player, Lee Sedol. AlphaGo won four of the five games and game two underscored its emergent capability. During game two AlphaGo made a surprising move, now referred to as 'move 37'. The move was unexpected, unconventional, and astonished Lee Sedol and the entire DeepMind design team. AlphaGo had exercised self-learning, innovation and creativity that exceeded the designers' imaginations. Since 2016, DeepMind has developed an AI called AlphaFold (DeepMind, 2021b) to assist with the problem of protein folding

which has delivered many medical research benefits, so AI efforts and results should not be thought of only as games.

AI development is careering in both senses of the word, and its reach and impact will be felt in most aspects of society including weapons' target decision-making in the form of Artificial General Intelligence (AGI). AGI "offers enormous benefits for humanity, yet it also poses great risk" (McLean et al, 2021, p. 1). Some sociotechnical risks result from selective data sets for example, such as when algorithms that instruct AI show a strong male bias (Winterson, 2021).

Many of the myriad outcomes from AI design decisions are the result of design teams' cognitive processes, coupled with the unexpected and unconventional behaviours of the system as evidenced by AlphaGo and exemplified by move 37. This sociotechnical uncertainty is the focus of the following sections, which are consolidated in a policy and value discussion in the conclusion.

Cooperative cognition

Design teams cooperate to build joint understanding, but this can be impoverished by inadequately pictured sociotechnical outcome scenarios. It is useful first to recognise the basic units of design team cooperation, namely groupwork and teams.

A work group can be defined as a set of more than one job holder in some organisational unit that may be permanent (Davis, 1969). Teamwork may come and go, existing only for as long as is required for a particular task (Kinlaw, 1991). Team performance emerges when tasks cease to be disjunctive and become more conjunctive or additive (Steiner, 1972). Disjunctive tasks are performed independently by group members and reflect individual choices among alternatives. Conjunctive tasks are performed together. The kind of task performed influences or even determines group performance and team development (Hackman and Oldman, 1980). The more complex the task, then often the more conjunctive, and hence cooperative.

Why do people cooperate? Holand and Danielsen (1991) suggest individuals make decisions based on a construction of their internal reality, and part of this construct includes cooperating agents, which are accommodated into a mental model or construct if the individual's self-interests are satisfied. The basis of a mental model or construct is developed by others such as Minsky (1974) and Johnson-Laird (1987) and illustrated by Hudson's account of self-determinism in committees.

Hudson (1983) suggests there are colonies of selves in each of us – such as the tactical, moral, civic, and capricious - and these are deployed to determine the internal construct used to identify, evaluate, and act on decisions. Justification becomes a process of developing the appropriate construct to accommodate the required decision. If the construct development process does not produce a conflict, the decision does not compromise the decision-makers.

A group exhibits "consensual rationality" (Lehrer, 1987, p. 87) when a group's members iterate their individual estimates of the likelihood that some proposition is true and then weigh them in terms of their various degrees of respect for one another (which ultimately leads them to a limit or consensual probability). Lehrer comments that the scheme should reliably model actual group activities, since "under expected conditions personal probabilities will coincide with consensual probabilities, and consensual probabilities will coincide with the truth" (Lehrer, 1987, p. 107).

Situation awareness is a necessary aspect for individual task-oriented situations. The individualist view of situation awareness is expanded when individual awarenesses jigsaw together at the system's level to yield distributed cognition. Distributed cognition describes cognition that "transcends the boundaries of an individual actor because the system's aggregate behaviour can be highly complex and adaptive" (Stanton, Salmon and Walker, 2019, p. 19).

Complex high-level functioning emerges from the combinations of low-level mechanisms and at its best is a cooperation between individuals and other artefacts, described here for design thinking as cooperative cognition, and is proposed as a vital component of design teams. These factors emphasise the need for group support design technologies that assist with making design values and principles explicit. Sound cooperative cognition also encourages a team's awareness of innate biases and the transparent treatment of uncertainty.

Individual and group biases

Heuristics are mental 'rules of thumb' that serve as useful everyday cognitive processes that bias our thinking to inform speed-error trade-offs when time is short. In this way, heuristics are useful for many decisions – but not all decisions. They are shorthand mind-sets that have been useful in "the environment of evolutionary adaptedness" (Bowlby, 1969, p. 58), but can be poorly suited to contemporary workplace design and decision-making.

Some common biases (Kahneman and Tversky, 1974) based on heuristics include: availability of information; representativeness (including how problem characteristics bias judgment on probability estimates); serial positioning effects (e.g., a tendency to recall first (primacy) and last (recency) items); anchoring and adjustments (e.g., an initial estimate sways judgment); and affect (e.g., when messages framed to evoke emotion can be biased either through dread or comfort).

If design team collaboration is bounded by innate psychological biases and sometimes these biases are features of group behaviour, such as groupthink (Janis, 1983), then organisational errors can be characterised as "upstream or latent errors acting in concert to magnify local individual errors" (Reason, 1995, p. 1708). Organisational errors include transport disruptions due to the commercial separation of maintenance from operations, and institutionally as the mismanagement of epidemics.

So bias and errors occur at various levels of sociotechnical systems (e.g., individual, group, organisation, and institution) and mostly in concert. System errors, after all, are human-induced design errors, and so the natural foibles and quirks of designers (at various levels) are likely to be reflected in designed system performance. Reports of such systemic bias include face respirator fitting "pass rates that are especially low in female and in Asian healthcare workers" (Regli et al, 2021, p. 1), "racial bias in pulse oximetry" (Sjoding et al, 2020, p. 1) and skin cancer diagnostic AI datasets with "substantial under-representation of darker skin types" (Wen et al, 2021, p. 1).

Another example of design-induced error is illustrated by Open AI's language-generating system called GPT-3. GPT-3 is a significant step towards AGI, and at first glance has an impressive ability to produce human-like text - but accuracy is not its strong suit. Although its output is grammatical, and even impressively idiomatic, it is untrustworthy. GPT-3 generates crucial failures such as this example of a prompt offered by researchers with GPT-3's continuation shown in **bold**:

> "At the party, I poured myself a glass of lemonade, but it turned out to be too sour, so I added a little sugar. I didn't see a spoon handy, so I stirred it with a cigarette. But that turned out to be a bad idea because **it kept falling on the floor. That's when he decided to start the Cremation Association of North America, which has become a major cremation provider with 145 locations**" (Marcus and Davies, 2020, p.1).

An emerging concern is that AI can also amplify biases as evidenced by a study examining natural language processing models: "Models generated many false answers that mimic popular misconceptions and have the potential to deceive humans" (Lin, Hilton and Evans, 2021, p. 1). As such, datasets are no more than selective stories (Winterson, 2021), and can lead to institutional errors if not treated as stories. A necessary design policy addressing such system design vulnerability is the treatment of uncertainty in sociotechnical systems development.

Treating uncertainty

Policy-makers suggest the effective linkage between science and risk decisions depends upon at least two goals. First, scientific uncertainties must be reduced (i.e., predictions must be more accurate). Second, technical specialists and subject matter experts must "effectively communicate the nature and magnitude of these uncertainties to people who must take action" (Andrew, 2016, p. 4). These intuitively attractive perspectives treat uncertainty as something to be overcome, and prediction as a "technical product that must be successfully integrated into the decision-making process" (Gallagher and Appenzeller, 1999, p. 79).

Examples of similar assumptions often erroneously applied to risk assessments in systems design include the following error types: future causality can be inferred from past incident antecedents; precursors can be isolated as single contributors to incidents; sufficient incident failure data exist; and controls can be identified on a value basis.

Key forces provoke such errors: "Risk ideology, which frames patterns of thinking for approaching analysis, and cultural norms for analysis, which evolve into a scheme of values adopted by risk analysts" (Andrew, 2014, p. 1). The measurement of a few risks can highjack attention away from searching for additional risks, leading to a triumph of precision over accuracy. A fuller search for risks (even if they are uncertain) is more useful than a scaled list of a limited number of risks: "Unfortunately, the problem due to uncertainty is compounded as existing hazard analysis techniques tend to ignore unknown uncertainties, and stakeholders involved in system development rarely track known uncertainties well through the system lifecycle" (Leong et al, 2017, p. 57).

One current pattern of thinking within system development is that risks are described as a product of the probability of an event or circumstance and the scale of the outcome or consequence. However, life rarely keeps still long enough to measure all important aspects and it is difficult to see beyond figures to embrace complex reality. The linear view of risk is inadequate to support a systems view of risk, where successful interventions occur by identifying leverage in systems. One approach to reframing risk is to model the construct to reflect its often non-linear, emergent properties. This more realistic view of risk as systemic, recognises multiple events leading to interacting exposures which may lead to changing consequences that impact in different ways depending on vulnerability. The impacts could be both positive and negative because reward is also linked to risk.

Systems thinking is a response to the technical focus of systems dynamics (Forrester, 1969) and has provided a language suitable for addressing complex design problems: "The systems thinking approach involves taking the overall system as the unit of analysis, looking beyond individuals and considering the interactions between humans and between humans and the artefacts within the system" (Stanton, Salmon and Walker, 2019, p. 3). It was motivated by a weakness of reductionist scientific analysis, which breaks a problem into parts and studies the parts in isolation to draw conclusions about the whole. This approach is ineffective for issues that are interrelated, exhibit emergence, and defy linear causation (often referred to as 'real life').

Circular causation, where a variable may be both the cause and effect of another, has become the norm rather than the exception in sociotechnical systems: "Behaviours emerge not from the decision or actions of individuals but from the interactions between humans and artefacts across the wider system" (Stanton, Salmon and Walker, 2019, p. 3).

True exogenous forces are rare. Recognition that the components of complex systems are fundamentally interconnected has emphasised the role of endogenous feedback loops, as illustrated in figure 1 which shows balancing and reinforcing feedback loops. When an arrow is used in

systems thinking diagrams, it does not denote linear cause but rather a circle of influence that may be both cause and effect (Senge, 1990).

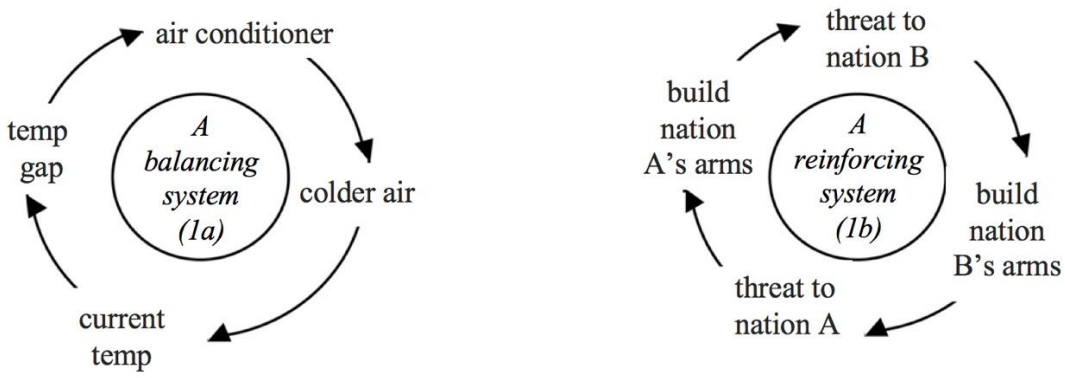

Figure 1: Systems thinking diagrams

Problem fragmentation can lead to an unexpected property that emerges because lack of recognition that any given element may be both a cause and an effect contributes to continual growth.

The emergent property construct is a powerful feature of systems thinking. Emergent properties in figure 1 are different because the systems have different structures given the behaviour of the elements when acting in concert. Although time delay is a feature of both systems, they are structurally different because the patterns of interaction result in different emergent properties - balance versus reinforcement, for example. Although a reinforcing system is illustrated here by an arms race, it may also describe a property of a well-designed sales and production system.

A systems perspective of probability is necessarily quite different to a historical view. A systems view of probability also recognises future scenarios and possibilities. This is different to data-centric models, which rely on defining reality tightly to allow measurement. In systems thinking terms probability emerges from a causal loop because patterns of events are better descriptions of systems behaviour than discrete events. Complex systems make data-centric limitations appear quite problematic, encouraging us to differentiate between detail complexity and dynamic complexity as they relate to probability judgements. Such complexities also highlight the need to distinguish resolvable uncertainty, which can be understood through discovery, from radical uncertainty, which cannot (Kay and King, 2020).

A systems model of consequence ideally uncouples consequence as a property that necessarily occurs because of probability. It is just as valid to say that consequences pull probability, as it is to say that probabilities push consequence. Definitions of probability or consequence being leading parts of risk are a matter of perspective and are not intrinsic to the meaning of risk.

One illustration of this effect is provided by prospect theory (Kahneman and Tversky, 1979), which describes our tendency to make different choices under different conditions, and to reframe the relationship between probability and consequence. Prospect theory suggests that when people are in a position of gain, they become increasingly risk averse (to maintain their gain); when people are in a position of loss and losses increase, they become more risk seeking (to reverse their loss). In systems thinking, people change how they deal with elements as patterns depending on context.

So, many influences are both cause and effect, and consequence interacts with probability. Figure 2 shows a systems view of probability and consequence comprising loops of elements that are both cause and effect.

The models of probability and consequence shown in figure 2 both involve uncertainty as one of their elements. Disruptions can be positive or negative (or both), as can vulnerability and impacts.

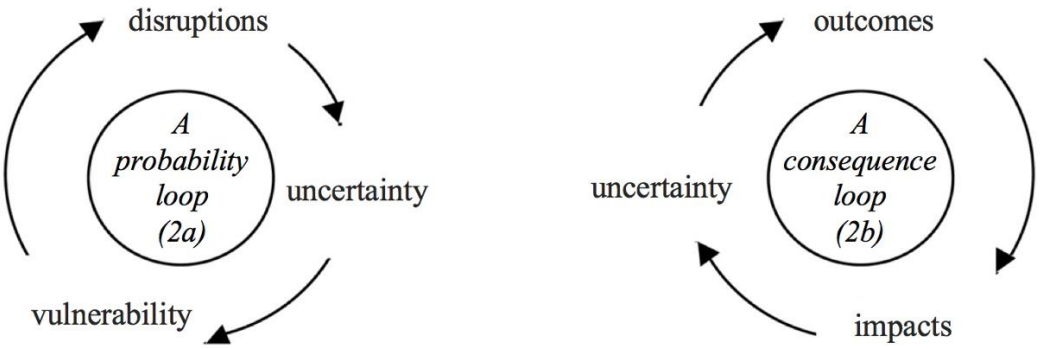

Figure 2: Probability and consequence systems diagrams

An uncertainty model integrates these factors into a cause-effect-cause picture. Figure 3 shows how retrospection and feedforward can be leveraged to yield foresight. In this uncertainty model risk acceptability is labelled risk appetite as this does not cast risk as a commodity to be avoided. Risk appetite describes how people feel about futures, providing insight for design decisions.

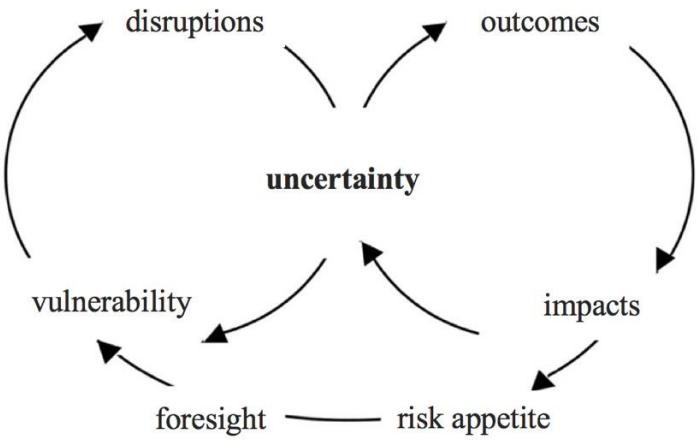

Figure 3: A systems model of uncertainty

Making risk appetite explicit is one of the key aims of the model in figure 3, because transparent stakeholder access to design decisions can be improved by recognising and valuing risk as an uncertain investment in potential reward and loss. For example, an unexpected outcome of AlphaGo's match with Lee Sedol was that it won the series but not every game. AlphaGo lost the fourth game (of five) after it won the previous games. This is significant because by the end of game three it had won the series – its goal. Is it possible that it lost game four because it explored more about its opponent Lee Sedol? Was AlphaGo using the opponent to learn more about itself because it could afford to? A long-term reward for AlphaGo could be to improve by risking losing a game, as it generated its own tactics within the goal of winning the series. Controlled mistake-making is after all one of the finest forms of learning. AlphaGo exhibited many behaviours of a designer as it mastered a form of 'deliberative uncertainty'.

The case for transparent stakeholder access to design decisions underpinning systems development has a significant role to play in sociotechnical systems' futures. Such transparency may be improved by considering uncertainty as modelled in figure 3, and the way it plays out in design.

Conclusions for design policy and values

This paper highlights selected limitations with cognition to illustrate how design policy and values may be improved. Design team decision-making has been examined regarding cooperative cognition goals in groupwork, and the innate biases of individuals and teams. Uncertainty multiplies these limitations, making the design of complex sociotechnical systems (such as AI) highly challenging. Cooperative cognition addresses these problems by focusing attention on dynamic versus detail complexity, and on resolvable versus radical uncertainty. By recognising these factors and suggesting one (of many) approaches to model uncertainty, cooperative cognition also makes biases easier to identify and treat: "All models are wrong but some are useful" (Box, 1979, p. 202).

How DeepMind operated as a team when creating AlphaGo is a complex question, but some insight is given by DeepMind's design strategy. The approach for AlphaGo was characterised by a decision architecture including a value network, a policy network, an evaluative network and network flexibility for scenario generation and self-learning. The design team operationalised a novel form of memory generated by creative and innovative percepts. Although some of AlphaGo's behaviours astonished the design team, their invention would likely not have been realised without the design team leveraging design policy and value transparency.

AlphaGo, as well as being a designed system, is also an accomplished designer. It exhibits the faculties of situation awareness, uncertainty modelling, and learns about bias inadvertently installed by DeepMind. As for AlphaGo's cooperative cognition, its policy, value, scenario, and evaluative networks are perhaps analogous to Hudson's colonies of selves (Hudson, 1983).

Similar AI systems can act as design process exemplars for design teams, gaming loss-making scenarios safely to expand learning potential. These lessons have been incorporated into the development of AlphaFold (DeepMind, 2021b) which accelerates medical research, enabling scientists to target and design cures for diseases more efficiently. The engineering of bacteria to secrete proteins that make waste biodegradable has also been demonstrated by AlphaFold.

This expanded design faculty also delivers the potential for unintended AI consequences with, for example, target-setting weapons systems, where military command being usurped is a critical uncertainty. Design policy and value principles explored in this paper are a contribution to making such design choices more transparent. Transparency helps recognise, for example, that datasets are selective stories and spotlights whose stories get to shape our reality (Winterson, 2021).

References

Andrew, M. (2014). Human errors endemic in risk analysis. *Proceedings of the Human Factors in Organisational Design conference*, Copenhagen.

Andrew, M. (2016). Modelling the dynamics of risk scenarios. *Proceedings of the IDC Second Safety in Design conference*, Melbourne.

Bowlby, J. (1969). *Attachment and loss*. New York: Basic Book.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, eds: Launer, R. L. and Wilkinson, G. N., Academic Press, 201-236.

Davis, J. (1969). *Group performance.* Reading, Mass; Addison-Wesley.

DeepMind (2021a). AlphaGo – the story so far. https://bit.ly/3uGGM6y (weblink accessed 01 October 2021).

DeepMind (2021b). AlphaFold – using AI for scientific discovery. https://bit.ly/3uHI6py (weblink accessed 01 October 2021).

Forrester, J. (1969). *Urban Dynamics*. MIT Press.

Gallagher, R. and Appenzeller, T. (1999). Beyond Reductionism. *Science*, 284.

Hackman, J. and Oldman, G. (1980). *Work design.* Reading, Mass; Addison-Wesley.

Holand, U. and Danielsen, T. (1991). Describing cooperation - the creation of different psychological phenomena. *Studies in CSCW*, eds: Bowers, J. & Benford, S., North Holland: Elsevier.

Hudson, L. (1983). Committed in Committee. *Times Literary Supplement*, Nov 11.

James, C. (2020). *The Fire of Joy*. Picador.

Janis, I. (1983). *Groupthink: psychological studies of policy decision and fiascos*. Boston: Houghton Mifflin.

Johnson-Laird, P. (1987). *Mental Models.* Cambridge University Press.

Kahneman, D. and Tversky, A. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, New Series*, Vol. 185, No. 4157, 1124-1131.

Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision making under risk. *Econometrica*, 47, 263-292.

Kay, J. and King, M. (2020). *Radical Uncertainty: Decision-making for an unknowable future*. The Bridge Street Press.

Kinlaw, D. (1991). *Developing superior work teams: building quality and the competitive edge*. San Diego; University Assoc. Inc.

Lehrer, K. (1987). Personal and social knowledge. *Synthese*, Vol 73.

Leong, C., Kelly, T. and Alexander, R. (2017). Incorporating Epistemic Uncertainty into the Safety Assurance of Sociotechnical Systems. *2nd International Workshop on Causal Reasoning for Embedded and safety-critical Systems Technologies.* EPTCS. *259*, 56-71.

Lin, S., Hilton, J. and Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Cs.CL arXiv:2109.07958* (preprint under review), Cornell University, 1.

Marcus, G. and Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://bit.ly/3DrrFAV (weblink access 01 Oct 2021).

McLean, S., Read, G. J. M., Thompson, J., Baber, C., Stanton, N. A. and Salmon, P. M. (2021). The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*.

Minsky, M. (1974). A Framework for Representing Knowledge. *Memos*, MIT-AI Lab, # 306.

Reason, J. (1995). A systems approach to organisational error. *Ergonomics*, Vol 38 Issue 8, 1708-1721.

Regli, A., Sommerfield, A. and vonUngern-Sternberg, B. S. (2021). The role of fit testing N95/FFP2/FFP3 masks: a narrative review. *Anaesthesia*, 76, 91-100.

Senge, P. (1990). *The Fifth Discipline*. Doubleday.

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E. and Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *The New England Journal of Medicine*, 383, 2477-2478.

Stanton, N. A., Salmon, P. M. and Walker, G. H. (2019). *Systems Thinking in Practice*. CRC Press, Taylor & Francis Group.

Steiner, I. (1972). *Group processes and productivity.* New York; Academic Press.

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zependa, L., de Blas Perez, C., Denniston, A. K., Liu, X. and Matin, R. N. (2021). Characteristics of publically available skin cancer image datasets: a systematic review. *Lancet Digital Health*, November 2021.

Winterson, J. (2021). *12 Bytes: How We Got Here. Where We Might Go Next*. Jonathan Cape.