University of Amsterdam

Research Master's Program Brain and Cognitive Sciences, Literature Thesis

Strategyproofness and Social Choice: a New Perspective for Cognitive Modelling

Matheus Boger
Student Number: 14163799

supervised by
Dr. Federico Fioravanti
Institute for Logic, Language and Computation, University of Amsterdam

examined by
Dr. Ulle Endriss
Institute for Logic, Language and Computation, University of Amsterdam

May 29, 2024

Abstract

Social choice theory is a subfield of economics and mathematics which models the aggregation of preferences. The most typical example of its applications is in political sciences, where the aggregations modelled are votes in an election. This thesis aims to serve as a mathematically simple introduction of this theory to cognitive scientists. This is because we believe the principles behind social choice are fertile ground for new cognitive modelling techniques. Thus, we present some central concepts and results of the field while making associations to how these could be applied in brain sciences. From its main application, social choice theory adopts the voting terminology to lay down an axiomatic structure to this aggregation process. As such, 'voter's ballots' are compiled into 'winners of elections'. The compilation procedure is called a social choice function (SCF), and the whole of the ballots is called a profile. All possible profiles under an election constitute a domain. When voters are not able to gain an advantage in the election by lying about their preferences under a certain rule, this SCF is said to be strategyproof. The Gibbard-Satterthwaite Theorem gives us the main result of this field: no SCF can be strategyproof without restrictions to the domain unless it is a dictatorship (where only one voter decides the result of the election). Therefore, studies have been conducted to prove which SCFs would be strategyproof under different domain restrictions. This thesis presents three different restrictions (single-peakedness, single-crossedness and separable preferences), along with the motivation why such restrictions were thought of, the group of strategyproof SCFs found for each one and a possible applicability of these mathematical structures for cognitive modelling. In this way, we tie the ideas up until now mainly applied in social sciences to new possibilities in the field of cognition. We finalize this presentation of social choice theory to cognitive scientists by pondering what difficulties might be faced in the adaptation of these structures to cognitive modelling, specially when we take in consideration the large diversity of data types we commonly find in cognitive science, ranging from behavioral to cellular.

Contents

Preface: Notation

This is a reference list for all important mathematical notations used throughout the thesis. In case a paper cited in the body does not make use of this notation, its notation is converted to this one in this thesis for consistency. Proper definitions and explanations of concepts pertaining to social choice are available at section 1.2.

$\mathfrak{L}(A)$: set of all linear orderings for a given set A

$\mathfrak{C}(A)$: all nonempty subsets of a given set A

A: unless stated otherwise, the set of alternatives on a ballot

n: unless stated otherwise, a natural number of voters in an election

$\succeq_i$: abbreviation for the linear ordering that represents the vote (ballot) cast by voter i

$b_i(A)$: the most preferred candidate in A by voter i

P: the profile of all ballots of an election

$\mathfrak{P}$: the set of all possible profiles in an election

O: the set of all objects whose subsets can be made into alternatives

$G(>_i)$: the set of all good candidates according to voter i

Chapter 1

Introduction

1.1 Motivation

Even though cognitive modelling is a widespread tool on the belt of many researchers of cognition, there are still many mathematical modelling techniques that have not come under the close inspection of the field. Particularly of interest in this thesis is the subarea of mathematics that is by convention referred to as "social choice". This naming might raise a few eyebrows for the unfamiliarized cognitive scientist, so let us first go over what social choice is not.

Social choice, in this context, is not an area of psychology or neuroscience. It does not concern how humans come to a judgement or decision regarding social environments, much less how that process is implemented in the brain. Social choice is also not a subfield of statistics, so its methods are not like the classical regressions and ANOVAs one might encounter in most empirical sciences. Social choice is, however, a subfield of economics, and can also be considered a subfield of mathematics, which attempts to model collective decision making, and has been applied, among others, to political science and sociology already, especially regarding voting in democracies and participatory budgets (Brandt et al., 2016). Another example of an application of social choice that could be of interest to the cognitive scientist is AI, where social choice was used to give insight into the alignment problem (more on that on Chapter 6). Basically, any field that has to deal with aggregation of different opinions is fertile soil for the application of social choice theory.

From our standpoint, social choice in political science stands in a similar position as the evidence accumulation models in psychology.[1] Both are rel-

[1] These are models which attempt to explain speed and accuracy distributions in behavioral experiments. For an introduction to them and their application in cognitive science,

atively simple mathematical stipulations of how decision making processes can be interpreted that are more descriptive of the process itself than pure statistics. However, a key difference between the two methodologies is that social choice has a longer history and a much more mathematically theoretical approach to it than evidence accumulation models.

It is possible to trace prototypical notions of what came to be social choice back into classical antiquity, but for practical purposes, it can be said that the mathematical study of collective decision making began during the Enlightenment, when the French Jean-Charles de Borda and the Marquis de Condorcet started to argue with mathematical substance over the flaws and strengths of different aggregation rules of votes, trying to determine which counting method could be said to better represent the collective will. These developments were later improved by Kenneth Arrow in the mid-twentieth century, who framed these notions of aggregation and fairness into axioms and theorems making use of the modern mathematical framework (Arrow, 1970). Thus, unlike most cognitive modelling approaches, social choice begins with a set of definitions of how a decision on the individual level is represented (this is called a vote) and a set of axioms that represent notions of fairness on the aggregation mechanism, which is submitted to analysis on how many of these axioms it is able to fulfill.[2]

We believe that this axiomatic structure, alongside the application mainly on political science and sociology, fields which are arguably further away from the general umbrella of cognitive sciences than psychology, has kept these notions out of the cognitive scientist's sight. It also does not help that the name "social choice" sounds to those more involved with neuroscience and cognition as something that directly models how an individual would go around with social situations (which is arguably better looked at from a game-theoretical and evolutionary point of view, like in J. M. Smith (1982)), and not the aggregation of multiple decisions that are already made. However, we also believe that the field of social choice has much to add to the cognitive sciences, mainly on two fronts: the collective and the neuronal. First, even though much has been studied on decision making on an individual level, with fascinating interplay of brain areas being revealed (De Martino et al., 2006; Martínez-Selva et al., 2006; Rolls et al., 2010), it is hard for researchers of cognition to scale up to the next level. Many times humans need to reach a

see Forstmann et al. (2016).

[2] For the mathematicians in social choice: it is common in cognitive modelling to start first with a (neuro)psychological phenomenon which is intended to be modeled and from there retroactively find mathematical structures that fit the data well, only then reaching an explanation and simplification phase of the models. This means that axiomatic structures rarely come into play.

consensus on a certain decision, and how that consensus is achieved is hard to boil down to mechanical equations that could model it. On the other hand, social choice has been dealing with the most stereotyped mechanism of such consensus reaching methods, namely vote casting. As rude as an approximation that might be to more subtle ways of collective decision making, we do not see why an attempt to model social psychology experimental results in a vote aggregation framework could not be made. It might be so that what seems like complex social interplay can be explained by an abstract analogue of vote aggregation rules that humans follow intuitively. Second, the word "vote" can make one instinctively think of a whole individual casting a vote, but as it has been seen in object recognition, the brain also can make use of distributed representation of stimuli throughout a neuron network, whose activation pattern as a whole determines the perceived percept (Ishai et al., 1999; Yamins et al., 2014). In this way, if the firing pattern of a specific neuron, that is, the frequency at which it sends signals to neighboring neurons, could be mathematically translated into a "vote", brain decoding methods utilizing social choice knowledge could be stipulated. This latter application is the one most sought after in this thesis, and how the mathematical methods appreaciated here could potentially be used for that goal will be a recurring topic. Thus, it seems to us to be a fruitful endeavor to bring both fields together, unlike it has been done to date.[3] In this way, the aim of this literature thesis is to bring a simple understanding of social choice as a field in mathematics to the world of cognitive science, and its scope is set as a presentation of key ideas in the field along with particular mathematical structures that are relevant in the field and could be applied in the brain sciences.

With the motivation for the exploration of the field of social choice set forth, in the rest of this introduction we will lay out the theoretical aspects necessary for the understanding of the work done in this thesis. All main mathematical notation is compiled in the preface. The goal of this brief summary is to clarify for those without the background specific to social choice the following research question which is tackled on this thesis: *what are the limits to strategyproofness for social choice functions? Are there domain restrictions which expand the total number of strategyproof social choice functions under them?*, and thus deatailed proofs will not be replicated in the main body of the thesis. The reader is advised to check the references if interested. This research question is useful for our goal of introducing social

[3] We did attempt to find published results on the cognitive sciences that made use of social choice methodology, to no avail. It is possible, however, that the fault is on our side or that this attempt has been made in the past but not publicized due to a negative result and publishing bias.

choice theory to cognitive scientists because it is a central question in social choice theory itself, still pursued today as more complex details are taken into consideration. How this question is still explored today should become clear throughout the thesis, but especially in Chapter 6, when we will have enough knowledge built up to guide the reader to more advanced readings.

1.2 Foundations of Social Choice

First and foremost, mathematical notation is used throughout this thesis. The preface has a list of all the notation used, so the reader is advised to refer to it whenever needed. Second, to limit the scope of the thesis, we will analyze only voting rules that return a (group of) winner(s) as a result, and not an ordering of preferences among the candidates, which is also a possibility studied in social choice. This type of voting rule we investigate here is called a *social choice function*, henceforth an *SCF*, and we start our introduction with its precise definition.[4] We then proceed to give some axiomatic properties SCFs ought to fulfill to be deemed fair (in a broad sense), and two important results of the field in the form of two theorems. The implications of the theorems are especially of interest to us, since they are the gateway to the exploration of how to get the so called *strategyproof* SCFs. All the definitions below are based on the definitions given in Brandt et al. (2016), and the theorems are directly reproduced from the same handbook.

Definition 1.1 (Social Choice Function, SCF). A *social choice function (SCF)* is a map $f : \mathfrak{L}(A)^n \to \mathfrak{C}(A)$, where A denotes the set of *alternatives* available in a ballot and n denotes the total number of *voters*.

Note that the use of linear orderings in the vote, as denoted by $\mathfrak{L}(A)$ implies that the voters must strictly order their preferences between alternatives, with no ties allowed inside one ballot. The combined set of all order of preferences is also called a *profile*. However, the range of the function is on nonempty subsets, as denoted by $\mathfrak{C}(A)$, which implies that a SCF might return draws between alternative candidates as a result of the *voting rule* and must also always declare at least one winner.

Now we turn to three properties that we expect fair SCFs to fulfill. In simple terms, we would like that there is a single winner, with no draws, that

[4] On the other hand, a voting rule which returns an ordering of preferences is called a *social welfare function* or *SWF*. However, as stated in the main text, these are left out of the scope of this thesis.

any candidate has a chance to be chosen as a winner and that no voter is better off lying about which candidate they prefer to get them elected. Thus,

Definition 1.2 (Resoluteness). A SCF is *resolute* if for all profiles it returns a set with only one winner.

That is, a resolute SCF does not admit any draws between candidates.

Definition 1.3 (Nonimposition). A SCF is *imposed* if there is an alternative x which cannot be elected, i.e., $\nexists P \,|\, f(P) = \{x\}$, where $P \in \mathfrak{L}(A)^n$. Thus, *nonimposition* means that there is always at least one profile under which f elects each alternative.

Definition 1.4 (Strategyproofness). A SCF is *strategyproof* if for any change in a ballot by a voter, the new outcome is not more preferred than the old one in the original ordering of the voter. That is, if voter i changes their ballot from $\succeq_i$ to $\succeq_i^{'}$, then $f(P) \succeq_i f(P^{'})$, where P is the old profile with $\succeq_i$ and $P^{'}$ is the new profile with $\succeq_i^{'}$.

This essentially means that if we consider the old vote to represent the genuine preferred result by i, then i cannot make such result happen more easily by changing the vote they cast. This type of behavior is called *strategic voting.*

Now, a characteristic that is not desirable but still important to define is a dictatorship.

Definition 1.5 (Dictatorship). A SCF is *dictatorial* if there is a voter i whose top-ranked alternative is always declared the winner under any profile P.

This voter i is called the *dictator.* A parallel can be traced to the common concept of a dictatorship, since only one agent determines the final result of the aggregation of the “collective” will.

Finally, with these definitions at hand it can be proven that:

Theorem 1.1 (Gibbard-Satterthwaite Theorem). *Any resolute, nonimposed, and strategyproof SCF for three or more alternatives must be a dictatorship.*

This theorem was first proven by Gibbard (1973) and Satterthwaite (1975), thus its name.

Theorem 1.1 shows that it is not possible to stipulate any aggregation rule that does not admit draws, can elect all possible alternatives and leaves no space for strategic voting or manipulation besides a dictatorship. Of course, a dictatorship by definition fulfills all three criteria, though it is not considered a fair voting rule, nor, in the case of an application of this methodology in the cognitive sciences, plausible to be a mechanism present in the brain. Consequently, this theorem implies that any SCF must abandon at least one of the axiomatic requirements we have set forth, and generally it is easier to let strategyproofness go. This means that the voting rule at hand will now be open to manipulation, that is, a voter will be able to make use of strategic voting in order to get a better result for themself, but at least every vote will count (i.e., it will not be a dictatorship). Nonimposition is harder to let go, since it ensures that every candidate has at least a chance to win, but depending on the situation it might be possible to consider a voting rule that is not resolute, if we are willing to accept more than one winner. This might be not an issue if we are considering the election of a committee, for example, but when precisely one person must be chosen for a position, or, hopefully in our future applications, a percept must be chosen to be acted upon, a resolute SCF is not only preferred, but also strictly necessary.

Concrete examples of voting rules will be given throughout the next chapters.

1.3 An Initial Parallel in Neuroscience

Now that a preliminary understanding has been reached, it would be useful to consider why the investigation of notions such as strategyproofness and manipulation could be fruitful for an application in neuroscience. First, as previously mentioned, our aim with this thesis is to introduce to cognitive scientists the field of social choice. A premise of this endeavor is that the firing behavior of a neuron could be mathematically interpreted as a ballot casting under this framework. This translation of neuron behavior onto votes could open doors to new ways of implementing cognitive models, from perception to decision making. This is mainly because the primordial task at hand in cognitive modelling at the neuronal level is to describe (and understand) how neurons can reach a consensus, given a determined input. In the framework of neural networks, this is modelled through a summation of the inputs multiplied by a weight, added to a bias factor and then passed through an activation function (Gurney, 2018). This process as a whole can be seen as an aggregation of pre-made decisions, and thus could be reviewed under the light of social choice.

It is clear then that different information processing patterns in the brain could be framed as a SCF. It is also valuable to note that one of the key findings about brain functioning is that different brain areas are connected to each other in order to modulate how they should process information in both bottom-up (motorperception into the higher faculties) and top-down (higher faculties into motorperception) directions (for example, Katsuki and Constantinidis (2014)). In this way, social choice applied to neuroscience could then be to think of these modulations as manipulations to a voting rule, and so the study of how to manipulate would be of much interest when thinking of combining these two fields. However, this line of research requires notions of computational complexity, which we deemed to be too advanced for an introduction. Instead, we propose that strategyproofness is parallel to safeguards that brain areas ought to have in order to sustain their typical functioning. This way, we can organize the presentation of new concepts here in a similar fashion that they are available in the social choice literature at the same time applications to neuroscience can be speculated.

As it will be later seen in this thesis, some domain restrictions do extend the quantity of strategyproof SCFs, allowing for non-dictatorial ways of robustly aggregating opinions. On the other hand, a strict adherence by the data to the properties needed to ensure it is contained in that restriction is often not realized. This is why there is also subfields in social choice which deal with the cases of partial or gradual adherence to these properties, exploring how vulnerable a class of SCFs gets as the profile data further deviates from the restricted domain (Bredereck et al., 2016; Chen & Finnendahl, 2018; Jaeckle et al., 2018; Lackner & Lackner, 2017). In this sense, it is possible to hypothesize brain functioning as a quasi-adherence to a domain restriction. This would leave open a door so that a limited amount of manipulation that is important for information processing is permitted. At the same time, this processing would not be vulnerable to chaos that might happen somewehere else in the brain. In this way, brain disorders could be framed as the amplified inability of a group of neurons to keep attained to a domain restriction, leaving them too open to manipulation by other areas. This kind of hyperinfluence could be key to understand brain activation patterns in disorders such as epilepsy, Parkinson's disease, and cerebral palsy, for example. Thus, understanding the underlying principles of strategyproofness is key in order to set up a productive interpretation of neuronal firing as voting.

1.4 Scope and Framework of the Investigation

It is then evident that the study of strategyproofness in social choice is fundamental, precisely because manipulations are mathematically proven to be unavoidable. There are, though, a few caveats. The first one being the position that Dowding and Van Hees (2008) claim that we should not worry about manipulations at all. This is mainly due to the observation that if a voter alters their vote according to the voting rule to achieve a result that they consider to be better than the original result that could be obtained by not changing it, then they are still being sincere on their preferences for an outcome. This means that although they manipulate the vote, such vote cannot be seen as insincere. The other observation that sustains the argument is that although manipulation may be taken to make the decision making process less transparent, the fact that the possibility of manipulation exists is in itself an incentive for voters to study how the decisions are made in the first place. This in turn could deepen the knowledge of the electorate over the processes adopted by the organizations they are part of, which can be seen as a democratic virtue.

Another point of analysis is how computationally costly a manipulation is, and how much information an agent would have to have in order to properly achieve its goal. In case of a sufficiently large number of voters, whose ballots are not known to voter i, the manipulator, how would they go about changing their ordering of candidates in the most optimal way for their interests? Even if the ballots were known, how complex would the computation be to determine this precise most optimal ballot? Would the time that it takes to churn this result out of a computer be considered worth the wait? Could it be made faster via a compromise where speed is obtained in sacrifice of optimality? Even worse, what if *every voter* behaves like a manipulator, how can the fact that voter j might change their ballot because they know voter i is doing the same be accounted for in i's decision making process? As it is possible to tell, if all computational details of a manipulation are taken into account, a succesfull manipulation by one or a group of voters can be made quite unattainable. And that is indeed the aim of a subfield of computational social choice: crafting voting rules which are not strictly strategyproof, but whose manipulation would be virtually impractical (Nisan, 2007; Procaccia & Tennenholtz, 2013). These computational and informational issues are, however, outside of the scope of this thesis. To concern here is to build a framework of classification of possibilities of strategyproofness under different domains, so that the main results of the field can be more easily assimilated by a newcomer cognitive scientist.

One additional aspect of a manipulation that can be taken into consideration is its doer. Are they the organizer of the election, a voter or an outsider? Traditionally, the literature calls the first type of manipulation *control*, and the last type *bribery*, leaving the term *manipulation* for manipulations that are done solely by voters. As it can be seen, *manipulation* can refer to both this restricted meaning or to the general concept which includes control and bribery. However, we have decided to leave matters related exclusively to control and bribery out of the scope of this thesis. This is because we believe that structures similar to control and bribery are not expected to be found in the mature brain. Control is often realized as the addition, deletion or partition of voters and candidates. Such a mechanism when taking the neuronal analogy we are discussing here into consideration would imply that a brain region could actively decide to either temporarily suppress or destroy and reform connections with other areas on the fly. This type of behavior is yet to be observed in the brain. On the other hand, the process of synaptic pruning[5] could probably be modelled through control-like structures, but this would lead us to investigate the maturing brain as well, complicating matters far beyond an introductory level for both the cognitive scientist and the mathematician. A similar issue is raised with bribery. Since it involves a third agent convincing that a voter should change their vote, it would imply either a significant modulatory effect of an area onto another in an overly complicated manner for an introduction or the effect of a technology such as TMS.[6] Thus, we decided to focus primarily on the issue of strategyproofness of manipulations in its most general sense, without reaching deeper into exclusive forms such as control and bribery. The reader interested in these two special types of manipulation can refer to Bartholdi et al. (1992), Hemaspaandra et al. (2007) and Hemaspaandra et al. (2009) for control and Faliszewski et al. (2009) and Elkind et al. (2009) for bribery.

Finally, the way we organize our investigation of strategyproofness in social choice and its applicability to the neurosciences is through the compilation of different results obtained through *domain restrictions*. A domain is the set of all profiles given a set of alternatives and a number of voters that fulfill a predetermined condition on what can be present on a ballot

[5] For the mathematician, *synapsis* is the name of the gap between two neurons through which electrochemical communication takes place. Thus, *synaptic pruning* is the process of getting rid of a synapsis that previously existed, disconnecting two neurons / areas. It is a common phenomenon throughout childhood and brain development (Sakai, 2020).

[6] Transcranial magnetic stimulation. Basically, a magnetic field is applied near to the region of interest, changing its behavior. This is a non-invasive technique and is also used for research on humans (Pascual-Leone, 1999).

(i.e., what is a possible vote). When such condition is well elaborated to mimic plausible outcomes of different types of decision making, restricting the analysis to a single domain can be beneficial, since this extra property that the profiles must abide to can be used to simplify or deepen the calculations around it. The domains we will be taking into consideration are the unrestricted, single-peaked, single-crossed, and separable preferences domains. These domains were chosen because they are the most relevant ones in the field of social choice theory for beginners. In this way, each of the following chapters will take a closer look on what are the typical motivations and results for each domain in social choice, followed by some guidance for papers one might want to read for a deeper understading from the mathematical perspective, and concluding with how these concepts could possibly represent some form of neuronal processing.

We hope that by the end of the thesis, it becomes clear that the use of social choice mathematics shows enough ground to expect it to provide a path for improvement for our current cognitive modelling methods, hopefully elucidating more of the mysterious working of the brain.

Chapter 2

Unrestricted Domains

To begin our analysis of strategyproofness in social choice and how the notions associated to the topic can be applied to cognitive science, we must first specify what is meant by *domain restriction*. Remember that the core of social choice is SCFs, which are mathematical functions, and thus by definition they have a *domain* and a *range*. The domain of a function is the set of values that are accepted as input to that function, and the range is the set of values to which the output of a function belong. In the case of SCFs, the domain is the profile of ballots that voters cast, while the range is non-empty sets of alternatives, meaning that ties are allowed but every election must have at least one winner. Thus, when we refer to domain restriction, we refer to the specification of a property to which the profiles that are accepted as an input must adhere. This does limit the scope of what a SCF could handle as a valid vote, but if the restriction is adequate, not only can it mimic actual informational patterns we might expect to occurr in the natural world, but it can also expand the number of strategyproof SCFs under that restriction.

Remember that without any kind of domain restriction, the Gibbard-Satterthwaite Theorem holds:

(Repost) Theorem 1.1 (Gibbard-Satterthwaite Theorem). Any resolute, nonimposed, and strategyproof SCF for three or more alternatives must be a dictatorship.

We already saw in the introduction that this means that the only way to implement a SCF that does not return draws, can return any alternative as an output and is not open to manipulation is through a dictatorship. In summary, not only is the general, unrestricted domain so broad as to give little information on what kind of structure we could expect to see in voting profiles, it also does not allow us to move away from dictatorships if we are looking for strategyproof SCFs.

Nevertheless, it is one of the premises of the scientific endeavor that our universe is ordered. In this way, it can be studied in order for us to understand what are its underlying structures that we can so easily miss without the tools of science. When social choice is applied into political science, for example, we could assume that each person votes randomly into an ordering of candidates, and thus the domain would remain unrestricted. However, it can easily be argued that that is not the case. Voters are assumed to have intent and preferences when casting a ballot. Not only that, in the real world, many other factors come into play. Political advertisement, local aggregations of like-minded individuals, fear of the unkown and belief in social causes all contribute to an individual's choice. This makes so that the ballots one individual may cast adhere to a certain structure that reflects both society and the perception of that individual of society. The same can be said of neuroscience.

If we intend to make social choice theory a new basis for cognitive modelling, then we must understand that domain restrictions are not only essential for practical reasons, but interesting for theoretical ones as well. Let's take vision as an example. Vision is said to be an ill-posed problem, since we must deduce what exists in a three dimensional world out of a two dimensional figure that is formed in our retina (Bertero et al., 1988). In order for that to be possible, natural selection found heuristics that are now known to shape visual information processing (for an example in vision itself see Cohen (2006), for an overall explanation of this principle in cognition in general, see Gigerenzer (2004)). These rely on assumptions about the world that are based on the laws of physics. For example, under darker contexts, even lighter colors assume a darker tone, sometimes even looking physically the same as darker colors under lighter contexts. However, having this knowledge about the world, the brain is able to counteract this and perceive color in a constant way. This is known as color constancy, and a classic example is reproduced in Figure 2.1 (Foster, 2011). This additional information that is not present in the stimulus itself, but is rather a restriction imposed on the calculation is what makes us able to adequately perceive the world in most contexts.[1] In a similar fashion, domain restrictions in social choice come into play to give order to a voting profile generating system (a collection of neurons) because they must either represent an ordered structure of the real world or implement a computatinal heurestic that allows us to navigate the world. Restrictions are a limitation to the scope of the domain of a function,

[1] The exceptions being natural environments with noisy characteristics, such as poor lightning, or artificial stimuli made to fool us, like optical illusions (Todorović, 2020).

but definitely not a limitation to the scope of scientific investigation. They are what will make this attempt possible.

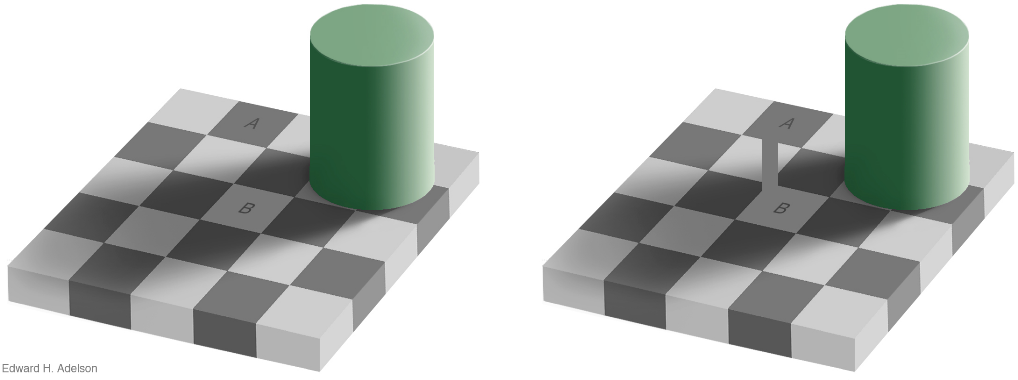

Figure 2.1: A classic example of color constancy, adapted from https://persci.mit.edu/gallery/checkershadow . While both areas A and B have the exact same hue in RGB values, which can be checked on the right image where a rectangle was plotted over the original, they are captured by our visual system as two different colors. This is because of the lighter/darker context around each tile, prompting the use of the color constancy principle by our brain to process correctly this stimulus.

Yet, before we consider restrictions on ballots and profiles, it would be productive to first consider how neuronal spiking patterns could be translated into votes in the framework of social choice. As seen, the core of social choice theory are SCFs that compile a series of votes into a decision. This means that this type of modelling is most suited for computations that require some kind of judgement, be it sensorial, as in determining what the impinging stimuli consist of, or behavioral, as in deciding what should be done next. In this context, it is stipulated that there is an area in the brain which implements this aggregation process. This does not mean that there is only one area that does that, only that for each type of judgement that must be made, there is an area that does it. It is certainly different for each modality of the senses as well as for more complex decision making, such as behavioral control. This notion of judgement can also be broken down or scaled up, as the level of analysis changes. For example, before determining that the image reflected on the retina is that of a dog, the brain must be able to determine that smaller patches of the dog belong to the same object, and are not part of the background. As for behavioral control, before action path A or B is chosen, the brain must be able to weight which one would be more beneficial for the individual considering its current state and goals. This calculation requires information such as how costly to an individual these two action paths are, and just the compilation of different cost factors into a final cost judgement is already a subdivision of the larger behavioral problem to which social choice theory could be applied. Thus, a brain region whose function is to make a certain judgement based on input from other regions is what social choice theory is best suited for. In this way, the raw brain data that must be somehow mathematically converted into votes are the activation patterns observed in the areas that serve as that input.

Of course, one of the major challenges of neuroscience is that different brain recording methods have different trade-offs. If our goal would be to translate single cell spiking into votes, then we would be restricted to invasive methods that can only be used in animal studies. However, it would be simple to realize votes as different firing frequencies to different stimuli. On the other hand, human research utilizing EEG and fMRI deals with much more elaborate signals, that represent the activity of whole populations of cells over longer periods of time. In this sense, event related potential[2] components and their intensity of EEG could be taken as a vote, for example. The complexity of this approach is that in any case, it would rely heavily on a task by task, case by case modelling. This is because the social choice framework requires that a set of candidates is pre-established before the SCF is run and an output is obtained. This means that for clear, and especially prototypical modelling, a set task, with clear expected outputs is needed. Fortunately, the majority of the cognitive psychology literature is already framed in this fashion, and this new approach would suit well to be tested with pre-existing methods.[3]

Thus, with this thesis, we aim to provide some general ideas on how each domain restriction discussed further could come up in the brain. We strive to excite the creativity of the cognitive scientist in their field and provide fertile soil upon which new ideas could be tested. We also give examples of why such a domain restriction was thought of inside of social choice and its relevance, in hope that this can further incentivize the reader to look beyond their subfield in the cognitive sciences.

[2] The change in voltage of an EEG signal that is recorded in relation to an experimental event.

[3] For the unfamiliarized with different types of brain recording and cognitive psychology tasks, Chapter 3 of Gazzaniga (2019) is a great introduction.

Chapter 3

Single-peaked Domains

One of the most relevant domain restrictions in social choice is single-peakedness. This domain is characterized by having the candidates lie on a linear scale, on which voter preferences must be single-peaked. I.e., between two candidates that lie both to the left or both to the right of the most preferred alternative, the one closest to the most preferred alternative must be preferred over the other. Such a domain restriction is useful in political science to model left- and right-wingedness, for example. Formally speaking (Ballester & Haeringer, 2011):

Definition 3.1 (Single peaked ballot). A *single peaked ballot* $\succeq_{i'}$ is one that, given that the candidates A are linearly ordered and the true preferences $\succeq_i$ for voter i over candidates A orders them in such a way that for any three candidates a', a'', and $b_i(A)$, where $b_i(A)$ represents i's most preferred candidate, $b_i(A) \succeq_{i'} a' \succeq_{i'} a''$ or $a'' \succeq_{i'} a' \succeq_{i'} b_i(A)$ implies $a' \succeq_i a''$.

This definition says that candidates should be able to be identified as lying on some kind of spectrum or range, so that some order between them can be established (here, the order does not mean preference; you could, for example, assign both left or right wing politicians as lower or higher on the scale and both dispositions would lead to the same mathematical results). It also states that wherever in the ballot $b_i(A)$ appears, a further away candidate from it in the ordering is always least preferred over a closer candidate when they both rank higher or lower than $b_i(A)$ in that ballot. This would be a natural assumption to make in a political example if one assumes that a voter who sees oneself as leftist would prefer a politician from the center rather than one from the right. The name single-peaked thus becomes clear if one visualizes the rankings in i's true preference as scores plotted over an axis that orders the candidates in $\succeq_{i'}$. It also becomes clear that such a profile assumption is sound if we assume that the candidates in the election represent different stances that lie on a one-dimension spectrum. For a graphical example of single-peaked preferences, see Figure 3.1.

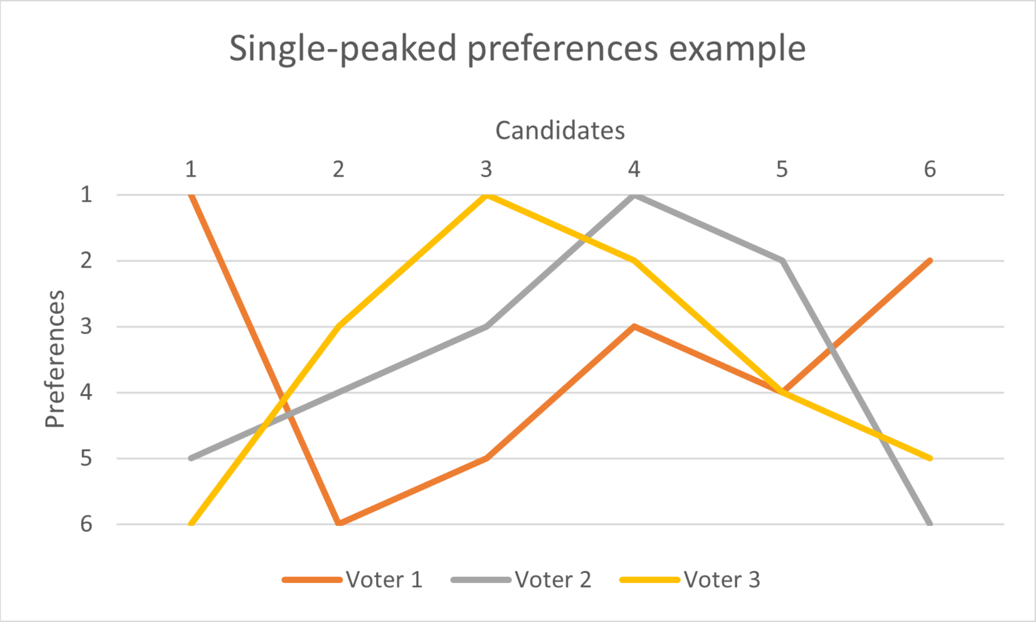

Figure 3.1: The preferences of voters 2 and 3 are single-peaked, while that of voter 1 is not.

Single-peaked domains are relevant to social choice literature because it has been shown that there are non-dictatorial SCFs that become strategyproof under this domain restriction, namely *generalized median voter schemes* (Arribillaga et al., 2020; Barberà et al., 1993). However, to achieve the definition of this class of SCFs, some other concepts must be first laid down. The following definitions were adapted from Arribillaga et al. (2020).

Definition 3.2 (Coalition). A *coalition* is a subset of voters. Thus, if n is the total number of voters and $N = \{i | i \in \mathbb{N}, 1 \leq i \leq n\}$, then a coalition is a member of 2^N.

A coalition is a straightforward concept: simply a set of voters in an election.

Definition 3.3 (Committee). A *committee* is a non-empty family C of non-empty coalitions that follows the following monotonicity restriction: given $S, T \subseteq N | S \in C, S \subsetneq T$, then $T \in C$.

On the other hand, a committee is a bit more complicated. Mathematically speaking, a family is a set whose members are sets. Therefore, a committee consists of different coalitions, which would represent different groups of voters with a certain tendency. Note that this definition implies through the monotonicity constraint that there exists for every committee a family of coalitions that define the committe. This constraint ensures that if a coalition is present by itself in the committee, any other coalition that includes every member of that first coalition must also be in the committee. For example, if $\{i_1, i_2\}$ is a coalition inside a committee, $\{i_1, i_2, i_3\}$ must also be. Just like committees in real life are organized to make decisions and conduct matters, this definition comes into play as a decision making body in the next definition.

Definition 3.4 (Coalition system). A *coalition system* is a monotonic family of committees $\{C_a\}_{a \in \mathbb{N}_A}$, where $\mathbb{N}_A = \{a | a \in \mathbb{N}, 1 \leq a \leq \#\{A\}\}$, where $\#\{A\}$ is the number of candidates available in the ballot. In this way, in a coalition system there is a committee assigned to each candidate in an ascending order, and its monotonicity rule is (i) $\forall a, a' \in \mathbb{N}_A | a < a', S \in C_a \to S \in C_{a'}$ and (ii) $C_{\#\{A\}} = 2^N \setminus \{\emptyset\}$.

Thus, a coalition system represents a build up of committees that would prefer a candidate at some point in the ordering of the candidates or lower. In this way, it is possible to count how much support a candidate has from each group of voters and keep track if a majority has been mustered already at some point or not. This is the key concept behind a generalized median voter scheme. Basically, a generalized median voter scheme can be viewed as a SCF where you have candidates disposed on an ascending order which would be the "checking" order for the function. The function then checks for each candidate if at that point cumulatively the coalitions of voters have gathered enough "strength", represented by their top alternatives being the one at hand or a previous one, to stop this checking process. The candidate whose number passes this "strength check" is then elected.

Definition 3.5 (Generalized median voter schemes). A SCF f is a *generalized median voter scheme* if there exists a coalition system $\{C_a\}_{a \in \mathbb{N}_A}$ such that for all profiles $P \in P^N$, $f(P) = a$ if and only if (i) $\{i \in N | b(P_i) \leq a\} \in C_a$ and (ii) $\forall a' < a, \{i \in N | b(P_i) \leq a'\} \notin C_{a'}$.

This is an admittedly hard to grasp concept. It is associated to the regular median in the sense that it returns as an outcome the candidate whose associated number $a \in \mathbb{N}_A$ is the smallest one where the numbers associated to the most preferred alternatives of all voters of the winning coalition at a are smaller than or equal to a. This is parallel to the median because it returns the smallest number in the dataset which cumulatively represents more than half of the data points.[1]

Lastly, the central result on strategyproofness in single-peaked domains:

Theorem 3.1. Generalized median voter schemes are strategyproof in single-peaked domains.

As stated in the Introduction, the proof of why these functions are strategyproof under single-peaked domains is ommitted, and the reader is advised to refer to the references if interested. The seminal work on this matter is Moulin (1980), who first proved this through other means than the definitions above. Another interesting read is Blin and Satterthwaite (1976), which shows that voters must be aware of the restriction placed by Definition 3.1 and actively structure their vote in that way to guarantee that the resulting profile will be single-peaked.

Moving further, although it is good news that there are non-dictatorial SCFs that are also strategyproof under single-peaked domains, is it feasible to expect that voting profiles show this property on actual data? Lackner and Lackner (2017) and Karpov (2020) have done combinatorial analyses under different modelling assumptions and came to the conclusion that single-peakedness is highly unlikely, specially as the number of voters and candidates grows, with its likelihood tending towards 0. This could be inferred to be so because matters tend to be not as simple as to have dispositions naturally laying in a one dimensional scale. However, on more practical matters, List et al. (2013) shows that deliberation of issues that are deemed not so polemic as to have had intense discussion before the experimental setting makes the voting profile come closer to single-peakedness than votes cast before the deliberation, especially if there is a clear one dimension with a left-right orientation to the issue. This is a relevant finding because the application of generalized median voter schemes to the talling of those votes is tempting not only because of strategyproofness, but also because due to its resemblance to the median, it can be seem as a promising 'fair compromise' from each side of the spectrum. On the other hand, Penn et al. (2011) is interesting because it gives an example of manipulation happening on single-peaked domains structured to resemble current political decision making bodies. Incrementally, Barberà et al. (2010) shows that the gener-

[1] Namely, $min\{x|\#\{i|i \leq x\} \geq k+1\}$, where # indicates the number of elements in the given set and $n = 2k+1$ is the dataset size.

alized median voter scheme is strategyproof even if a whole group of voters try to manipulate the election together. For a broader analysis on single-peakedness and strategyproofness, the reader is advised to refer to Barberà (2001). A deeper insight can be obtained through Nehring and Puppe (2007).

However, what could single-peakedness mean for the brain sciences? Is it plausible to expect a single-peaked domain in any area of the brain? This question seems even harder to solve taking into account that combinatorial analyses have deemed such profiles to be unlikely and work from Blin and Satterthwaite (1976) has shown that such a restriction must be kept in mind by the voters at all times to assure its realization. Remembering that of interest to us is also a quasi-adherence to the restrictions by the neuron populations rather than a strict one, since a bit of manipulation is useful for interarea communication, we would argue that, yes, it is plausible to expect single-peakedness in the machinery of the brain. The argument for this presence is twofold. First, previous analyses were not concerned about the specific patterns observed on the brain, which are also not random and thus highly unlikely under non-specialized combinatorial analyses such as the ones mentioned before. The same principle applies to the result by Blin and Satterthwaite (1976), since neurons can have been naturally selected to 'vote keeping in mind' the restriction of single-peakedness. Second, single-peaked preferences can be visualized as a peak of preferredness over a linear ordering of the alternatives, which resembles bell shaped distributions, such as the normal distribution. A similar distribution of firing frequency is observed in the primary visual cortex, for example (Hubel & Wiesel, 1962; Ringach, 2002). When a neuron is exposed to a stimulus it is adjusted to detect, such as the inclination of a line segment at a specific angle, it fires with the highest frequency possible. Then, as the angle deviates from its most preferred, it fires with progressively lower frequencies, returning to base level at a certain point where the angle is deemed to be not similar enough to elicit a reaction. This pattern can easily be interpreted as single-peaked ballot over different (angle) alternatives. A similar behavior is observed for other visual characteristics, as well for other perceptual modalities (Bandyopadhyay et al., 2010; Conway, 2009; Rothschild et al., 2010). In this way, it is plausible to expect profiles which are close to single-peakedness on neuronal data at least on primary sensory cortexes, as well as in any other neuron population which shows a bell curve firing pattern. Regarding vision, specifically, an open question in cognition is how different ensembles of properties are bound together to form a cohesive whole as a recognizable object (Hummel, 2001). Considering that generalized median voter schemes resemble the median, which is a mathematically robust aggregator, since it is hardly altered by extreme values, such SCFs should also be able to capture robustly the signal within the firing of neuronal populations. Thus, perhaps such SCFs could be applied in an aggregation process of percept characteristics, functioning well even on dubious and noisy contexts. This hypothesis could be initially tested with data analysis on figure-ground dissociation tasks done by monkeys under single-cell recording procedures, which is already available (Yamane et al., 2020).

Chapter 4

Single-crossed Domains

Another relevant domain in social choice theory is the single-crossed domain. In the last chapter we saw a domain where the candidates had an ordering and the ballots of the voters had to follow a certain restriction according to that ordering. Now, we will think that not only the candidates, but the voters are ordered as well. To make this possible in a way that will not interfere with the main result of this chapter, we will consider only the top candidate $b_i(A)$ of each ballot when ordering the voters. As a practical example why one would be insterested in an ordering of the voters, we can take the left to right political spectrum once more. It might be the case that some interesting mathematical properties can be deduced for this type of analyses if we are able to order the voters on how left or right wing they are based on their most preferred candidate. In this way, we start with the necessary definitions for this domain, all adapted from Saporiti (2009). Another reliable introductory paper to single-crossed domains is Saporiti and Tohmé (2006).

Definition 4.1 (Single-crossedness). Let $\mathfrak{P}$ be all possible profiles for a given set of candidates A ordered by $>$. A profile $P \in \mathfrak{P}$ is *single-crossing* on A if it is possible to order the voters $\succeq_i \in P$ through a liner order $\succ$ such that $\forall a_1, a_2 \in A$ and $\forall \succeq_{i_1}, \succeq_{i_2} \in P$: (i) $[a_1 > a_2, \succeq_{i_1} \succ \succeq_{i_2}, a_1 \succeq_{i_2} a_2] \Rightarrow a_1 \succeq_{i_1} a_2$, and (ii) $[a_1 > a_2, \succeq_{i_1} \succ \succeq_{i_2}, a_1 \succeq_{i_1} a_2] \Rightarrow a_2 \succeq_{i_2} a_1$.

This definition says that if you can order the voters in such a way that a higher-ranked voter on $\succ$ will always rank a higher-ranked candidate on $>$ at least as high as or higher than a lower-ranked voter on $\succ$ would and vice-versa, then you get a single-crossed domain. In a political example, this would mean that a voter that leans more to the right than another one will prefer a candidate that is at least as right-wing as the one that the other voter voted for if not more (and vice-versa for left leaners). A graphical example is given in Figure 4.1.

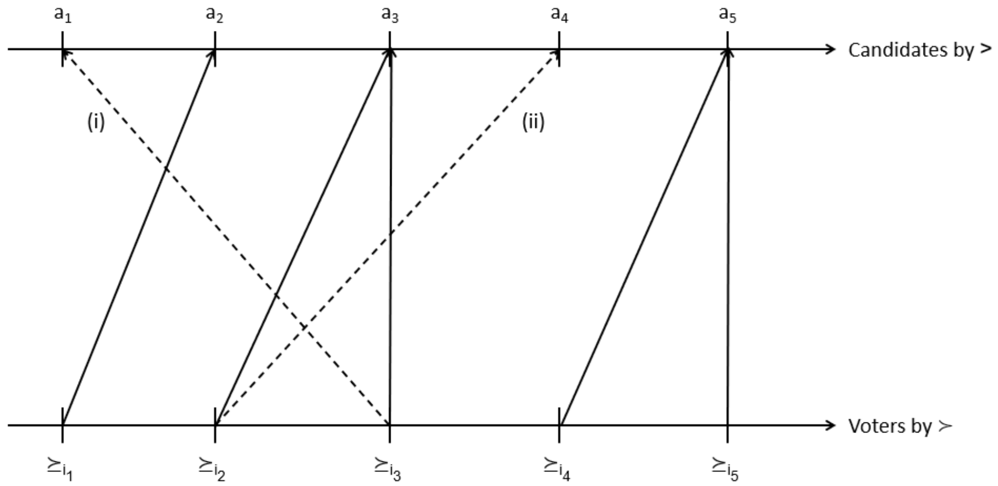

Figure 4.1: Example of a single-crossing profile and violations of conditions (i) and (ii) from definition 4.1. The top line orders the candidates in A by $>$, and the bottom line orders the ballots $\succeq_i$ in P by $\succ$. An arrow from the bottom to the top line indicates the most preferred candidate in a ballot (only the top candidate is taken in consideration in this example for simplicity). All solid lines taken together show a possible single-crossing profile on A, while each of the dashed lines represent a violation in conditions (i) and (ii) respectively when that dashed line is assumed to be the new vote cast by that voter keeping all other votes constant.

Note that it is not strictly necessary to restrict the ranking of voters to consider only their top preferences in order to achieve single-crossedness. However, not only does that make the analyses easier, it will also be essential for strategyproofness. Nevertheless, Bredereck et al. (2013) provides a method to determine if a given profile satisfies single-crossedness or not, which is not trivial. Following on the definitions, when a SCF only considers the top alternative in a ballot for its functioning, it is said to be *tops-only*. This is an interesting property for a SCF to have because it means that less information from each ballot is needed to compute the winner(s).

Definition 4.2 (Tops-only). A SCF f is *tops-only* if for any profiles P and P' where $b_i(A)$ is the same for every i on both profiles, then $f(P) = f(P')$.

A tops-only rule is also interesting for the voter, because it means that they are only required to signal who they think is best instead of having to rank every candidate in a specific order. Going further, before we get to the main theorem of the chapter, let us define two other properties of SCFs in general. Those will also appear in the main result.

Definition 4.3 (Unanimity). A SCF f is *unanimous* if $\forall a \in A$ and $\forall P \in \mathfrak{P}$ such that in every vote $\succeq_i$, $b_i(A) = a$, then $f(P) = a$.

This basically gives the mathematical definition that one would expect from the word unanimous: if all of the voters rank one candidate on top of their ballots, then that candidate is elected. This property can be seen as a minimal responsiveness to the ballots cast.

Definition 4.4 (Anonymity). A SCF f is *anonymous* if $\forall P, P' \in \mathfrak{P}$ where P and P' differ only in the labeling of each voter but contain the exact same ballots, then $f(P) = f(P')$.

This means that if the only difference between two profiles is the order by which the votes are cast, which can be seen as an attribution of who casted which vote, then an anonymous SCF would elect the same candidate. One of the implications of this property is that the SCF is not a dictatorship, and that voters have an equal say amongst themselves.

We now follow with the definition and examples of the SCFs that will end up in the main theorem of this chapter.

Definition 4.5 (Extended median rule). A SCF f is an *extended median rule* if there are $n+1$ fixed ballots with tops $a_1, a_2, ..., a_n, a_{n+1} \in A$ such that for $\forall P \in \mathfrak{P}$, $f(P) = med(b_1(A), b_2(A), ..., b_{n-1}(A), b_n(A), a_1, a_2, ..., a_n, a_{n+1})$, where *med* represents the median when A is ordered by $>$.

At first this artificial introduction of $n+1$ ballots might seem strange, but it is done in order to guarantee that there will be an odd number of votes for a well defined median. If one adds (about) half of the votes for the lowest candidate in $>$ and the other half for the highest, then the standard median is achieved, with ties in even numbers of voters being broken in favor of the lower or the higher ranked candidate around the center of the profile according to which candidate (the lowest or the highest) got the extra fixed ballot in $n+1$. Notice that, by definition, extended median rules are tops-only. Using a similar logic and restricting which candidates might be chosen for $a_1, a_2, ..., a_n, a_{n+1}$, we get *peak rules*, which are also tops-only.

Definition 4.6 (Peak rules). A SCF f is a *peak rule* if there are $n-1$ fixed ballots $a_1, a_2, ..., a_{n-2}, a_{n-1} \in \{b_i(A) | 1 \leq i \leq N, \succeq_i \in P\}$ and

$$f(P) = med(b_1(A), b_2(A), ..., b_{n-1}(A), b_n(A), a_1, a_2, ..., a_{n-2}, a_{n-1}).$$

Notice that peak rules have fixed ballots cast only for candidates that have already been voted for by an actual voter.[1] With this, we can state:

Theorem 4.1. A SCF f is unanimous, anonymous and strategyproof in a single-crossing domain if and only if it is a peak rule.

Again, proofs are left out of the main text (but can be found in Saporiti (2009)).

It is also relevant to mention that a domain is not limited to having but one restriction. As long as the restrictions are not mutually exclusive, a new domain restriction can be defined from their combination. For example, Elkind et al. (2020) characterizes domains that satisfy both single-peakedness and single-crossedness at the same time. In fact, in the specific case of these two domains, they can also be interpreted as lesser instances of more general restrictions, as proposed by Barberà and Moreno (2011). Another consideration we had in mind for the application of social choice theory to neuroscience is that it is probably good for interarea communication in the brain that some degree of manipulation is still allowed, and thus an incomplete adhesion to a restriction would probably suit its machinery better than a strict one. Fortunately, for the case of single-crossedness, Jaeckle et al. (2018) offers a method to measure how deviant a set configuration is from single-crossedness through various distance metrics. These could be used in cognitive modelling as an assesment tool of robustness in functionality of the group of neurons at hand. Finally, as discussed with single-peakedness as well, there have been combinatorial analyses on how likely single-crossing domains are to emerge (Bruner & Lackner, 2014). Once again, they seem to be quite unlikely, especially as the number of voters and candidates grow. However, the same point stands that this takes no innate configuration shaped by evolutionary trial and error at all, and thus does not hinder the purposes of this thesis to suggest the development of new toolboxes for cognitive scientists.

Moving on, for single-peaked domains, we had only to consider a 'natural' ordering of the alternatives. In contrast, now, when looking on how to apply the mathematics behind single-crossed domains to the brain sciences, we must also keep the voter side of things in mind. Considering that we are assigning ballots to represent firing patterns of neurons, an ordering of

[1] Notice also that instead of adding $n + 1$ fixed ballots, $n - 1$ fixed ballots are added. This is done in order to achieve what is called *Pareto efficiency*. This, however, falls out of the scope of this introduction and the interested reader is advised to refer to Saporiti (2009).

voters could represent different degrees of intensity of a stimulus associated to a specific neuronal population. To avoid being repetitive and consider only perceptual processes, this time we will propose an application in the neuroscience of learning.[2]

It is believed that learning takes place in the brain through the processing of prediction error signals between the expected outcome of an action and the observed outcome (Daw & Tobler, 2014). These signals are deemed to be implemented in the brain through the activation of dopaminergic neurons. Studies utilizing fMRI techniques have found that that activity can be captured in the BOLD signal[3] recorded from the striatum,[4] which seems to be a central processing station for reward and overall motor and cognitive control. Thus, it is plausible to expect to find a population of striatal neurons who would interpret the signal from dopaminergic neurons and assign different levels of positive or negative motivation to execute certain behaviors. As such, initially in an experimental setting, where the response repertoire is restricted, we could test whether an almost single-crossed profile can be found in the neuronal chain that processes decision-making.

To make this example more concrete, we will first go over a very simple experiment akin to those used in the cognitive psychology literature for decision-making and learning (akin to the one in Miletić et al. (2021), for example). Suppose different symbols are shown in a screen to a participant, each associated with a certain probability and intensity of reward. In this case, points that would be converted into monetary recompensation by the end of the experiment. However, some symbols are also associated with a certain probability and intensity to losing points. These are not previously disclosed to the participant. Every trial the participants are presented with a set of symbols, and they must pick one as their response. It is expected that through the trials they will learn which symbols are better than the other in whatever way they might optimize their responses, be it through small but certain rewards or large and uncertain ones. A possible model for the learning of this association could be implemented via social choice theory.

As per usual, we are postulating the neuronal firing patterns can be trans-

[2] For the mathematician, it must be clarified that in typical cognitive psychology fashion, learning denotes a sustained change in behavior, especially towards a specific stimulus, and is thus a broader concept than the common usage of the word "learning" in daily life.

[3] Blood-oxygen-level-dependent signal. This is what is actually recorded in fMRI scans. This signal is based on the change of magnetic properties in the blood around a brain area that has been active. For more details, see Drew (2019).

[4] A subcortical are of the brain that is involved with motivation, movement and other functions. Its name is due to the striped pattern that can be observed in it due to alternated segments of gray and white matter. For more, see Graybiel and Grafton (2015).

lated into votes. In this case, the alternatives would then be which symbol to choose. Arguably, there are good outcome and bad outcome symbols, and they can be ordered in a line of how good they are through an utility function, for example. The ordering of the voters could then be on a processing hierarchy basis. Considering the regular flow of information being from bottom to top in decision-making related areas, we could align the recorded neurons according to area, from the mid-brain, through the striatum onto the prefrontal cortex. It is expected that neurons higher up in the processing chain have access to more information than the ones lower down and are thus able to make better comparisons among the shown stimuli. In this way, we could model this interaction as a single-crossed domain, where neurons further up in hierarchy vote for higher utility symbols. This configuration of social choice theory would allow us to delve into more precise questions regarding the decision-making per se, rather than only speed and accuracy as it has been traditionally done. For example, the modelling of a SCF that compiles the neuronal information into this behavioral choice would give us new variables to explore correlations with brain activation. Not only that, this close to single-crossed conformation can give us insight on the robustness of choice, since we can model a strategyproof peak rule and test how suitable it is for different populations. It is not too farfetched to expect it to fit worse for people with decision-making related disorders, such as gambling disorder and substance use disorder (Grant & Chamberlain, 2015; Potenza et al., 2019; Tarter & Vanyukov, 2001). In this case, one possible explanation for the mechanism of the disorder could be that either predisposition to or addiction itself flips the voting profile around, breaking the single-crossed assumption that would guarantee regular functioning of these areas.

More generally, once these bases are more solid, a cognitive model based on the single-crossed domain could be used in non-experimental contexts. If a broader range of attitudes and behaviors could be assigned a good and bad moral orientation, then ethical processing could be implemented in a similar way as previously described. Of course, this assumes that more complex social interactions are stored in the mind as some kind of symbol in a "language of thought", which might not be how the brain works in the first place (Ayede, 2010; Fodor, 1975). But, we, as cognitive scientists, should be open to trying new methods whenever feasible, and social choice theory demonstrates a solid background to work upon.

Chapter 5

Domains with Separable Preferences

The last important domain in social choice that we will analyze in this introduction is the so called separable preferences domain. This domain is characterized by having voters vote for their preferred set of candidates, instead of one single candidate (or an ordering of subsets of all possible combinations of candidate sets instead of an ordering of singleton candidates). Furthermore, each voter in this domain must judge each candidate as good or bad in their view. This is mathematically implemented as the answer to the question: "Would you rather there was no winner at all instead of this candidate being the sole winner of the election?". Thus, unlike the two previous domains where we thought of orderings of candidates and voters, this time we will look at each voter separately. We will also take a step back and redefine what is considered an alternative for the SCFs we will investigate in this chapter. So far we have considered a set of candidates A, which was the same as the alternatives, where voters' ballots $\succeq_i$ were linear orderings over A. In this chapter, however, we will consider objects belonging to a set O, which would be equivalent to candidates in an election. These are the singleton candidates over whom the voters will have to answer the aforementioned question about them being a good or bad candidate. However, we are now interested in posing the power set 2^O as the alternatives whereover a ballot would be cast, thus serving the role of the alternatives set. This will give all the possible subsets of candidates that could be elected as winners of the election. In this manner, in this chapter, the A we will take into consideration is a family instead of a simple set, and namely, $A = 2^O$. Since, after all, now candidates and objects are the same, while alternatives and sets of objects are the same.[1] We will do this precisely because the separable preferences domain is based on the notion of good and bad candidates that represent what the voter thinks of each object. The point of having separable preferences, then, is to list groups of candidates in a way that a voter would be more satisfied with a result that has more good candidates. This can be expected in a political setting, for example. Say we have a leftist voter that has to give their say over the formation of the parliament. Then, that voter would prefer a parliament with more left-leaning politicians over one with more right-leaning ones. So, to this voter, the goodness of the subset of candidates at hand should improve by the addition of a left-wing politician. This voter preference ranking that takes into account improvement of the subset by the addition of a good candidate is what constitutes the core of separable preferences. As expected, this *preference* that each voter has is a technical term as well. It is the equivalent of a ballot listing subsets instead of singleton elements in a linear ordering. The definitions in this chapter are adapted from Barberà et al. (1991).

Definition 5.1 (Preference). A *preference* associated to voter i is a linear order $>_i$ over $A = 2^O$ where O is the set of all objects. This represents how pleased a voter would be with a particular result of an election. Note that the empty set is also placed somewhere in this linear order, meaning that the voter would prefer that there were no winner appointed rather than having that particular set of objects as the winners.

Since both preferences and the ballots up so far are linear orderings, both can be used to make a profile that serves as input to a SCF. On top of that, though, we will use this concept of preferences to analyze how to inform a SCF to be strategyproof under what is called the *separable preferences* domain. To categorize it, we need the notion of *good* and *bad* candidates. Notice that in the other chapters the words "candidates" and "alternatives" were mutually interchangeable, which is not the case in this chapter. Here, "objects" and "candidates" are interchangeable, while "alternatives" refer to sets of objects.

Definition 5.2 (Good and bad candidates). For each voter i exhibiting preference $>_i$ over A, a candidate $o \in O$ is said to be *good* according to i if $\{o\} >_i \varnothing$. Alternatively, a candidate $o \in O$ is said to be *bad* according to i if $\varnothing >_i \{o\}$. The set of all good candidates according to i is represented as $G(>_i)$, with the complement $G^c(>_i)$ being the set of all bad candidates

[1] Thus, in this chapter, candidates and alternatives mean different things.

according to i.

In this way, we have an initial separation between good and bad objects for each voter based on an individual evaluation of each candidate. Now, we need to define *separability* so this concept can be put into use when considering non-singleton elements of A.

Definition 5.3 (Separability). A preference $>_i$ is *separable* if $\forall O_s \subseteq O$ where $o \notin O_s$, $O_s \cup \{o\} >_i O_s$ if and only if $o \in G(>_i)$.

Thus, *separability* means that a set of candidates can only be improved in $>_i$ through the addition of a good candidate to it. This implies that the best set possible according to i is $G(>_i)$, while the worst is $G^c(>_i)$. The *separable preferences* domain is that where all voters exhibit a separable preference. The set of all possible separable preferences given O is denoted here as $sp(O)$. Thus:

Definition 5.4 (Separable Preferences Domain). The domain restriction for *separable preferences* is that for every voter i a preference $>_i$ is assigned where $>_i \in sp(O)$.

This is a very reasonable domain to consider. Take the political example once more. If for a given election there are some candidates that a voter likes, e.g., considers good, and some that this voter doesn't like, e.g., considers bad, then a preference order that is separable can be expected. On the other hand, there are contexts in which separability cannot be expected. These are the ones where interaction effects are prominent. For example, let's say two excellent professors p_1 and p_2 are to be chosen for leading a new institute in a university, but they really don't get along well and cannot work together without bringing the institution apart. In this case, given that voter i is aware of this situation, both p_1 and p_2 would belong to $G(>_i)$, but $\{p_1\} >_i \{p_1, p_2\}$, which goes against separability.[2] To illustrate better the concept of separability, some examples of separable and non-separable preferences are given in Table 5.1.

This domain restriction is also mathematically interesting because there is a family of SCFs that become strategyproof under separable preferences. To define them, we have to return to the definition of coalitions and committees, as is done in Chapter 3.

[2] This example is adapted from Barberà et al. (1991)

Table 5.1: Examples of separable and non-separable preferences. In this example, $O = \{o_1, o_2, o_3\}$ and $G(>_i) = \{o_1, o_2\}$. The brackets to denote sets are abbreviated for cleanliness. Alternatives are listed from top to bottom, where a set higher above in the table is more highly ranked in $>_i$ than one listed underneath.

Separable	Separable	Separable	Non-separable	Non-separable
o_1, o_2	o_1, o_2	o_1, o_2	o_1	o_1, o_2
o_1	o_1, o_2, o_3	o_2	o_2	o_1
o_2	o_1, o_3	o_1	$\varnothing$	o_2
$\varnothing$	o_2, o_3	o_1, o_2, o_3	o_1, o_3	o_2, o_3
o_1, o_2, o_3	o_1	$\varnothing$	o_1, o_2, o_3	$\varnothing$
o_2, o_3	o_2	o_1, o_3	o_2, o_3	o_3
o_1, o_3	$\varnothing$	o_2, o_3	o_1, o_2	o_1, o_3
o_3	o_3	o_3	o_3	o_1, o_2, o_3

(Repost) Definition 3.2 (Coalition). A *coalition* is a subset of voters. Thus, if n is the total number of voters and $N = \{i | i \in \mathbb{N}, 1 \leq i \leq n\}$, then a coalition is a member of 2^N.

(Repost) Definition 3.3 (Committee). A *committee* is a non-empty family C of non-empty coalitions that follows the following monotonicity restriction: given $S, T \subseteq N | S \in C, S \subsetneq T$, then $T \in C$.

For a detailed consideration of these concepts, refer back to their original definitions.

Now we can once again use the idea of a committee to establish a different voting rule than the generalized median voter scheme from the single-peaked domain.

Definition 5.5 (Voting by Committees). A SCF f is *voting by committees* if for each candidate $o \in O$ there exists a committee C_o such that for all profiles $P \in \mathfrak{P}$, $o \in f(P)$ if and only if $\{i | o \in b_i(A)\} \in C_o$.

Remember that $b_i(A)$ represents the top alternative in i's ballot. Important to note in this rule is that one winning committee associated to an alternative is determined beforehand based on the voter population. Then, if a coalition of this committee lists the alternative on the top of their ballot (which is presumed to be the same as their preference), the alternative is declared a winner of the election. The reciprocal statement must also hold; a winner of the election has a winning coalition of its associated committe present in the profile listing it on top of their ballots.

The term “winning” was used unpretentiously in the preceding explanation, but it is relevant to be aware that it is also a technical term.

Definition 5.6 (Minimal Winning Coalition). Coalitions in a committee C that elects a candidate through voting by committees are called *winning*. Given two coalitions $M \in C$ and M', M is a *minimal winning coalition* if and only if $M' \subsetneq M \to M' \notin C$, where $M' \subsetneq M$ means $M' \subset M$ and $M' \neq M$.

Notice that the monotonicity rule of committes make so that supersets of minimal winning coalitions are all included in C. Another property that must be pointed out over voting by committees is that this rule is not necessarily anonymous, since a committee to elect an alternative might be composed of exactly one voter considered an authority on the subject. However, the main theorem of this chapter does state two interesting properties to voting by committees.

Theorem 5.1 Under the separable preferences domain, voting by committees is the only SCF that is both strategyproof and nonimposed.

It is good to note that in the separable preferences domain, literature can also refer to the nonimposition property as *voter sovereignty.* Recall also that nonimposition means that all alternatives have at least one profile that can make them a winner. Per usual, the proof is skipped, but it can be found in Barberà et al. (1991).

Another interesting property of the separable preferences domain is that it also conforms SCFs that have other anti-manipulation properties besides strategyproofness that we have not discussed in this brief introduction, one example being that a voter does not benefit from voting two or more times (Fioravanti & Massó, 2024). It is also curious how an interesting property can be found going the other way around, as voting by committees is actually strategyproof in a larger set of preferences than just separable (Serizawa, 1995). On the other hand, partitions that create subdomains of separable preferences taking into account biases voters might have towards specific candidates seem not to extend the family of strategyproof SCFs from voting by committees (Martínez & Moreno, 2013). Furthermore, Barberà et al. (2005) is also an interesting paper to consider, since it touches upon how voting by committees can work if some subsets of the object set are not allowed as winners. It must also be noted that, unfortunately, separable preferences are not enough to guarantee strategyproofness for a whole group attempting to manipulate an election (Barberà et al., 2010). Interestingly enough, however, separable preferences are the core to strategyproofness in more elaborate domains, mainly those composed out of the product of other ones (Breton & Sen, 1999). Finally, Nehring and Puppe (2007), already briefly mentioned in Chapter 3 can now be of more value, since the core of its message is that there are hyperstructures that can be defined over SCFs' domains that bind single-peaked properties and separable preferences. This could be interesting to keep in mind for a later step in the application of social choice theory to the cognitive sciences, since if the more specific models work out to fit well to specific data patterns found in behavior or the brain, it might be that there is a more potent structure behind it that is also more generalizable.

On that note, we now come to the part of the chapter where we speculate how the presented piece of mathematics could be used in neuroscientific investigation. The intriguing part of the separable preferences domain is that it is based on the power set of a set of objects. This means that the voting by committees rule is appropriate for the aggregation of multiple different parts that will be yielded into one whole. And it is known that one of the biggest mysteries regarding perception for now is how separate characteristics are brought together into one cohesive and meaningful whole percept (Hummel, 2001). How neurons extract simple, lower layer information such as direction of an edge and color is already known, but how these are combined into one entity has been so elusive that scientists have even come up with the grandmother cell parody (Gross, 2002). The joke is that if for every characteristic and every object we need a neuron specialized to identify it, then to recognize our own grandmothers we would also need a cell specialized only for that. Of course, that is not plausible, since we must learn through experience how to recognize her, unlike lower characteristics which are partially innate. Thankfully enough, research has given us the concept of distributed representation, which is also applied in technologies born out of cognitive science, such as neural networks (Goller & Kuchler, 1996; Ishai et al., 1999; Rissman & Wagner, 2012; Thrun, 1994).

Distributed representation at its core basically says that a neuron higher up in the hierarchy would only fire when it senses information pertaining to different input neurons that represent specific patterns in the data that it is trained to identify. In perception, this means that we expect neurons from higher areas in the tract where object recognition is implemented to emit action potentials only when all neurons that code information about that type of object fire at the same time. For example, a neuron that is good at apple recognition would fire under the combination of "round" neurons and "red" neurons. This simple addition mechanism is, however, not enough. Apples can also be green, after all. To be able to efficiently collect all types of data and reject alternative hypotheses, the perceptual system must take in consideration factors that might not be so obvious at first. Voting by committees could be such a model.

Let's take mugs as an example. They come in all different shapes, sizes and colors, but they all have some very distinct features. They have a holder big enough for a few fingers to pass through, but small enough so that the holder itself can rest on top of the fingers and still be able to hold the rest of the mug against gravity. They also have a concave surface big enough to fit what is considered a regular amount of beverages, such as coffee and tea, but never too big to be meant for sharing. These constraints require previous world knowledge to be able to be computed, but are really good ways to heighten priority in perceptual decision towards a mug being recognized than any other kind of object. With voting by committees, we are able to select precisely which neurons heavily influence this type of task and give them higher value when a mug is to be declared the "winner of an election". Furthermore, the power set nature of voting by committees allows us to stack the same SCF if we make parallel elections with lower characteristics that are combined into elements of higher-order families (of sets). This way, information can be redirected efficiently towards where it is most useful in discrimination and be given higher or lower priority by higher-order neurons at the same time.

Of course, the setting up of these committees that would compose the SCF is not trivial. The brain must learn which sets of neurons to make into coalitions for each type of object it might encounter. On the other hand, the flexibility that such a method allows (of just switching around components of a coalition) makes it also very open for learning new associations into objects. This property, along with the aforementioned efficacy, makes this piece of social choice theory a suitable candidate for this type of modelling. Perhaps this could also help even with the area of explainable AI, where researchers try to unpack what exact calculations trained models compute so that they can be explained not only to customers, but also to courts so that companies can be held responsible for the products they produce (Holzinger, 2018; Holzinger et al., 2022). A decoding algorithm based on voting by committees might be a new dent into cracking this problem of our current hyperconnected society.

Chapter 6

Discussion & Conclusion

6.1 Recap

This thesis has the purpose to showcase social choice theory to cognitive scientists interested in expanding the toolboxes of cognitive modelling. We believe this to be important because social choice theory gives well fundamented mathematical tools for the aggregation of decisions, having already been applied to the social sciences. This link is possible to be traced because one can observe higher order neural calculations as judgements over compiled information - precisely what aggregation of decisions means. Thus, it is of relevance in the scientific spirit to at least pose this hypothesis of usefulness of social choice theory to the brain sciences in order to test it when convenience allows. A test of this hypothesis was not conducted as part of this study due to its scope. As part of a literature thesis in the program Brain and Cognitive Sciences of the University of Amsterdam, it is not the goal of the current study to conduct empirical experimentation, but rather to compile literature in the interstice of interdisciplinary fields that could shine a light on new research directions. Hopefully, not only the author, but the cognitive science community in general will soon be able to test it with the material at hand.

Overviewing the material up to this point, we see a pattern present through the chapters of the thesis. First, on Chapter 1, the Introduction, a general gist of the logics of social choice theory is presented. Next, on Chapter 2, we see the motivation for domain restriction in both social choice theory and its application to neuroscience. These two chapters are the opening to what this thesis is: an introduction to social choice theory, especially on the topic of strategyproofness, for cognitive scientists. Then, the bulk of the work is laid on the following three chapters. All of them present a motivation behind a certain domain restriction from the social choice perspective, followed by the presentation of the mathematical model that has been developed on that field. Next come its implications, further literature for intricacies that had to be left out of the scope of the current work for simplicity's sake, and finally a light speculation on how these models could be applied to neuroscience and the cognitive sciences more broadly.

These three mathematical models, seen as the combination of the domain restriction and the presented strategyproof SCF, were, namely: the *single-peaked domain* and *generalized median voter schemes*, the *single-crossed domain* and *peak rules*, and the *separable preferences domain* and *voting by committees*. Remember that a strategyproof SCF is one where no voter can benefit from manipulation or voting insincerely.

Briefly recapping, in single-peaked domains there is a predisposed order of the candidates, over which each voter's ballot must be single-peaked. This structure was mainly associated to the lower levels of perception, with how neurons fire to very determinate stimuli. In single-crossing domains, there is a predisposed order of both candidates and voters, such that the most preferred candidate of a voter could not be ranked lower or higher on the overarching order to the most preferred candidate of a voter that themself ranks lower or higher than the original voter. This structure was mainly associated to the learning mechanisms present in the brain, where prediction error values are used to adjust long-term decision-making. Finally, with separable preferences, the voters must now rank subsets of the candidates as alternatives following a preliminary division of whom they consider to be good or bad candidates. This structure was associated to the higher levels of perception, with how neurons must take into consideration different properties of the percept already captured on lower levels to reach a final understanding of what that percept is. It was also associated briefly to AI technology.

This link to AI also links to the division of matters still to be discussed before this introduction of social choice to the cognitive scientists can be concluded. Namely, speculation over the aforementioned hypotheses of possible applications of the mathematical models to brain data and where possible difficulties might lie, along with the relevance of social choice theory to the larger group of disciplines often associated to the cognitive sciences, such as AI itself. After these two topics are covered, we can finally get back to our final considerations and consider our job done - for now.

6.2 Difficulties Ahead

For the length of this thesis, we have seen mathematical structures with detail when regarding only themselves, but not when combined with actual data we obtain from studies on cognition. The reason for the former is evident: an introduction to social choice theory must present its basic definitions in order to serve its goals. However, considerations over the latter have been swept under the rug. It has come the time to address this.

The main issue here is the sheer variety of data we have at our disposal - from behavioral tasks to brain imaging. Naturally, most techniques utilized in cognitive modelling generally only tackle one or a few of them at a time. For example, evidence accumulation models (Forstmann et al., 2016). They are a set of models made to describe distributions of accuracy and reaction time (both behavioral measures) obtained from cognitive psychology empirical studies. The only configuration needed to fit one of these models is the separation between conditions of a study and a cleaning of the data that were poorly recorded. During the fitting process, the generally Bayesian algorithm then deduces which latent variables the model assumes have which values to best describe participant data. In this process, then, even if the dynamics of the model and the fitting procedure itself might be somewhat complicated, the raw input that goes into the model is not. Surely, that raw data can be more or less informative and quite expensive to obtain depending on the experimental setup, paradigm and recruitment of participants, but in the end, it is just a list of two associated values: reaction time and correctness. These models were thought up for that specific goal and now they have shown they are much more valuable than that, with studies linking the latent variables they assume to brain activations (Mulder et al., 2014). The same is true for studies that deal with said brain activation to begin with. Data collected from an fMRI or EEG scan is inherently hard to deal with, because of the many specificities each methodology brings. It is not a surprise then that methods tailored to each type of data exist separately of each other, and an overarching theme in the cognitive sciences is how to combine them to achieve a more holistic explanation (Parsons, 2001; Sanei & Chambers, 2013; S. M. Smith, 2004).

This proposal of “let’s try to use social choice to study cognition” in the shape of a thesis, however, starts the other way around. Specifically, out of interest for a field that at first sight seems to share many important properties with work that is commonly done in cognitive modelling. The main goal was to understand its methodology apart of the more traditional cognitive methodology to then attempt to bring them together. Thus, the exact type of data that could be handled was left as an afterthought. Also because neuronal firing and voting share in abstract a fundamental characteristic: they are both the revealing of preference towards a set of stimuli. The devil is in the details, though. A vote in social choice theory is, usually, a linear ordering.[1] Also, most of the work done in social choice theory is axiomatic. These two facts combined mean that we are bound to the structure of a linear ordering, or else the theory that already exists cannot be used, and a new one must be investigated. However, the output recorded from behavioral experiments or brain scans are never linear orderings. This is why the word 'translated' has been used throughout the speculations presented in the thesis. We must be able to identify first how to precisely do this transformation from traditional cognitive data to linear orderings if we are to purposefully apply social choice theory to our branches of scientific investigation.

How to tackle this problem, then? The answer is, like ever so commonly in science, 'it depends.' Unsatisfying as this answer may be, let's remember this is already the best answer we can give a majority of the times in more traditional brain sciences as well. For example, when a study wants to investigate a specific area of the brain using fMRI scans, it is seldom obvious which parcellation procedure would best fit the purposes of the study in extracting that area from the data.[2] There is always trade-offs to be considered (Eickhoff et al., 2018). This is the case again with social choice theory.

A linear ordering is not the same as a distribution of reaction times. A linear ordering is not the same as the continuous voltage signal of a local field potential. Inherently, data will be lost in the translation process. For example, single cell recordings can be transformed into linear orderings by the combination of firing frequency at the adequeate latency to the exhibition of a certain stimulus. This pairing of which stimulus elicited a higher frequency of action potentials can be made into a linear ordering exhibiting the preferences of that neuron. More elaborate data structures like EEG signals, whose components are hard to even trace back to a specific source in the brain, will require more interpretation to be meaningfully transformed into linear orderings. The argument to be made against this complication is that even with this loss of data we will still be able to gain more insight into the

[1] Modern social choice theory started with the work by Arrow (Arrow, 1970), which made use of linear orderings as votes. Later, expansions on his work were conducted, which include not only the domain restrictions we considered in this thesis, but also structural modifications such as changing the accepted format of a ballot. For examples of such work, in these cases, where a ballot is not a linear ordering, but a partial ordering, the reader can refer to Acevedo (2022) and Cullinan et al. (2014).

[2] For the mathematician, parcellation in this case means the assignment of a label representing a brain structure (such as 'prefrontal cortex' or 'amygdala') to a voxel obtained from the scan.

machinery of the brain via the application of such methodology. We believe this to be true.

Why? Well, let us step back to one of the setting stones of how cognitive scientists are trained to think of cognition and refer to Marr's levels of analysis. In his influential work, *Vision* (Marr, 2010), Marr proposed that we should think of three different levels when studying cognitive systems: the computational, the algorithmic, and the implementational. At the computational level, we are mainly concerned with input and output, not delving much deep on how exactly such output is achieved. This is the level of analysis most cognitive psychologists stay at with behavioral experiments. At the algorithmic level, we are concerned on what calculations are done on the input so that the output is obtained. This is the level of analysis cognitive modelling tends to stay. At the implementational level, we are concerned with how these calculations are materialized in the physical world. This is the level of analysis the traditional neurosciences focus on. And the harsh reality is that bridging all of the three levels is extremely hard (Carandini, 2012). And one complication factor that interests us here now is the type of data that is dealt with at each level. They are all varied, ranging from the behavioral to the molecular, and not easily transmutable. But what is easily overlooked is that all levels of analysis are common to one single cognitive aspect of interest, and thus the structures they deal with must all share the same final goal: whatever that module is there for. If that cognitive process can be interpreted in the most comprehensive way possible as an aggregation of preferences - then social choice theory can provide a mechanism to bridge these different levels. As long as a meaningful interpretation of the different signals we can record from different levels can be cast to reshape this data into linear orderings, we have at our disposition a rich and robust mathematical toolbox to serve as an interface between these different levels.

This is also an argument in favor of abstraction. Daily in our labs we are faced with concrete data and concrete experimentation, which is good for the construction of evidence, but not necessarily the best guide to understand overarching themes in such an interdisciplinary field as the cognitive sciences. It is productive to look at abstractions of ideas and see how they match in their most simplified version to be able to make new connections and gain new insights. This is the main goal of this thesis. We hope that it reaches out the appropriate reader that will be able to make use of the work showcased here to set the abstraction back into the concrete.

Which domain and SCF will fit your experimental data the best?

6.3 Cognitive Sciences, AI and Social Choice Theory

Moving further, another good mental exercise is to consider how social choice theory can contribute to the cognitive sciences as a whole, not only specifically modelling of brain activation. The field of cognitive science relies on the coming together of many disciplines. Of these, psychology and neuroscience are the most preminent, but we should never forget how philosophy, linguistics, computer science and so on have left their mark on the history of the studies of cognition. An exemplary remark on that is how the now so popular tools of neural networks and AI have originally started in the cradle of the cognitive sciences. It was a psychologist in the 50s who first introduced the concept of a perceptron (Rosenblatt, 1958), which is the basis of modern neural network technology, and again a psychologist who first used backpropagation for learning in the 80s (Rumelhart et al., 1986). It is impressive how little these facts are known outside of the cognitive sciences community, given the attention AI has received in the past years.

A recurrent problem in AI is decision making (Phillips-Wren, 2012). Specially for systems which are not solely used to identify an object, like in computer vision, but whose outcome will directly impact a human life. This is the example of algorithms used in loan concession or self-driving cars, for example. In case of an accident with a self-driving car, what should be the car's priority? To save the people inside it to the prejudice of a pedestrian? Does that depend on the number of people affected, age, gender, and other variables we can consider for humans? This type of ethical questioning is not trivial for us, much less for machines (Kazim & Koshiyama, 2021). Such difficulty can be felt first-hand with the moral machine experiment by Awad et al. (2018), where they gathered and analyzed millions of decisions over this type of situtation. Furthermore, the inclusion of social choice theory topics in AI conferences (Torra et al., 2014) shows that researchers are aware of the possibility of application of social choice theory to this type of machine decision making.

The problem of being able to conceive human ethics into machines and limit their actions to this scope is known as the alignment problem (Yudkowsky, 2016). Although a hard problem to tackle, fortunately there are many researchers investigating it. Some are using social choice as a tool. It is not hard to envision why this is seen as a possibility. To be able to instruct a computer on the most common choice a human would make in different situations, the best one could do is ask participants about their preferences over the courses of actions and compile them. This is precisely what social choice theory was developed for. Baum (2020) offers a brief overview of this work, while Conitzer et al. (2024) gives arguments over why social choice must be considered as part of the implementation of AI systems. Finally, Fish et al. (2023) extends the notion of social choice theory so that agreements over text renditions of ideas can be come upon as well, not only simple preference aggregation. This can be used to support democratic processes that require more than just a selection of candidates, but ratification of statements. As it can be seen, social choice theory is versatile in its applications.

Outside the scope of AI, there is still much that can be modelled besides brain activation patterns. Behavior, even though characterized by way less output variables than what is traditionally collected from brain scans, never ceases to amaze us. As it has already been pointed out with the example of evidence accumulation models, at times cognitive models begin at the behavioral level to then link its implicit dynamics to brain activation. Given the versatility of the mathematics of social choice, as long as there is enough creativity and boldness by the researcher, there is no reason why an application in behavioral data analysis cannot be found. For example, many studies on gambling have a behavioral task that asks for participants to make choices between different bets (Brevers et al., 2013; Buelow & Suhr, 2009). Is it not possible to fit a SCF on a group level to compare how strategyproof the collective choices of people with gambling disorder would be compared to healthy controls? What could this difference mean if found? As long as there is an aggregation of preferences, a social choice theory model can be thought of. This is the main takeaway from this thesis.

6.4 Further Reading and Final Remarks

Finally, to conclude our exploration into social choice theory, let's point out a few studies that look into strategyproofness that might be of interest to the reader who wishes to move further into the field of social choice theory. First, let's remember how this investigation started: the Gibbard-Satterthwaite theorem.

(Repost) Theorem ??.1 (Gibbard-Satterthwaite Theorem). Any resolute, nonimposed, and strategyproof SCF for three or more alternatives must be a dictatorship.

This theorem holds whenever there is no restriction on the domain of voting profiles. We have seen so far different restrictions that lax this theorem giving us strategyproof SCFs for each domain. It is important to point out that the different domain restrictions that social choice theory studies are not limited to the ones presented here, as one can think of all sorts of constraints on ballot format, voter ordering, and so on (Barberà, 2007; Barberà et al., 2012; Chatterji et al., 2013; Endriss & Grandi, 2018). This line of thought also traces back to the notion presented in Chapter 5 with Nehring and Puppe (2007): that some domains are subdomains of hyperstructures of other domains. This, however, is a more advanced topic in social choice theory. Finally, as far as domain classification goes, Chatterji and Zeng (2023) presents us a framework on how to classify different domains that accept a strategyproof SCF. This work specially might be of interest to the advanced cognitive modeller who wants to understand all the mathematical intricacies of different domains so that this knowledge can be applied in their work.

To conclude, we would like to present one promising study that while does not use social choice theory methods per se, it does use adjacent mathematical concepts. Lefgren et al. (2018) used notions of utility and preferences to model behavior observed in people experiencing depression. They only cite social choice theory briefly as a similar toolbox used for group decision making, unlike the individual level they theorize on. It should be clear to the reader by now that we do not agree with this stance, since an individual's decision making is still the aggregation of votes of many neurons. However, what is interesting about their work is that they applied this simple model of utility over preferences to actual data, obtaining results that confirmed their model. Seeing this different configuration of cognitive modelling also succeding is promising because it shows us that the effort on developing new toolboxes can be well paid off. This boosts our encouragement towards the application of social choice theory to actual experimental data.

Unfortunately, due to the scope set by the Brain and Cognitive Sciences program for this work, we limited ourselves to reviewing literature and writing a piece that could be accessible and thought provoking to cognitive scientists, without being able to delve into a hands-on trial of our ambitious idea. However, we are eager to see this possibility tested out, and urge the reader to enter in contact in case they make such an attempt. As mentioned in the Introduction, we were not able to retrieve any work on the brain sciences that makes use of the framework of social choice theory, and partially the reason might be that it has actually been done in the past, but unpublished due to an unsuccesfull model. Publication bias is a current unfortunate reality in academia, and to avoid that our efforts go to waste, we must be able to share our ideas and failed attempts among each other as well. Although we hope social choice theory provides novel and useful mathematical modelling paradigms to the cognitive sciences, and believe it will, it is not guaranteed.

Let us know of your failures as well, and let us help each other in the cognitive scientific community to further deepen our understand of the mysterious mechanisms of the brain and the ever fascinating logics of human behavior.

Thank you for taking this thesis at hand.

References

Acevedo, M. (2022). An exploration of voting with partial orders. *HMC Senior Thesis*, *267*.

Arribillaga, R. P., Massó, J., & Neme, A. (2020). On obvious strategy-proofness and single-peakedness. *Journal of Economic Theory*, *186*, 104992.

Arrow, K. (1970). *Social Choice and Individual Values* (2nd ed.). Yale University Press.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64.

Ayede, M. (2010). The language of thought hypothesis. In *Stanford Encyclopedia of Philosophy*.

Ballester, M. A., & Haeringer, G. (2011). A characterization of the single-peaked domain. *Social Choice and Welfare*, *36*(2), 305–322.

Bandyopadhyay, S., Shamma, S. A., & Kanold, P. O. (2010). Dichotomy of functional organization in the mouse auditory cortex. *Nature Neuroscience*, *13*(3), 361–368.

Barberà, S. (2001). An introduction to strategy-proof social choice functions. *Social Choice and Welfare*, *18*(4), 619–653.

Barberà, S. (2007). Indifferences and domain restrictions. *Analyse & Kritik*, *29*(2), 146–162.

Barberà, S., Berga, D., & Moreno, B. (2010). Individual versus group strategy-proofness: When do they coincide? *Journal of Economic Theory*, *145*(5), 1648–1674.

Barberà, S., Berga, D., & Moreno, B. (2012). Domains, ranges and strategy-proofness: The case of single-dipped preferences. *Social Choice and Welfare*, *39*(2), 335–352.

Barberà, S., Gul, F., & Stacchetti, E. (1993). Generalized median voter schemes and committees. *Journal of Economic Theory*, *61*(2), 262–289.

Barberà, S., Massó, J., & Neme, A. (2005). Voting by committees under constraints. *Journal of Economic Theory*, *122*(2), 185–205.

Barberà, S., & Moreno, B. (2011). Top monotonicity: A common root for single peakedness, single crossing and the median voter result. *Games and Economic Behavior*, *73*(2), 345–359.

Barberà, S., Sonnenschein, H., & Zhou, L. (1991). Voting by committees. *Econometrica*, *59*(3), 595.

Bartholdi, J. J., Tovey, C. A., & Trick, M. A. (1992). How hard is it to control an election? *Mathematical and Computer Modelling*, *16*(8), 27–40.

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & SOCIETY*, *35*(1), 165–176.

Bertero, M., Poggio, T., & Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE*, *76*(8), 869–889.

Blin, J. M., & Satterthwaite, M. A. (1976). Strategy-proofness and single-peakedness. *Public Choice*, *26*(1), 51–58.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

Bredereck, R., Chen, J., & Woeginger, G. J. (2013). A characterization of the single-crossing domain. *Social Choice and Welfare*, *41*(4), 989–998.

Bredereck, R., Chen, J., & Woeginger, G. J. (2016). Are there any nicely structured preference profiles nearby? *Mathematical Social Sciences*, *79*, 61–73.

Breton, M. L., & Sen, A. (1999). Separable preferences, strategyproofness, and decomposability. *Econometrica*, *67*(3), 605–628.

Brevers, D., Bechara, A., Cleeremans, A., & Noël, X. (2013). Iowa Gambling Task (IGT): Twenty years after – gambling disorder and IGT. *Frontiers in Psychology*, *4*, 61282.

Bruner, M.-L., & Lackner, M. (2014). The likelihood of structure in preference profiles. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the Iowa Gambling Task. *Neuropsychology Review*, *19*(1), 102–114.

Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Neuroscience*, *15*(4), 507–509.

Chatterji, S., Sanver, R., & Sen, A. (2013). On domains that admit well-behaved strategy-proof social choice functions. *Journal of Economic Theory*, *148*(3), 1050–1073.

Chatterji, S., & Zeng, H. (2023). A taxonomy of non-dictatorial unidimensional domains. *Games and Economic Behavior*, *137*, 228–269.

Chen, J., & Finnendahl, U. P. (2018). On the number of single-peaked narcissistic or single-crossing narcissistic preference profiles. *Discrete Mathematics*, *341*(5), 1225–1236.

Cohen, A. L. (2006). Contributions of invariants, heuristics, and exemplars to the visual perception of relative mass. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 574–598.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., & Zwicker, W. S. (2024). Social choice for AI alignment: Dealing with diverse human feedback. *arXiv*.

Conway, B. R. (2009). Color vision, cones, and color-coding in the cortex. *The Neuroscientist*, *15*(3), 274–290.

Cullinan, J., Hsiao, S. K., & Polett, D. (2014). A Borda count for partially ordered ballots. *Social Choice and Welfare*, *42*(4), 913–926.

Daw, N. D., & Tobler, P. N. (2014). Value learning through reinforcement. In *Neuroeconomics* (pp. 283–298). Elsevier.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*(5787), 684–687.

Dowding, K., & Van Hees, M. (2008). In praise of manipulation. *British Journal of Political Science*, *38*(1), 1–15.

Drew, P. J. (2019). Vascular and neural basis of the BOLD signal. *Current Opinion in Neurobiology*, *58*, 61–69.

Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, *19*(11), 672–686.

Elkind, E., Faliszewski, P., & Skowron, P. (2020). A characterization of the single-peaked single-crossing domain. *Social Choice and Welfare*, *54*(1), 167–181.

Elkind, E., Faliszewski, P., & Slinko, A. (2009). Swap bribery. In *Algorithmic Game Theory: Second International Symposium, SAGT 2009, Paphos, Cyprus, October 18-20, 2009. Proceedings 2* (pp. 299–310). Springer Berlin Heidelberg.

Endriss, U., & Grandi, U. (2018). Graph aggregation. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 447–450.

Faliszewski, P., Hemaspaandra, E., & Hemaspaandra, L. A. (2009). How hard is bribery in elections? *Journal of Artificial Intelligence Research*, *35*, 485–532.

Fioravanti, F., & Massó, J. (2024). False-name-proof and strategy-proof voting rules under separable preferences. *Theory and Decision*, 1–18.

Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wüthrich, M. (2023). Generative social choice. *arXiv*.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions [Number: 1]. *Annual Review of Psychology*, *67*(1), 641–666.

Foster, D. H. (2011). Color constancy. *Vision Research*, *51*(7), 674–700.

Gazzaniga, M. S. (2019). *Cognitive neuroscience: The biology of the mind* (Fifth edition.). W.W. Norton & Company.

Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, *41*(4), 587.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In *Blackwell handbook of judgment and decision making* (pp. 62–88).

Goller, C., & Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. *Proceedings of International Conference on Neural Networks (ICNN'96)*, *1*, 347–352.

Grant, J. E., & Chamberlain, S. R. (2015). Gambling disorder and its relationship with substance use disorders: Implications for nosological revisions and treatment. *The American Journal on Addictions*, *24*(2), 126–131.

Graybiel, A. M., & Grafton, S. T. (2015). The striatum: Where skills and habits meet. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021691.

Gross, C. G. (2002). Genealogy of the "Grandmother Cell". *The Neuroscientist*, *8*(5), 512–518.

Gurney, K. (2018). *An Introduction to Neural Networks*. CRC Press.

Hemaspaandra, E., Hemaspaandra, L. A., & Rothe, J. (2007). Anyone but him: The complexity of precluding an alternative. *Artificial Intelligence*, *171*(5), 255–285.

Hemaspaandra, E., Hemaspaandra, L. A., & Rothe, J. (2009). Hybrid elections broaden complexity-theoretic resistance to control. *Mathematical Logic Quarterly*, *55*(4), 397–424.

Holzinger, A. (2018). From machine learning to explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55–66.

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., & Samek, W. (Eds.). (2022). *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (Vol. 13200). Springer International Publishing.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, *8*(3), 489–517.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*(16), 9379–9384.

Jaeckle, F., Peters, D., & Elkind, E. (2018). On recognising nearly single-crossing preferences. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Karpov, A. (2020). The likelihood of single-peaked preferences under classic and new probability distribution assumptions. *Social Choice and Welfare*, *55*(4), 629–644.

Katsuki, F., & Constantinidis, C. (2014). Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, *20*(5), 509–521.

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *2*(9), 100314.

Lackner, M.-L., & Lackner, M. (2017). On the likelihood of single-peaked preferences. *Social Choice and Welfare*, *48*(4), 717–745.

Lefgren, L., Stoddard, O., & Stovall, J. (2018, September). *Are Two Bads Better Than One? A Model of Sensory Limitations* (w25060). National Bureau of Economic Research. Cambridge, MA.

List, C., Luskin, R. C., Fishkin, J. S., & McLean, I. (2013). Deliberation, single-peakedness, and the possibility of meaningful democracy: evidence from deliberative polls. *The Journal of Politics*, *75*(1), 80–95.

Marr, D. (2010, July 9). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Martínez, R., & Moreno, B. (2013). Strategy-proofness on restricted separable domains. *Review of Economic Design*, *17*(4), 323–333.

Martínez-Selva, J. M., Sánchez-Navarro, J. P., Bechara, A., & Román, F. (2006). Brain mechanisms involved in decision making. *Revista de neurologia*, *42(7)*, 411.

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *eLife*, *10*, e63055.

Moulin, H. (1980). On strategy-proofness and single peakedness. *Public Choice*, *35*(4), 437–455.

Mulder, M., Van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences – A model-based review. *Neuroscience*, *277*, 872–884.

Nehring, K., & Puppe, C. (2007). The structure of strategy-proof social choice — Part I: General characterization and possibility results on median spaces. *Journal of Economic Theory*, *135*(1), 269–305.

Nisan, N. (2007). Introduction to mechanism design (for computer scientists). In E. Tardos, N. Nisan, T. Roughgarden, & V. V. Vazirani (Eds.), *Algorithmic Game Theory* (pp. 209–242). Cambridge University Press.

Parsons, L. M. (2001). Integrating cognitive psychology, neurology and neuroimaging. *Acta Psychologica*, *107*(1), 155–181.

Pascual-Leone, A. (1999). Transcranial magnetic stimulation: Studying the brain-behaviour relationship by induction of 'virtual lesions' (A. Howseman, S. Zeki, D. Bartres-Fazf, & J. P. Keenan, Eds.). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *354*(1387), 1229–1238.

Penn, E. M., Patty, J. W., & Gailmard, S. (2011). Manipulation and single-peakedness: A general result. *American Journal of Political Science*, *55*(2), 436–449.

Phillips-Wren, G. (2012). AI tools in decision making support systems: A review. *International Journal on Artificial Intelligence Tools*, *21*(2), 1240005.

Potenza, M. N., Balodis, I. M., Derevensky, J., Grant, J. E., Petry, N. M., Verdejo-Garcia, A., & Yip, S. W. (2019). Gambling disorder. *Nature Reviews Disease Primers*, *5*(1), 51.

Procaccia, A. D., & Tennenholtz, M. (2013). Approximate mechanism design without money. *ACM Transactions on Economics and Computation*, *1*(4), 1–26.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, *88*(1), 455–463.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: Insights from functional brain imaging. *Annual Review of Psychology*, *63*(1), 101–128.

Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Decision-making, errors, and confidence in the brain. *Journal of Neurophysiology*, *104*(5), 2359–2374.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rothschild, G., Nelken, I., & Mizrahi, A. (2010). Functional organization and population dynamics in the mouse primary auditory cortex. *Nature Neuroscience*, *13*(3), 353–360.

Rumelhart, D. E., Hintont, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323(6088)*, 533–536.

Sakai, J. (2020). How synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceedings of the National Academy of Sciences*, *117*(28), 16096–16099.

Sanei, S., & Chambers, J. A. (2013, May 28). *EEG Signal Processing*. John Wiley & Sons.

Saporiti, A. (2009). Strategy-proofness and single-crossing. *Theoretical Economics*, *4(2)*, 127–163.

Saporiti, A., & Tohmé, F. (2006). Single-crossing, strategic voting and the median choice rule. *Social Choice and Welfare*, *26*(2), 363–383.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, *10*(2), 187–217.

Serizawa, S. (1995). Power of voters and domain of preferences where voting by committees is strategy-proof. *Journal of Economic Theory*, *67*(2), 599–608.

Smith, J. M. (1982). *Evolution and the theory of games*. Cambridge University Press.

Smith, S. M. (2004). Overview of fMRI analysis. *The British Journal of Radiology*, *77*, S167–S175.

Tarter, R. E., & Vanyukov, M. M. (2001). Introduction: Theoretical and operational framework for research into the etiology of substance use disorders. *Journal of Child & Adolescent Substance Abuse*, *10*(4), 1–12.

Thrun, S. (1994). Extracting rules from artificial neural networks with distributed representations. *Advances in neural information processing systems*, *7*.

Todorović, D. (2020). What are visual illusions? *Perception*, *49*(11), 1128–1199.

Torra, V., Narukawa, Y., Navarro-Arribas, G., & Megías, D. (Eds.). (2014). *Modeling Decisions for Artificial Intelligence: 10th International Conference, MDAI 2013, Barcelona, Spain, November 20-22, 2013. Proceedings* (Vol. 8234). Springer Berlin Heidelberg.

Yamane, Y., Kodama, A., Shishikura, M., Kimura, K., Tamura, H., & Sakai, K. (2020). Population coding of figure and ground in natural image patches by V4 neurons (C. R. Fetsch, Ed.). *PloS one*, *15*(6), e0235128.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111(23)*, 8619–8624.

Yudkowsky, E. (2016). The AI alignment problem: Why it's hard, and where to start. *Symbolic Systems Distinguished Speaker*, *4*, 1.