Motivation

- The goal of the selection theorems program is to find theorems that 'narrow down the type signature of an agent'. Coherence theorems in particular assume a coherence property, and derive other properties: for example, the complete class theorem shows (roughly) that a decision rule (of type $observation \to action$) that is not dominated minimizes loss with respect to some probability distribution.
    - Ryan: I think it's not just this. I think it's also narrowing down the type signature of systems that are strongly selected for *but aren't actually agents*
- The type of result we're looking for in the economics literature is *nonexistence of a representative agent*: showing that a market consisting of subagents does not have a utility function, or equivalently shows incoherent behavior (for some definition of incoherent).
    - Ryan: I think the reason we want to establish a general criterion for determining whether a market is an agent is so:
        - We can determine the type signature of human preferences and know how it will behave as a market if we e.g. allow contracts
        - We know how superintelligence markets will behave
- Ideally, we want to find a coherence theorem from the economics literature that assumes the market is coherent (has a utility function with some restricted domain), and derives some property p of the subagent utility functions. If p is very restrictive, this means that most collections of subagents are incoherent, and thus will not be selected for. We looked mainly at Jackson's paper, which references earlier work by Gorman.

Gorman (1953, 1961)

- Gorman analyzes Marshallian demand functions; that is, functions from income I and prices p_i to consumption levels x_i
    - $D(p_1, \ldots, p_L, y) = (x_1, \ldots, x_L)$
- When there are n agents in the market, each with demand $D_i(p, y_i)$, the total demand is just $\sum_i D_i(p, y_i)$.
- Gorman asks the question: When is this demand function purely a function of the total income $\sum_i y_i$?
    - $$D\left(p, \sum_i y_i\right) = \sum_i D_i(p, y_i)$$
- and derives that the utility functions of each subagent are Gorman aggregable. Examples of Gorman aggregable functions: all linear in income, or homothetic and identical.[1] This is a classic result in economics.

[1] Jackson (I think mistakenly) says that *all* Gorman aggregable functions are either linear in income, or homothetic and identical.

    - Linear in income means the utility function for each agent i is of the form $u_i(x, y) = u_i(x) + y$
    - identical (up to normalization) and homothetic: that is, satisfying $x_i = f_i(p) + ug_i(p)$ where $f = \sum p_i f_i$, $g = \sum p_i g_i$ are homogeneous of degree 1.

Jackson (2020)

- Jackson and Yariv prove a similar result with a somewhat more natural assumption. Instead of assuming that the *demand functions* of subagents can be aggregated into a representative demand function, they assume that the *utility functions* $V(\cdot, a_i)$ of subagents can be aggregated (by linear combination) into a single utility function $V(\cdot, a)$ that is purely a function of the average resources allocated to all subagents.
    - $V\left(\sum \lambda_i x_i, a\right) = \sum_i \lambda_i V\left(x_i, a_i\right)$
- This assumption is inspired by (but not implied by) the fact that Pareto-efficient markets behave like they are maximizing some linear combination of subagent utilities.
    - Also inspired by the notion of an “average agent”, relevant to the utilitarian welfare function, where $$\lamdba_I$$ represents the proportion of population with preferences $$a_i$$
- Jackson and Yariv again obtain strong restrictions on the subagent utility functions: they must be of the form
    - $V(x; a) = c \cdot x + h(a)$
- i.e. linear with respect to all goods!

Relevance to agent foundations

- We think that neither result is terribly meaningful for the selection theorems program. Both Gorman and Jackson-Yariv make unrealistic assumptions regarding the form of the representative agent's utility function. For the selection theorems program, the only assumptions we want to make are coherence properties.
- I also asked a microeconomics PhD student (in a different subspecialty) and he thinks that most papers about representative agents are taking a welfare economics lens or a modeling-simplification lens, not a coherence lens.
- Could other utility functions exist that don’t take the Jackson-Yariv form but do represent the market’s preferences?
    - Is generally a function on internal world states (i.e. subagent resource allocations)