"The New Keynesian Cross: Understanding Monetary Policy and Forward Guidance with Heterogeneous Households"

Paper 1  "The New Keynesian Cross: Understanding Monetary Policy with Hand-to-Mouth Households"

p. 2 to p. 44

Paper 2  "The Puzzle, the Power, and the Dark Side: Forward Guidance Redux"

p. 45 to p. 72
THE NEW KEYNESIAN CROSS:
UNDERSTANDING MONETARY POLICY
WITH HAND-TO-MOUTH HOUSEHOLDS

Florin Ovidiu Bilbiie

MONETARY ECONOMICS AND
FLUCTUATIONS
THE NEW KEYNESIAN CROSS: UNDERSTANDING MONETARY POLICY WITH HAND-TO-MOUTH HOUSEHOLDS

Florin Ovidiu Bilbiie

Discussion Paper DP11989
Published 23 April 2017
Revised June 2017

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre’s research programme in MONETARY ECONOMICS AND FLUCTUATIONS. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Florin Ovidiu Bilbiie
THE NEW KEYNESIAN CROSS: UNDERSTANDING MONETARY POLICY WITH HAND-TO-MOUTH HOUSEHOLDS

Abstract

The New Keynesian Cross describes aggregate demand through a planned expenditure PE curve and captures a key amplification mechanism and decomposition of heterogeneous-agent New Keynesian (HANK) models à la Kaplan, Moll and Violante, 2015. In response to monetary policy, PE’s shift is the direct effect (intertemporal substitution), while its slope (marginal propensity to consume) is the share of the indirect effect in total. There is amplification (dampening) when hand-to-mouth’s income elasticity to aggregate is more (less) than unity; This elasticity depends chiefly on income (including fiscal re-distribution. The effects are magnified by self-insurance when households are hand-to-mouth only occasionally: the aggregate Euler equation now features discounting (McKay, Nakamura and Steinsson, 2015) when the elasticity of hand-to-mouth income to aggregate is lower than unity, but compounding when larger. This matters most for forward guidance (FG), whose power is reduced in the former case, thus resolving the "FG puzzle" (Del Negro et al, 2013) - but amplified in the latter (Werning, 2015), thus aggravating the puzzle.

JEL Classification: E21, E31, E40, E44, E50, E52, E58, E60, E62

Keywords: hand-to-mouth; heterogenous agents; aggregate demand; optimal monetary policy; liquidity trap; Keynesian cross; forward guidance.

Florin Ovidiu Bilbiie - florin.bilbiie@gmail.com
Paris School of Economics and CEPR

Acknowledgements

I am grateful in particular to Jordi Galí, as well as Jess Benhabib, Edouard Challe, Lawrence Christiano, Davide Debortoli, Virgiliu Midrigan, Benjamin Moll, Emi Nakamura, Xavier Ragot, Kenneth Rogoff, Jon Steinsson, Gianluca Violante, Mirko Wiederholt, and Michael Woodford for discussions and encouragement, and participants in several seminars and conferences for comments. I gratefully acknowledge without implicating the support of Banque de France via the eponymous Chair at PSE, and of Institut Universitaire de France, as well as the hospitality of New York University and CREI during part of writing this paper.
The New Keynesian Cross: Understanding Monetary Policy with Hand-to-Mouth Households

Florin O. Bilbiie

June 2017 (First draft April 2016)

Abstract

The New Keynesian Cross describes aggregate demand through a planned expenditure PE curve and captures a key amplification mechanism and decomposition of heterogeneous-agent New Keynesian (HANK) models à la Kaplan, Moll and Violante, 2015. In response to monetary policy, PE’s shift is the direct effect (intertemporal substitution), while its slope (marginal propensity to consume) is the share of the indirect effect in total. There is amplification (dampening) when hand-to-mouth’s income elasticity to aggregate is more (less) than unity; This elasticity depends chiefly on income (including fiscal re-)distribution. The effects are magnified by self-insurance when households are hand-to-mouth only occasionally: the aggregate Euler equation now features discounting (McKay, Nakamura and Steinsson, 2015) when the elasticity of hand-to-mouth income to aggregate is lower than unity, but compounding when larger. This matters most for forward guidance (FG), whose power is reduced in the former case, thus resolving the "FG puzzle" (Del Negro et al, 2013)—but amplified in the latter (Werning, 2015), thus aggravating the puzzle.

JEL Codes: E21, E31, E40, E44, E50, E52, E58, E60, E62

Keywords: hand-to-mouth; heterogenous agents; aggregate demand; monetary policy; liquidity trap; Keynesian cross; forward guidance.

\footnote{I am grateful in particular to Jordi Galí, as well as Jess Benhabib, Edouard Challe, Lawrence Christiano, Davide Deportoli, Virgiliu Mierigan, Benjamin Moll, Emi Nakamura, Xavier Ragot, Kenneth Rogoff, Jon Steinsson, Gianluca Violante, Mirko Wiederholt, and Michael Woodford for discussions and encouragement, and participants in several seminars and conferences for comments. I gratefully acknowledge without implicating the support of Banque de France via the eponymous Chair at PSE, and of Institut Universitaire de France, as well as the hospitality of New York University and CREI during part of writing this paper.}

\footnote{Paris School of Economics, UP1 Panthéon Sorbonne, and CEPR; 48 Boulevard Jourdan 75014 Paris. florin.bilbiie@parisschoolofeconomics.eu. http://florin.bilbiie.googlepages.com.}
"The amount that the community spends on consumption obviously depends on [...] the principles on which income is divided between the individuals composing it (which may suffer modification as output is increased)." 

"[...] we may have to make an allowance for the possible reactions of aggregate consumption to the change in the distribution of a given real income between entrepreneurs and rentiers resulting from a change in the wage-unit".

(Keynes [1935], Chapter 8, Books I and III).

1 Introduction

The Keynesian cross of the baseline New Keynesian model is not very Keynesian at all: the slope of aggregate demand, or the planned expenditure (PE) curve is very close to zero. In other words: consumption is almost insensitive to current income. This is blatantly in contrast with mounting evidence—reviewed in great detail elsewhere—obtained using a wide spectrum of (micro and macro) data and econometric techniques.1

The demand side of the NK model has been slow to evolve to meet this challenge, but a new wave of research into heterogeneous-agent New Keynesian models (labelled HANK by one of the main references in this literature, Kaplan, Moll and Violante 2015—hereinafter KMV) is changing this. This burgeoning literature uses heterogeneous-agent models with financial imperfections and nominal rigidities to analyze the transmission of monetary policy and its redistributive effects.2

The models used are often quantitatively plausible and are solved numerically: they use plausible idiosyncratic income processes, and can fit distributions of wealth, asset holdings, and various aspects of household finances.

A major theme of the HANK model is the decomposition of the effect of monetary policy proposed by KMV: into a "direct" effect, driven essentially by intertemporal substitution, and an "indirect effect" consisting of the endogenous amplification on output through general-equilibrium effects. KMV show that while in the representative-agent model most of the total effect of monetary policy is driven by the former component, in their HANK model a

1See Mankiw (2000), Gali, Lopez-Salido and Valles (2007), Bilbiie and Straub (2012, 2013) for reviews. Five streams of evidence emerge: (i) direct evidence on zero net worth from micro data (i.a. Wolff, 2000; Bricker et al, 2014); (ii) the problem of zero elasticity of intertemporal substitution in estimated consumption Euler equations, which is almost as old as the consumption Euler equation itself (from Hall, 1978 to Hall, 1988; Campbell and Mankiw, 1989, 1990, 1991; Mankiw and Zeldes, 1991; Yogo 2003; Hurst 2004; Vissing-Jorgensen 2003, Bilbiie and Straub, 2012 and many others); (iii) the sensitivity of consumption to fiscal transfers and rebates (Parker, 1999; Johnson, Parker and Souleles 2006); (iv) the recent literature on "wealthy hand to mouth"—a significant share of agents with illiquid wealth behave as H (Kaplan and Violante 2014; Surico et al 2016); (v) worldwide evidence (World Bank, 2014).

large portion (as much as 80 percent) is due to the latter.\textsuperscript{3}

An earlier wave of analysis of heterogeneous agents and aggregate demand in sticky-price models used a simpler, two-agent setup—what KMV call TANK, label that I will embrace here too; in those models, a fraction of agents are "hand-to-mouth" (indexed $H$): they consume all their disposable income. The pioneering paper was Galí, Lopez-Salido and Valles (2007),\textsuperscript{4} which studied government spending multipliers in a model where the distinction between optimizing and hand-to-mouth agents is modelled through access (or lack thereof) to physical capital—as was suggested by Mankiw (2000) in a different context. The very important contribution of the paper was to show for the first time that in this model, if enough agents are $H$, government spending can have a positive multiplier on private consumption; this "Keynesian" conclusion is in line with some empirical findings and unlike then-existing flexible-price and sticky-price models that implied negative consumption multipliers.\textsuperscript{5}

Bilbiie (2008) studied monetary policy using as a starting point GLV's framework and simplifying it by modelling the distinction between agents based on participation in asset markets (thus abstracting from physical investment): savers hold shares in firms. The novel element afforded by this simplification was an analytical, closed-form expression for the aggregate Euler-IS curve, which clarified that the elasticity of aggregate demand to interest rates is increasing with the share of constrained $H$ agents (the economy becomes "more Keynesian"), up to some threshold (beyond that, the elasticity changes sign and the economy becomes "non-Keynesian": interest rate cuts become contractionary, for reasons explained in that paper in detail). The paper also derived optimal monetary policy in this framework, and studied the determinacy properties of interest rate rules.\textsuperscript{6}

\textsuperscript{3}Auclert (2016) performs a different but related decomposition into three channels that account for households’ financial positions.

\textsuperscript{4}The first version of GLV’s paper was dated 2002; a companion paper (GLV, 2004) analyzed determinacy properties of interest rate rules, and showed numerical simulations suggesting that the Taylor principle is not sufficient for determinacy (i.e. it is too weak) in their model where only a subset of agents hold the stock of physical capital.

\textsuperscript{5}Other work extended this paper: Bilbiie and Straub (2004) analyzed a model with distortionary taxation and competitive labor market. Bilbiie Meier, and Mueller (2008) showed that an estimated version of the GLV model can reproduce the decrease in consumption multipliers during the Great Moderation period through a combination of more widespread participation in financial markets (lower hand-to-mouth share) and more active monetary policy. Monacelli and Perotti (2013) studied the role of redistribution for the multiplier, and Bilbiie, Monacelli, and Perotti (2012, 2013) the effect of public debt, and spending and tax cuts’ effect on welfare in a liquidity trap, respectively.

\textsuperscript{6}The earliest version is the working paper Bilbiie (2004). Bilbiie and Straub (2012, 2013) present empirical evidence consistent with the "Keynesian" region since the 1980s and with the non-Keynesian region during the Great Inflation. Colciago (2012) and Ascari, Colciago, and Rossi (2015) show that sticky wages make the occurrence of the "non-Keynesian" region less likely. Nistico (2016) generalizes this to Markov switching between types and studies financial stability as an objective of monetary policy. Eggertsson and Krugman (2012) use a very similar aggregate demand structure but with savers and borrowers (instead of spenders). They show that a deleveraging shock can throw the economy into a liquidity trap; the amplifica-
In this paper, I first propose a "New Keynesian Cross" for the analysis of heterogeneous-agent models that consists of a Planned Expenditure curve, PE for short (pictured in Figure 1 further below). I will show that the slope of this curve is the share of the indirect effect in total, while the shift of the curve (in response to monetary policy changes) is the direct effect. I analyze several heterogeneous-agent models through this lens and calculate in closed form these effects and decompositions—starting from the RANK benchmark whereby, as emphasized by KMV already, the indirect effect is virtually zero.

I first show that (and how) the earlier, 2000s-vintage TANK models are useful for understanding some key mechanisms of the new generation of 2010s-vintage HANK models. Specifically, the baseline TANK model (whereby hand-to-mouth households’ income is endogenous because they are employed) features amplification of monetary policy shocks, and this amplification is driven by the indirect effect (in KMV’s terminology). For the NK cross, the TANK model implies a steeper PE curve—much like the old Keynesian cross implies a steep PE curve when the marginal propensity to consume (MPC) increases. This holds true here too: when we add households with unit MPC (out of their own income), aggregate MPC increases.

The keystone for this mechanism of indirect-effect driven amplification is the "their own" qualification in the previous bracket. For what delivers this amplification, in addition to the mere addition of hand-to-mouth agents, is that their income respond to the cycle more than one-to-one. How constrained households’ income is related to aggregate income depends crucially on income distribution generally, and on fiscal redistribution in particular—thus echoing Keynes’ insights cited at the outset. In short, the NK-cross amplification occurs by the interaction of (i) hand-to-mouth behavior and (ii) an income distribution such that hand-to-mouth income rises more than one-to-one with aggregate income; point (ii) requires that there be not too much redistribution in favor of the hand-to-mouth, i.e. the tax system not be too progressive.

I then provide a generalization of this model that can be seen as a simplified version HANK: not without tongue in cheek, I label it SHANK (from "simple HANK"). It is inspired by McKay, Nakamura, and Steinsson (2015, 2016—hereinafter MNS) although it is different in some key respects (in particular: income of constrained agents depends on aggregate, as in the TANK model). It applies the framework of Krusell, Mukoyama, and Smith (2011): agents are subject to idiosyncratic risk against which they attempt to self-insure by using (here) liquid bonds; to simplify and obtain analytical solutions, I then assume (following Krusell et al, Ravn and Sterk, 2013, and McKay et al among others) that these bonds are not traded in
equilibrium.\textsuperscript{7} Thus in equilibrium a fraction of agents are hand-to-mouth (as in the TANK model), while the others are savers (and stockholders) and have an Euler equation. This Euler equation now takes into account the possible transition to the "constrained", hand-to-mouth state—unlike the TANK model (nested here when idiosyncratic shocks become permanent, which eliminates self-insurance).\textsuperscript{8} This model is also related to a more general framework analyzed by Werning (2015); and finally it draws on Bilbiie and Ragot (2016)—which focuses on equilibrium liquidity used to self-insure, and the optimal design of monetary policy.

This model delivers a consumption function and aggregate Euler equation with discounting, just as in MNS—but only when the elasticity of hand-to-mouth income to aggregate income is less than unity (a special case of which is MNS’s, where income of constrained is a fixed unemployment benefit or home production and thus invariant to the cycle). Whereas when said elasticity is larger than unity, there is instead compounding, or "inverse discounting": today’s consumption increases more than one-to-one in response to good news about future aggregate consumption. The intuition for this compounding relies on the self-insurance mechanism inherent in these models: good news about aggregate income in the future mean disproportionately more good news in the hand-to-mouth state, and thus dis-saving (less demand for self-insurance). With zero equilibrium savings, today’s consumption must go up and income adjusts upwards to deliver this. Finally, the unitary elasticity case (whereby hand-to-mouth income elasticity to aggregate is one) is trivially equivalent to full insurance, just as in the TANK model. This holds true in the more general framework studied by Werning (2015), where amplification also occurs generally when income risk is countercyclical and liquidity procyclical.

The implications for monetary policy shocks of the SHANK model are nevertheless very similar to the TANK model. There is more amplification, but only for persistent shocks. Where the distinction between the two models (and hence: the self-insurance against idiosyncratic risk margin) matter most is when it comes to future monetary policy, aka forward guidance FG. In the TANK model, the amplification of FG does not go beyond the amplification of contemporaneous policy changes: with more hand-to-mouth, monetary policy is more powerful uniformly, at all horizons.

With idiosyncratic risk and self-insurance (SHANK), we need to distinguish again the

\textsuperscript{7}Gornemann, Kuster, and Nakajima (2013), Den Haan, Riegler, and Rendahl (2016), Ravn and Sterk (2013), Bayer et al (2016), and Challe et al (2016) built models with endogenous unemployment risk based on search and matching. I abstract from this important (complementary) channel here. See Ravn and Sterk (2016) for some analytical results in such a model.

\textsuperscript{8}Curdia and Woodford (2009) and Nistico (2016) also study NK models with this "infrequent participation" metaphor due to Lucas (1990); their focus is different, and in their models constrained agents are in fact borrowers (they borrow in equilibrium subject to a spread).
two cases. With "discounting" in the Euler equation, I recover MNS's result that the power of FG is dampened: in particular, the total effect on present activity decreases, the further FG is pushed into the future—thus solving what Del Negro, Giannoni and Patterson (2013) called "the FG puzzle" (that in the representative-agent NK model FG power increases when it is pushed into the future). But with compounding in the Euler equation, the opposite is true: FG power increases when pushed into the future, so the FG puzzle is aggravated (this also holds in Werning's more general framework). In a companion paper (Bilbiie, 2017) I analyze this in detail and show that despite this radically different amplification (and aggravation of the FG puzzle), the optimal welfare-maximizing duration of FG is not much affected—the intuition being that there is a "dark side" to FG: when its power increases, so does its welfare cost.

2 The New Keynesian Cross and Direct-Indirect Effects: the Representative-Agent Benchmark

To set the stage and introduce the key concepts, consider the representative-agent model first. I show in Appendix A using standard intertemporal budget constraint algebra that the "consumption function" for an agent $j$ who takes as given the interest rate and her income is, loglinearized around a steady-state equilibrium:

$$c_j^t = (1 - \beta) y_j^t - \sigma \beta r_t + \beta E_t c_{t+1}^j.$$  

The other key equation is the Euler equation, or IS curve, obtained by further imposing market clearing—which with a representative agent is also the definition of income, $c_j^t = y_j^t$:

$$c_i^t = E_t c_{t+1}^i - \sigma r_t$$

Following KMV I compute the total effect, and the decomposition between direct and indirect effects, of an exogenous change in monetary policy summarized by a decrease in the real interest rate $r_t$ meant to capture for instance more expansionary monetary policy. The

---

9 There is no pretension of novelty regarding this section: it reiterates some findings of Kaplan, Moll and Violante (in continuous time), and casts it into the NK cross framework.

10 The first examples I know of of such loglinearized intertemporal budget constraints leading to consumption functions are Campbell and Mankiw (1989, 1990, 1991) and Gali (1990, 1991)—although the idea has an illustrious history the details of which can be found in Campbell and Mankiw, 1989. For other recent uses in different contexts see Garcia-Schmidt and Woodford (2014), Gali (2016), and Farhi and Werning (2017).

11 In Appendix A.2 I solve the complete forward-looking model with a Phillips curve and Taylor rule (this is standard textbook material, for instance Gali (2008), and Woodford (2011) in the context of fiscal multipliers). Here, I abstract from the exact equilibrium mechanism by which the real interest rate is
results for a shock with exogenous persistence $p$ are in Proposition 1.\textsuperscript{12}

**Proposition 1** In the representative-agent NK model, in response to an interest rate cut of persistence $p$, the total effect $\Omega$ and indirect effect share $\omega$ are:

$$
\Omega = \frac{\sigma}{1 - p},
\omega = \frac{1 - \beta}{1 - \beta p}.
$$

This preliminary proposition is essentially a discrete-time version of KMV’s decomposition in the representative-agent model. The total effect, denoted by $\Omega$, is obtained by imposing market clearing, or in other words directly from the Euler-IS equation, as: $\Omega \equiv \frac{dc_t}{\bar{d}(-r_t)}$ which leads to the above expression. The direct effect ($\Omega_D$) is the partial derivative of the consumption function, keeping $y^j_t$ fixed: $\Omega_D \equiv \frac{dc_t}{\bar{d}(-r_t)}|_{y^j_t=\bar{y}} = \frac{\sigma^2}{1 - \beta p}$. Conversely, the indirect effect ($\Omega_I$) is the derivative along the path where $c^j_t = y^j_t$, but the interest rate is kept fixed: $\Omega_I \equiv \frac{dc_t}{\bar{d}(-r_t)}|_{r_t=\bar{r}} = \frac{1 - \beta}{1 - \beta p} \frac{\sigma}{1 - p}$; naturally, this is also given by the difference between total and direct, $\Omega_t = \Omega - \Omega_D$. Finally, the relative share of the indirect effect is $\omega \equiv \frac{\Omega_I}{\Omega}$ given in the Proposition. Notice that as $p$ increases, the indirect effect becomes stronger.

A useful benchmark is that of iid shocks—which allows to abstract from the effects of persistence and use these concepts in order to gauge endogenous amplification.\textsuperscript{13} When $p = 0$, the total effect is $\Omega = \sigma$, and the indirect share $\omega = 1 - \beta$; this is the first result emphasized by KMV: with discount rate close to 1, the indirect effect is almost absent in the representative-agent NK model.

Consider the following picture, a familiar-looking Keynesian cross.\textsuperscript{14} The key equation throughout the paper is the one delivering the upward sloping line labelled PE: like the planned expenditure line from the standard textbook ("old") Keynesian cross diagram, it expresses consumption (aggregate demand) as a function of current income, for a given real interest rate:

$$
c_t = \omega y_t - (1 - \omega) \Omega r_t,
$$

\textsuperscript{12}Since there is no endogenous state variable and hence no endogenous persistence, it follows that $p$ is also the persistence of any endogenous variable.

\textsuperscript{13}This is the case considered by Auclert (2016) for a different decomposition in a richer HANK framework.

\textsuperscript{14}The genesis of this representation is in a handwritten comment by Jordi Gali on my 2004 PhD thesis; I included this in a revision of the paper (Bilbiie, 2004), using the terminology "Keynesian cross", but that did not make it into the 2008 JET published version at the insistence of a referee who demanded that "Keynesian" (which appeared in the title) and "Keynesian cross" be eliminated altogether.
I will show that not only in the baseline RANK model just studied but also in several heterogeneous-agent models reducible to this form, $\Omega$ is generally the total effect of an interest rate change on aggregate demand, while the slope $\omega$ captures the share of what KMV call the indirect effect in total $\Omega$. The shift of the PE curve will hence be the direct effect $(1 - \omega)\Omega$. A cut in interest rates translates the PE curve upwards (by $(1 - \omega)\Omega$) and the equilibrium moves from the origin to the intersection of the dashed PE curve and the 45 degree line. The rest of the paper is devoted to the analysis of the key objects $\omega$ and $\Omega$ and their determinants in a series of two-agent models. I then look at a topical application: the effects of news about future monetary policy changes, aka forward guidance FG.

The results for the RANK model are naturally interpreted through the lens of this NK cross: because the slope of PE is very close to zero, and almost all the effect of monetary policy comes from the direct shift of the PE curve, we can conclude that there is very little Keynesian about the representative-agent NK model. We now move to a model that has a very Keynesian flavor.
3 TANK: A Keynesian Model with Amplification and (through) Indirect Effect

The exposition here follows closely Bilbiie (2008) and I refer the reader to that paper for details of a more general setup where the same mechanism occurs, and for a detailed comparison with GLV. There are two key ingredients: first, one class of agents with total mass $\lambda$ is excluded from asset markets and hence has no Euler equation. Second, these same agents do participate in labor markets and make an optimal labor supply decision—their income is therefore labor income. I label these agents "hand-to-mouth", denoted by $H$. The rest of the agents $1 - \lambda$ also work and trade a full set of state-contingent securities, including shares in monopolistically competitive firms (thus receiving their profits from the assets that they price).

The linear approximation of the model is as follows. Both types’ labor supply decision $j = S, H$ is governed by (where everything is expressed in percentage deviations of steady-state aggregates): $\varphi n_t^j = w_t - \sigma^{-1} c_t^j$, with $\sigma^{-1}$ relative risk aversion, $\varphi$ the inverse elasticity of labor supply, and $n$ are hours worked, $w$ the real wage, and $c$ consumption. Assuming that elasticities are identical across agents, the same equation also holds on aggregate with the same elasticity and income effect, $\varphi n_t = w_t - \sigma^{-1} c_t$. All output is consumed and produced only by labor with constant returns $c_t = y_t = n_t$, which implies $w_t = (\varphi + \sigma^{-1}) c_t$. $H$ agents only have labor income, and they consume all of it, $c_t^H = w_t + n_t^H$. Combining this with their labor supply, we obtain their consumption function in closed-form $c_t^H = [(1 + \varphi) / (\sigma^{-1} + \varphi)] w_t$.

Hand-to-mouth thus consume all their income $c_t^H = y_t^H$, and the key word is "their": for while their consumption comoves one-to-one with their income, it comoves more than one-to-one with aggregate income. In particular,

$$c_t^H = y_t^H = \chi y_t$$

where $\chi \equiv 1 + \varphi \geq 1$

(4)

denotes the elasticity of $H$ agents’ consumption (income) to aggregate income. As will become clear momentarily, this parameter is key for the amplification effects of monetary policy in this model. It is also what distinguishes my model from earlier analyses such as Campbell and Mankiw (1989) and the literature that followed it—where it is assumed that hand-to-mouth (or, in their terminology, "rule-of-thumb") agents consume a fraction ($<1$) of aggregate income; the implications of this are discussed in more detail below. Finally, note that in general this parameter depends on fiscal redistribution: I give one particular
example in the next section, and a more general setup in the Appendix.

For S agents, we need to worry about distributional effects. The income of savers is, $y_t^S = w_t + n_t^S + \frac{1-\chi}{1-\lambda}d_t$, recognizing that they hold all the shares and thus internalizing the effect of profit income, where $d_t$ is expressed as a share of steady-state output. I approximate around a "full-insurance" steady-state whereby an optimal sales subsidy induces marginal-cost pricing and is financed by taxing firms (and thus, implicitly, savers). This further generates zero steady-state profits and, hence, full insurance (see Appendix for a more general, arbitrary redistribution scheme). Assuming for simplicity and without loss of generality constant returns to labor $y_t = n_t$, we then obtain that profits vary inversely with the real wage:

$$d_t = -w_t,$$

Savers S face an extra income effect of the real wage (which for them counts as marginal cost and reduces profits) that is the keystone for monetary transmission. Replacing $d_t$ and S agents’ labor supply schedule into their income definition, we obtain:

$$y_t^S = \frac{1 - \lambda \chi}{1 - \lambda} y_t.$$  

(5)

The additional negative income effect of wages captures the externality imposed by H agents on S agents: when demand goes up, the real wage goes up (because prices are sticky), income of H agents goes up, and so does their demand. Total demand goes up, thus amplifying the initial expansion; S agents "pay" for this by working more, which is an equilibrium outcome because their income goes down as profits fall (marginal cost goes up and, insofar as labor is not perfectly elastic $\varphi > 0$, sales do not increase by as much). By this intuition, income of savers is less procyclical the more H agents there are—and the more so, the more inelastic is labor supply.

Using the consumption function for savers, which is of the form (1) with $j = S$, we are in a position to write the aggregate consumption function:

$$c_t = [1 - \beta (1 - \lambda \chi)] y_t - (1 - \lambda) \beta \sigma r_t + \beta (1 - \lambda \chi) E_t c_{t+1}.$$  

(6)

It is important to notice that this is very different from the equation obtained by Campbell and Mankiw (1989, 1990, 1991) in their model with savers and spenders. The spenders in their model consume a constant fraction of aggregate income; this is equivalent to assuming, in my model, that $\chi = 1$, either because labor is infinitely elastic or because fiscal

---

15 This is not strictly necessary for any of the results but it simplifies the algebra.
redistribution perfectly insures agents (see below).\footnote{In their latest paper, Campbell and Mankiw (1991) do acknowledge, in a different context (of utility costs from following rule-of-thumb behavior—see their footnote 26), that under the assumption that spenders consume their own income the model behaves differently; That is the only mention of this alternative assumption (maintained throughout this paper) that is crucial for the amplification emphasized here. See Bilbiie and Straub (2012, 2013) for the implications of this different assumption for empirical estimates of $\lambda$.}

The aggregate Euler-IS equation is obtained by imposing good market clearing $c_t = y_t$ in (6):

$$c_t = E_t c_{t+1} = \frac{1 - \lambda}{1 - \lambda \chi} \sigma r_t.$$  \hfill (7)

The aggregate elasticity of intertemporal substitution—the elasticity of aggregate demand to interest rates—is increasing with the share of $H$ agents, as long as $\lambda < \chi^{-1}$. The reason is the Keynesian spiral already emphasized above: an interest rate cut implies an aggregate demand expansion, through intertemporal substitution of $S$ agents; with sticky prices, this translates into a labor demand shift, which increases the real wage. Since the wage is the $H$ agents’ income, this increases their demand further, which amplifies the initial demand expansion. This is an equilibrium: the extra output is optimally produced by $S$ agents, who face a negative income effect coming from profits (recalling that the real wage is marginal cost).\footnote{Equation (7) and this amplification mechanism are analyzed for the first time in Bilbiie (2004, 2008). But the mechanism also holds in GLV’s (2004, 2007) framework—it is just somewhat hidden because convoluted with physical capital, which in itself affects monetary transmission non-trivially (Dupor, 2001) and because the model then needs to be solved numerically. As $\lambda$ increases the amplification gets larger and larger: when $\lambda > \chi^{-1}$ an expansion cannot be an equilibrium any longer, as the income effect on $S$ agents starts dominating. That "non-Keynesian" region, whereby interest rate cuts are contractionary, is analyzed in detail in Bilbiie (2008); here we concentrate on the standard, Keynesian region throughout.}

This same intuition makes it so that the total effect of an interest rate cut is increasing with the share of $H$ agents. In terms of the NK cross in Figure 1, a higher share of $H$ agents $\lambda$ or a higher elasticity of their income to aggregate income $\varphi$ increase the slope of the PE curve—just like an increase in the marginal propensity to consume does in old-Keynesian models. The following proposition spells out the exact expressions for the total effect and indirect share in this model.

**Proposition 2** In the TANK model, in response to an interest rate cut of persistence $p$, the total effect and indirect effect share are

$$\Omega = \frac{\sigma}{1 - \lambda} \frac{1 - \lambda}{1 - \frac{1 - \lambda \chi}{1 - \lambda \chi}}.$$  

$$\omega = \frac{1 - \beta (1 - \lambda \chi)}{1 - \beta p (1 - \lambda \chi)}.$$  

The total effect $\Omega$ and the indirect share $\omega$ are both (potentially very) large in the TANK
model, and increasing with the share of constrained agents. Cast in the NK cross framework, the explanation is that the slope of the PE curve is increasing with $\lambda$: this can be seen as a reinterpretation of Bilbiie (2008) using KMV’s decomposition in this framework. The total effect $\Omega$ is evidently also increasing with the inverse labor elasticity $\varphi$. The indirect effect share is also increasing with the share of hand-to-mouth and with the inverse labor elasticity (which captures the elasticity of hand-to-mouth’s income to aggregate income). It follows directly that the direct and indirect effects are given by $\Omega_D = (1 - \omega) \Omega$ and $\Omega_I = \omega \Omega$ respectively. Notice that the direct effect decreases with $\lambda$—even though the total effect is increasing, the share $\omega$ increases faster. In other words, as $\lambda$ increases the PE curve gets steeper and steeper. These effects depend crucially on $\chi$, so we turn to its role and key determinants.

3.1 Redistribution and Monetary Policy: Amplification or Dampening?

The TANK model has a distinctly Keynesian flavor, in particular when we recall Keynes’ view that the marginal propensity to consume (the equivalent of $\omega$ in our framework) depends on income distribution, as clear in the quotes from the General Theory provided at the outset. The point that income distribution matters is a general one, and the generalization is straightforward.

Suppose that in the TANK model, some arbitrary redistribution scheme—two examples of which I provide first below, and in the Appendix—makes it so that the income of hand-to-mouth depends on aggregate income as in (4) but with $\chi$ arbitrary (to anticipate, $\chi$ will be a function of fiscal redistribution parameters, and possibly smaller than 1). The following Corollary summarizes the effects.

**Corollary 1** In the TANK model, there is amplification ($\frac{\partial \Omega}{\partial \lambda} > 0$) if and only if the elasticity of hand-to-mouth income to aggregate is greater than unity $\chi > 1$ (and dampening $\frac{\partial \Omega}{\partial \lambda} < 0$ otherwise). The indirect effect share $\omega$ is increasing with the share of hand-to-mouth $\lambda$ regardless of $\chi$.\(^{18}\)

To give one example of a fiscal redistribution scheme shaping $\chi$, consider taxing profits at rate $\tau^D$ and rebating the proceedings in a lump-sum fashion to hand-to-mouth agents;\(^{19}\) the Appendix outlines a more general redistribution. The model being otherwise unchanged (notably, steady-state profits are still zero due to the optimal subsidy—a convenient but largely innocuous assumption), we have in loglinearized form that per-capita transfers to

\[^{18}\]The derivatives are $\frac{\partial \Omega}{\partial \lambda} = \frac{\sigma}{1 - p(1 - \chi)} \chi^{-1}$ and $\frac{\partial \omega}{\partial \chi} = \frac{\beta \chi(1 - p)}{(1 - \beta p(1 - \lambda \chi))^{1/2}}$.

\[^{19}\]See Section 4.3 and Proposition 3 in Bilbiie (2008) for a first analysis of the link between the redistribution of profits and Keynesian logic.
hand-to-mouth are \( t^H_i = \frac{\tau^D}{\lambda} d_i \). It is straightforward to show that H agents’ consumption elasticity to aggregate income is then:

\[
\chi = 1 + \varphi \left( 1 - \frac{\tau^D}{\lambda} \right).
\]

The key parameter \( \chi \) (the elasticity of hand-to-mouth, constrained agents’ income to aggregate income) is lower than \( 1 + \varphi \) but higher than 1 inasmuch as \( \tau^D < \lambda \); in other words, if there is not too much redistribution, amplification still occurs. When \( \tau^D = \lambda \), all the endogenous redistributive effects emphasized here are undone, and the economy is back to the perfect-insurance, representative-agent case. Finally, when H get a share of profits higher than their share in the population \( \tau^D > \lambda \) (an example of progressive taxation) we have \( \chi < 1 \) and there is dampening instead of amplification; indeed, in the limit when \( \chi = 0 \) the total effect is scaled down by \( 1 - \lambda \) and the indirect share is the same as in the RANK model. Notice that as emphasized in the Corollary, while the total effect is decreasing in \( \lambda \) when \( \chi < 1 \), the indirect effect is still increasing in \( \lambda \): monetary policy "works" less, but it does so disproportionately more through the general equilibrium response of H agents’ income, made of labor income and fiscal redistribution. All these effects are illustrated in Figure 2 which plots the total effect and indirect share for the TANK model under the two different assumptions concerning \( \chi \) (> 1 and < 1) and distinguishing transitory and persistent policy changes.
Fig 2: $\Omega$ and $\omega$ in TANK: $\chi = 2$ (thick), 0.5 (thin), $p = 0$ (solid) and 0.8 (dash).

The more general point is that, given an income function of aggregate income for H agents, say $C^H_t = \Gamma (Y_t) + T$, a transfer will always reduce the elasticity of their after-tax income to aggregate income. In particular, the loglinearized consumption function is now, letting $\chi^0$ denote the elasticity to aggregate income without the transfer $\chi^0 = \frac{\Gamma Y}{\Gamma Y + T}$ and $\chi^T = \frac{\Gamma Y}{\Gamma Y + T}$ the elasticity with a transfer, it follows immediately that as long as the transfer is positive $\chi^T < \chi^0$. If the transfer is high enough, it can bring the model to the "dampening" region even if without a transfer it were in the "amplification region: $\chi^T < 1 < \chi^0$.

An important observation is that, when the elasticity of H income to the cycle is higher than one, the indirect effect is potentially much larger than in the RA model, even at small $\lambda$. Take for example a purely transitory interest rate shock, so that in the representative-agent model the indirect effect share is $1 - \beta$; with KMV’s calibration ($\beta = 0.95$), this indirect share is merely 0.05, while the calibrated HANK model gives an indirect share of 0.8. What is the TANK model’s indirect effect share? The key point is that it is not only proportional to the share of hand-to-mouth in the population $\lambda$.\(^{20}\) Instead, as I have just shown, the indirect share is also proportional to their income’s elasticity to aggregate income: $\omega = 1 - \beta (1 - \lambda \chi)$; thus, the TANK model delivers KMV’s HANK model’s $\omega = 0.8$ for: $\lambda = 0.4$ if labor elasticity is 1; for $\lambda = 0.26$ if $\varphi = 2$; and for a mere $\lambda = 0.13$ if $\varphi = 5$.

\(^{20}\)This is true with Campbell and Mankiw’s assumption that hand-to-mouth consume a proportional share of aggregate income, but not in the TANK model where they consume their income.
Figure 3 illustrates this further by plotting a "HANK surface": the combination of $\lambda$ and $\varphi$ that delivers $\omega = 0.8$. The solid line is under no fiscal redistribution, and the dotted line for $\tau = 0.5$; for reasons by now clear, more redistribution of the type that makes H agents’ income less cyclical implies a lower indirect effect share (and hence a higher $\lambda$ at given $\varphi$ to get the same indirect share).

Figure 3: HANK surface ($\varphi, \lambda$): $\omega = 0.8$, without (thick) and with redistribution $\tau^D = 0.5$ (thin).

A related way to understand the inherently indirect-effect-driven amplification of TANK models is emphasized in the following Corollary.

**Corollary 2** "Indirect amplification". If a TANK model gives $M$ times higher total effect than the RANK model (i.e. *amplification*), $\Omega(\lambda) = M \ast \Omega(0)$, then the indirect share is at least (for iid shocks):

$$\omega \geq 1 - \frac{1}{M}.$$  

In other words, if the total effect of a TANK model is *twice* as much as that of a RANK model, at least *half* of it is indirect; if it is four times, then at least three quarters is indirect, and so on. Note that the above is a *lower bound*, and is invariant to $\lambda$ and $\chi$. The proof is immediate: with iid shocks the ratio of the two total effects is $M = \frac{1-\lambda}{1-\lambda \chi}$. Replacing in the indirect share we have $\omega = 1 - \beta \frac{1-\lambda}{M} > 1 - \frac{1-\lambda}{M} \geq 1 - \frac{1}{M}$.\footnote{I am grateful to Davide Debortoli who suggested this interpretation for the useful special case $M = 2$.} \footnote{For persistent shocks, the lower bound is $\omega \geq \frac{1 - \frac{1}{M}}{1 - p \frac{1}{M}}$.}
4 Amplification and Dampening, Magnified: Discounting and Compounding through Self-Insurance in SHANK

McKay, Nakamura, and Steinsson (2015) in a recent influential contribution, outlined a model with incomplete markets and idiosyncratic (unemployment) risk that implies a form of "discounting" in the aggregate Euler equation. The same authors built a simplified version of their model in which the aggregate Euler equation with discounting can be solved analytically, using a set of assumptions first used by Krusell Mukoyama Smith (2011) for asset pricing and Ravn and Sterk (2012) in the context of a New Keynesian model with incomplete markets and endogenous unemployment risk. Werning (2015) uses a similar assumption to derive an aggregate Euler equation under a more general risk structure. Curdia and Woodford (2009) and Nistico (2016) are two other examples of the use of the "infrequent participation" device (introduced by Lucas, 1990) in models with nominal rigidities, albeit in a different context and for different questions. Bilbiie and Ragot (2016) build a model with three assets—of which one ("money") is liquid and is traded in equilibrium while the others are accessed only infrequently—and study optimal monetary policy in that framework.

Here, I outline a simple HANK model that builds on those contributions in order to perform the same analytical exercise we saw in the TANK model; in particular, I use an "infrequent participation" structure similar to Bilbiie and Ragot (2016) but, as in the other papers cited in the previous paragraph, with no trade (no equilibrium liquidity)—even though it distinguishes, like the HANK model, between liquid assets (bonds) and illiquid assets (stock). Inequilibrium, there is thus infrequent (limited) participation in the stock market.

There are two states, as in the TANK model: savers S and hand-to-mouth H. But unlike in the TANK model, there is now idiosyncratic risk: agents switch states following a Markov chain. The probability to stay type $S$ is $s$ and the probability to stay type $H$ is $h$ (while the transition probabilities are respectively $1 - s$ and $1 - h$), and by standard results the mass of $H$ is:

$$\lambda = \frac{1 - s}{2 - s - h},$$

with the stability condition:

$$s \geq 1 - h.$$  

Notice that this nests the TANK model when idiosyncratic shocks are permanent, $s = h = 1$: the share of $H$ $\lambda$ stays at its initial value and is a free parameter. At the other extreme, idiosyncratic shocks are iid when $s = 1 - h$: the probability for a household to be $S$ or $H$ is

\[^{23}\text{The authors use this to argue that a calibrated version of their model resolves the Forward Guidance puzzle, the unrealistically large power of forward guidance in the the RANK model. The mechanism through which this happens is precisely the discounting in the aggregate Euler equation.}\]
tomorrow is independent on whether it is S or H today.

There are two assets: liquid public bonds (that will not be traded) and illiquid stock that can only be accessed when S. S households can thus infrequently become H and self-insure through bonds (liquidity), leaving their illiquid stock portfolio temporarily. The price for self-insurance is the interest rate on bonds that are not traded. The following Euler equation governs the bond-holding decision of S households who self-insure against the risk of becoming H:

\[ (C_t^S)^{-\frac{1}{2}} = \beta E_t \left\{ (1 + r_t) \left[ s \left( C_{t+1}^S \right)^{-\frac{1}{2}} + (1 - s) \left( C_{t+1}^H \right)^{-\frac{1}{2}} \right] \right\}. \]

I therefore assume that in the H state the equivalent Euler equation holds with strict inequality: households are constrained, or impatient, and become hand-to-mouth thus consuming all their income \( C_t^H = Y_t^H \).

Loglinearizing around the same symmetric steady state \( C^H = C^S \) as in the HANK model, the self-insurance equation is:

\[ c_t^S = s E_t c_{t+1}^S + (1 - s) E_t c_{t+1}^H - \sigma r_t. \]

Replacing the consumption function of H that is identical to previously (4): \( c_t^H = y_t^H = \chi y_t \) (whatever the redistribution scheme determining \( \chi \), be it the one Section 3.1 or the one in the Appendix) we obtain the the aggregate Euler-IS:

\[ c_t = \delta E_t c_{t+1} - \sigma \frac{1 - \lambda}{1 - \lambda \chi} r_t, \]

where \( \delta \equiv \frac{s + (1 - \lambda - s) \chi}{1 - \lambda \chi}. \)

Several remarks are in order. First, in the TANK limit of the previous section (permanent idiosyncratic shocks \( s = h = 1 \)) we have no discounting \( \delta = 1 \), and \( \lambda \) is then an arbitrary free parameter. In the other extreme (the iid idiosyncratic uncertainty special case \( s = 1 - h \) of HANK, e.g. Krusell Mukoyama Smith, McKay, Nakamura and Steinsson, etc.) we have \( \lambda = h \) and \( \delta = \frac{1 - \lambda}{1 - \lambda \chi} \). The striking implications for the aggregate Euler equation are summarized in the following Proposition.

**Proposition 3** "Discounting and Compounding." In the SHANK model, distinguish

---

24 One justification for this could be that the idiosyncratic shock is a preference shock to \( \beta \) rendering households impatient "enough" to make the constraint bind.

25 For a HANK model with endogenous unemployment where unemployment benefits are cyclical depending on the real wage, see Den Haan et al, 2016.
two cases according to whether the elasticity of H’s income to aggregate is less or greater than unity:

Case 1. \( \chi < 1 \) \( \rightarrow \) the aggregate Euler equation features **discounting** (\( \delta < 1 \)); the discounting effect is magnified by idiosyncratic risk \( \partial \delta / \partial s > 0 \).26

Case 2. \( \chi > 1 \) \( \rightarrow \) the aggregate Euler equation features **compounding** (\( \delta > 1 \)); the compounding effect is magnified by idiosyncratic risk \( \partial \delta / \partial s < 0 \).

Case 1 corresponds to the finding of MNS, which is strictly nested here for \( \chi = 0 \) (implying \( \delta = s \)) and iid idiosyncratic shocks (so \( s = 1 - h = 1 - \lambda \)). When good news about future aggregate income/consumption arrive, households recognize that in some states of the world they will be constrained and seek to self-insure against this idiosyncratic risk; but this "precautionary" increase in saving demand cannot be accommodated (there is no asset), so the household consumes less today. Income adjusts accordingly to give the household the right incentives for this allocation. The higher the risk \((1 - s)\), the more discounting (the lower is \( \delta \)); in the limit as idiosyncratic shocks become permanent the self-insurance channel disappears and we recover the TANK model \( \delta \rightarrow 1 \).

The opposite holds in case 2, when \( \chi > 1 \): the effect of monetary policy is amplified—on the one hand through the elasticity to interest rate (as previously emphasized in Bilbiie, 2008 and above) but also, more surprisingly, through overturning the "discounting" effect discovered by MNS. The endogenous amplification through the Keynesian cross now holds not only contemporaneously, but also for the future: good news about future aggregate income increase today’s demand because they imply less need for self-insurance, precautionary saving. Since future consumption in states where the constraint binds over-reacts to good "aggregate news", households internalize this by attempting to self-insure less. But the precautionary saving still needs to be zero in equilibrium, so households consume more and income increases to deliver this, thus delivering amplification. This effect is magnified with higher risk (higher \(1 - s\)): the highest compounding is obtained in the iid case, because this corresponds to the strongest self-insurance motive (at given \( \chi \)).

Lastly, note that like the TANK model, this model too trivially nests the representative-agent NK model when the constrained agents’ income elasticity to aggregate income is unitary, \( \chi = 1 \)—for instance because labor is infinitely elastic \( \varphi = 0 \) or the redistribution scheme implies \( \tau^D = \lambda \). In that case, agents are perfectly insured through either labor supply or the tax system. This result, as well as the finding emphasized in the previous paragraph (of more amplification with more constrained agents, partly because of the intertemporal, self-insurance dimension) also echoes results recently obtained by Werning (2015) in a more

---

26 More precisely: \( \partial \delta / \partial s = \frac{1 - \chi}{1 - \lambda \chi} \).
general environment where this amplification arises when income risk is countercyclical and liquidity procyclical (my simple framework abstracts from the latter).

While the Euler-equation representation seems particularly useful to understand the possibility of compounding, in this model too we can recover the PE curve, or consumption function (whose cumbersome derivation is in the Appendix):

\[ c_t = [1 - \beta (1 - \lambda \chi)] y_t - (1 - \lambda) \sigma \beta r_t + \beta \delta (1 - \lambda \chi) E_t c_{t+1}. \]

(9)

Using this, the following Proposition summarizes the effect of monetary policy in this model.

**Proposition 4** In the SHANK model, in response to an interest rate cut of persistence \( p \), the total effect and indirect effect share are:

\[
\Omega = \frac{\sigma}{1 - \delta p} \frac{1 - \lambda}{1 - \lambda \chi},
\]

\[
\omega = \frac{1 - \beta (1 - \lambda \chi)}{1 - \beta \delta p (1 - \lambda \chi)}.
\]

Figure 4 illustrates and summarizes these findings; it plots the total and indirect effect in the SHANK model as a function of the share of hand-to-mouth, for several cases, assuming that the persistence of the policy change is \( p = 0.8 \) (with iid monetary policy shocks the two models trivially coincide). With red dashed line we have the TANK limit of the SHANK model \( (s = h = 1) \), distinguishing between \( \chi > 1 \) and \( \chi < 1 \): as we saw above, in the former case there is amplification and in the latter dampening, and the share of the indirect effect increases with \( \lambda \). These effects are amplified when moving towards higher risk (higher \( 1 - s \)). In the limit when \( 1 - s = h = \lambda \), represented by blue dots, we have the highest compounding and the fastest discounting.
It appears as though, when it comes to the effects of monetary policy changes, the difference between the SHANK and TANK models is mainly quantitative. The main difference is that in the "compounding" case of the SHANK model, there are two sources of amplification: the first is as in the TANK model, through increasing the contemporaneous elasticity of aggregate demand to interest rates (conversely, the slope of the PE curve in the NK cross is unchanged). The second is through the compounding/inverse discounting effect $\delta$, which only applies to future changes (i.e. if policy changes are persistent). This is of particular importance when studying news shocks and announcement of future policy changes, aka "forward guidance": we study this below, after briefly touching upon the implications of the discounting/compounding for interest rate rules.

4.1 Self-insurance and the Taylor principle

To study the stability properties of interest rate rules succinctly, I add back the simplest possible supply block: a contemporaneous Phillips curve $\pi_t = \kappa c_t$ and a Taylor rule $i_t = \phi E_t \pi_{t+1}$. While clearly over-simplified, this setup nevertheless captures a key mechanism of the NK model, i.e. a trade-off between inflation and real activity; results are conceptually very similar when considering a more standard Phillips curve adding future expected inflation.
Replacing in the aggregate Euler equation we obtain:

\[ c_t = \left( \delta - \sigma \frac{1 - \lambda}{1 - \lambda \chi} (\phi - 1) \kappa \right) E_t c_{t+1} \]

The requirement for existence of a (locally) unique rational expectations equilibrium, aka "determinacy", is that the root be outside the unit circle, i.e.:

\[ \phi > 1 + \frac{\delta - 1}{\sigma \kappa} \frac{1 - \lambda \chi}{1 - \lambda}. \]

It is evident that in the discounting case, the threshold \( \phi \) is \textit{weaker} than the Taylor principle \( (\phi > \phi^{TP} = 1) \), while in the compounding case it is \textit{stronger}. Written differently:

\[ \phi > 1 + (\chi - 1) \frac{(1 - s)}{\sigma \kappa (1 - \lambda)} \]

The intuition is clear: in the "compounding" case, there is a more powerful demand amplification to sunspot shocks; this raises the need for a more aggressive response in order to rule out sunspot equilibria. The higher the risk \((1 - s)\) and the higher the elasticity of \(H\) income to aggregate \(\chi\) the higher this endogenous amplification, and the higher the threshold. The opposite is true in the "discounting" case, since the transmission of sunspot shocks on current demand is dampened.

5 Future Monetary Policy (Forward Guidance)

The difference between the TANK and SHANK models (i.e. the persistence of idiosyncratic shocks) matters most when it comes to future monetary policy, aka forward guidance FG. This is natural, since we saw that the key logic explaining discounting/compounding goes through the effect of "news", and FG is nothing else than a special type of news. I therefore briefly characterize the implications for the effects of \textit{forward guidance}, taking the iid case for simplicity: at \(t + T\) there is a shock that lasts for one period. In a separate paper (Bilbiie, 2017), I consider a more general case and characterize analytically the effects of arbitrary FG in a liquidity trap, and the optimal design of FG policy in these models.

Since it nests the TANK (and, trivially, RANK) model, let us work directly with the encompassing SHANK model. To find the effect of FG, we iterate the PE curve or consumption function of this model (9) to obtain:

\[ c_t = - (1 - \lambda) \beta^\sum_{i=0}^{\infty} \left[ \beta \delta (1 - \lambda \chi) \right]^i E_t r_{t+i} + [1 - \beta (1 - \lambda \chi)] \sum_{i=0}^{\infty} \left[ \beta \delta (1 - \lambda \chi) \right]^i E_t y_{t+i}. \quad (10) \]
Direct differentiation with respect to a one-time interest rate cut at $t+T$ delivers the following Proposition.

**Proposition 5** In response to FG (an interest rate cut in $T$ periods) the total effect and indirect effect share are:

$$\Omega^F = \frac{1 - \lambda}{1 - \lambda \chi} \delta^T, \quad \omega^F = 1 - [\beta (1 - \lambda \chi)]^{1+T}. $$

Specifically, for any $k$ from 0 to $T$ the total effect is (by direct differentiation of the forward-iterated Euler equation (8)) $\Omega^{F(k)} \equiv \frac{d^c_{t+k}}{d(-r_{t+T})} \delta^{1+k},$ for any $k$ from 0 to $T$. The direct FG effect $\Omega^F_D$ corresponds to the derivative of the first sum in (10): $\Omega^F_D \equiv \frac{d^c_{t+k}}{d(-r_{t+T})} \bigg|_{y_{t+k}=y} = \beta \sigma (1 - \lambda) [\delta \beta (1 - \lambda \chi)]^T$. The indirect FG effect corresponds to the second term in (10): $\Omega^F_I \equiv \frac{d^c_{t+k}}{d(-r_{t+T})} \bigg|_{r_{t+k}=r} = \frac{1 - \lambda}{1 - \lambda \chi} \sigma \delta^T \left[1 - [\beta (1 - \lambda \chi)]^{1+T}\right]$, which delivers the indirect share in the Proposition.\(^\text{27}\)

To understand the results, it is useful to start from the RANK limit ($s=1$ and $\lambda=0$). Notice that the total effect of one-time FG is invariant to time, which is one instance of the FG puzzle emphasized by Del Negro et al (2013): the interest rate cut has the same effect regardless of whether it takes place next period, in one year, or in one century. Furthermore, the indirect effect’s share increases, the further FG is pushed into the future ($\omega^F$ is increasing with $T$).

Take now the TANK special case ($s=h=1$, $\lambda$ arbitrary). As for within-period policy changes, the total effect $\Omega^F$ is larger—but it is still time-invariant, i.e. it is the same for any $k$ from 0 to $T$. The same insights as for iid monetary policy shocks apply: higher $\lambda$ results in higher total effect, higher indirect effect and lower direct effect, and higher indirect effect share. In addition, the indirect effect share is increasing with time, just as—but at a faster rate than—in the RANK model. The key point is that in the TANK model forward guidance is more powerful than in the RANK model, but this has no impact on the way in which the total effect depends (not) on the horizon of FG.

The main novel insight from the SHANK model, as found by MNS is to break this invariance: the effect of forward guidance is no longer time-invariant, because of discounting. However, as holds true in Werning’s (2015) more general setting, this insight is overturned.

\(^{27}\)Garcia-Schidt and Woodoford (2014) also use a version of the forward-iterated consumption function to compute the effects of FG under finite planning horizon using a notion of "reflective equilibrium". See also Farhi and Werning (2017) for combining incomplete markets with a version of that information imperfection, i.e. "k-level thinking", that delivers a complementarity. The last paper also derived independently the analytical expressions found here for the simple RANK case.
if, as in the TANK model, the income of hand-to-mouth covaries with aggregate income more than one-to-one. Direct inspection of the expressions in Proposition 5 unveils that when $\chi < 1$ and $\delta < 1$ (and decreasing with $\lambda$), the total, direct, and indirect effect are all decreasing with $\lambda$; furthermore, the total effect is lower when pushed far into the future, thus resolving the FG puzzle as discussed above. The indirect share increases when the horizon $T$ increases, but at the same rate as in the TANK model. Matters are different when $\chi > 1$ because of the two mechanisms: the contemporaneous (TANK) amplification, and the compounding discussed above. As a consequence, the total effect of FG now increases with $T$, which this delivers a novel side of (and thus aggravates) the "FG puzzle."\textsuperscript{28}

Figure 5 illustrates the findings pertaining to FG effects. The left column plots the total effect and the right column the indirect share, as a function of $\lambda$ (top, for $T = 10$) and $T$ (bottom, for $\lambda = 0.2$). I distinguish the two cases according to whether $\chi$ is larger (thick) or lower (thin) than unity, and plot for each case the TANK model with dash and the iid SHANK model with dots (recall that the former is the limit as $s = h = 1$ and the latter the limit as $s = 1 - h = 1 - \lambda$; the analysis and discussion in the previous section still apply). The total effect of FG increases steeply with $\lambda$, relatively more when there is more idiosyncratic risk ($1 - s$ higher), when $\chi > 1$; and it decreases with $\lambda$—relatively more when there is more idiosyncratic risk—when $\chi < 1$.

The same is true with respect to the horizon of FG: the further FG is pushed into the future, the more powerful it is. The more risk, the larger is this amplification (it disappears with no risk, i.e. in the TANK model). Conversely, when $\chi < 1$, there is dampening: the total effect decreases with the horizon, and the more so the higher the risk (it is again invariant in the TANK limit, even though $\chi < 1$ makes the effect lower in levels). The share of the indirect effect, on the other hand, is invariant to the level of idiosyncratic risk: it is increasing with both $\lambda$ and $T$; the speed with which it does so depends on $\chi$, as noted before.

\textsuperscript{28}In the companion paper Bilbiie (2017), I show that while the FG puzzle is much aggravated in both the TANK and SHANK models, optimal monetary policy in a liquidity trap nevertheless contains the duration of extra accommodation. The reason is that while FG becomes more powerful (and more needed because recessions are, by the same amplification token, also larger), there is a dark side of FG: its welfare cost is high precisely when its power is also high.
6 Conclusions

What can we learn about the workings of monetary policy in modern HANK models based on earlier TANK models? This paper proposed a New Keynesian cross for the analysis
of heterogeneous-agent models, centered on a planned expenditure PE curve that captures aggregate demand. The slope of PE is the indirect effect share (the part of the total effect that is due to the endogenous response of output), and its shift in response to a monetary policy change is the direct effect (the pure intertemporal substitution effect). This representation unveils an important amplification mechanism when hand-to-mouth households’ income is endogenous and responds to aggregate income more than one-to-one: the more constrained agents there are, the higher the slope of the planned expenditure line, and the larger the effect of monetary policy. This amplification is thus driven by the indirect effect (in Kaplan, Moll and Violante’s terminology). This mechanism is overturned—and thus there is dampening instead of amplification—when the income of hand-to-mouth agents responds to the cycle less than one-to-one. Whether that key elasticity (of hand-to-mouth income to the cycle) is larger or smaller than one depends chiefly on the details of the labor market (how much of an aggregate expansion goes to labor income) and on fiscal redistribution (how progressive is the tax system). This is an example of a more general insight on the role of redistribution, which has its origins in Keynes’ General Theory as illustrated by the opening quotes: that the marginal propensity to consume (the slope of planned expenditure in the old Keynesian cross) depends on the distribution of income.

Adding a self-insurance channel (another defining feature of HANK models), magnifies these effects further. When income of hand-to-mouth responds to aggregate income less than proportionally, there is further dampening through discounting in the Euler equation (of the type first identified in this type of models by McKay, Nakamura and Steinsson). But when income of hand-to-mouth responds more than one-to-one to the cycle, amplification is magnified—there is now compounding in the aggregate Euler equation, for future aggregate expansions imply an incentive to dis-save (less self-insurance) and thus consume disproportionately more today. This captures in a simple and intuitive way (through one parameter in the aggregate Euler equation) an amplification mechanism emphasized in a more general framework by Werning. While this further amplification does not change the effects of unanticipated monetary policy shocks significantly (even though it does of course imply that effects are different when shocks are expected to persist), it has stark implication for the effects of announcements of future monetary policy (forward guidance).

In the TANK model, the effects of forward guidance do not depend on its horizon—a property shared with the baseline representative-agent model and called by Del Negro et al "the FG puzzle". As McKay, Nakamura and Steinsson have shown, the HANK model alleviates this "forward guidance puzzle" in the discounting case: this dampening holds when the income of hand-to-mouth varies cyclically, as long as its elasticity to the cycle is
But in the compounding case, the puzzle is in fact aggravated: the power of forward guidance increases with its horizon, a direct consequence of the logic explained above. Yet despite this amplification of FG’s effects, and despite there being more scope for using FG in a liquidity trap (because the same amplification also makes ZLB recessions deeper), I show in a companion paper (Bilbiie, 2017) that the optimal duration of FG is not much increased. That is because there is a dark side to more FG power: the welfare cost of inefficient volatility once the liquidity trap is over.

References


29I abstract from FG issues pertaining to commitment and communication in environments with information imperfections, emphasized i.a. by Bassetto (2016), Wiederholt (2016), and Garcia-Schmidt and Woodford (2014); such features give rise to further dampening, as shown in the last paper.


[30] Den Haan, W. J.; P. Rendahl, and M. Riegler 2016, "Unemployment (Fears) and Deflationary Spirals" Mimeo LSE


[58] Oh, H. and R. Reis, 2012, "Targeted transfers and the fiscal response to the great recession." Journal of Monetary Economics, 59(S), S50-S64


[65] Werning, I., 2015, "Incomplete markets and aggregate demand", Mimeo MIT


A Intertemporal Budget Constraint, Euler Equation, and Consumption Function

An agent $j$ chooses consumption, asset holdings and leisure solving the standard intertemporal problem:

$$\max E_0 \sum_{t=0}^{\infty} \beta^t U(C_{j,t}, N_{j,t}),$$

subject to the sequence of constraints:

$$B_{j,t} + \Omega_{j,t+1}V_t \leq Z_{j,t} + \Omega_{j,t} (V_t + P_t D_t) + W_t N_{j,t} - P_t C_{j,t}.$$  

$C_{j,t}, N_{j,t}$ are consumption and hours worked, $B_{j,t}$ is the nominal value at end of period $t$ of a portfolio of all state-contingent assets held, except for shares in firms—likewise for $Z_{j,t}$, beginning of period wealth.\(^{30}\) $V_t$ is average market value at time $t$ of shares, $D_t$ their real dividend payoff and $\Omega_{j,t}$ are share holdings. Absence of arbitrage implies that there exists a stochastic discount factor $Q_{t,t+1}$ such that the price at $t$ of a portfolio with uncertain payoff at $t+1$ is (for state-contingent assets and shares respectively):

$$\frac{B_{j,t}}{P_t} = E_t \left[ Q_{t,t+1} \frac{Z_{j,t+1}}{P_{t+1}} \right] \text{ and } \frac{V_t}{P_t} = E_t \left[ Q_{t,t+1} \left( \frac{V_{t+1}}{P_{t+1}} + D_{t+1} \right) \right],$$

which iterated forward gives the fundamental pricing equation: $\frac{V_t}{P_t} = E_t \sum_{i=1}^{\infty} Q_{t,t+i} D_{t+i}.$ The riskless gross short-term REAL interest rate $R_t$ is a solution to:

$$\frac{1}{R_t} = E_t Q_{t,t+1}$$  \hspace{1cm} (12)

Note for nominal assets we have the nominal interest rate $\frac{1}{I_t} = E_t \frac{P_t}{P_{t+1} Q_{t,t+1}}.$

Substituting the no-arbitrage conditions (11) into the wealth dynamics equation gives the flow budget constraint. Together with the usual 'natural' no-borrowing limit for each state, and anticipating that in equilibrium all agents will hold a constant fraction of the shares (there is no trade in shares) $\Omega_j$ (whose integral across agents is 1), this implies the usual intertemporal budget constraint:

$$E_t [Q_{t,t+1} X_{j,t+1}] \leq X_{j,t} + W_t N_{j,t} - P_t C_{j,t}.$$  

\(^{30}\)We distinguish shares from the other assets explicitly since their distribution plays a crucial role in the rest of the analysis.
\[ X_{j,t} = Z_{j,t} + \Omega_j (V_t + P_t D_t) = Z_{j,t} + \Omega_j \left( E_t \sum_{i=0}^{\infty} P_t Q_{t,t+i} D_{t+i} \right) \]

\[ E_t \sum_{i=0}^{\infty} Q_{t,t+i} C_{j,t+i} \leq \frac{X_{j,t}}{P_t} + E_t \sum_{i=0}^{\infty} Q_{t,t+i} \frac{W_{t+i}}{P_{t+i}} N_{j,t+i} \]

(13)

\[ = E_t \sum_{i=0}^{\infty} Q_{t,t+i} Y_{j,t+i} \]

where

\[ Y_{j,t+i} = \Omega_j D_{t+i} + \frac{W_{t+i}}{P_{t+i}} N_{j,t+i} \]  

(14)

is income of agent \( j \). Maximizing utility subject to this constraint gives the first-order necessary and sufficient conditions at each date and in each state:

\[ \beta \frac{U_C (C_{j,t+1})}{U_C (C_{j,t})} = Q_{t,t+1} \]

along with (13) holding with equality (or alternatively flow budget constraint holding with equality and transversality conditions ruling out Ponzi games be satisfied: \( \lim_{i \to \infty} E_t [Q_{t,t+i} Z_{j,t+i}] = \lim_{i \to \infty} E_t [Q_{t,t+i} V_{t+i}] = 0 \)). Using (13) and the functional form of the utility function the short-term nominal interest rate must obey:

\[ \frac{1}{R_t} = \beta E_t \left[ \frac{U_C (C_{j,t+1})}{U_C (C_{j,t})} \right] . \]

### A.1 Loglinearized equilibrium

Denote by small letter log deviations from SS. Notice

\[ Q_{t,t+i} = \beta^i \frac{U_C (C_{j,t+i})}{U_C (C_{j,t})} \]

and in SS: \( Q_i = \beta^i \). Thus we have

\[ q_{t,t+i} = \ln \frac{Q_{t,t+i}}{Q_i} = \ln \frac{U_C (C_{j,t+i})}{U_C (C_{j,t})} = -\gamma (c_{t+i} - c_t) , \]

where we notice

\[ q_{t,t+i} = q_{t,t+1} + q_{t+1,t+2} + \ldots + q_{t+i-1,t+i} \]
The Euler equation is:

\[ r_t = -E_t q_{t,t+1} \]

Rewrite, with \( \sigma = \gamma^{-1} \)

\[ c_t = E_t c_{t+1} - \sigma r_t \]

and iterate forward, using \( q_{t,t+i} = -\sum_{k=0}^{i-1} r_{t+k} \)

\[ c_t = E_t c_{t+i} + \sigma E_t q_{t,t+i} \]

(15)

Now loglinearize intertemporal budget constraint

\[ E_t \sum_{i=0}^{\infty} \beta^i \left( q_{i,t+i}^j + c_{i,t+i}^j \right) = E_t \sum_{i=0}^{\infty} \beta^i \left( q_{i,t+i}^j + y_{i,t+i}^j \right) \]

Add to each side \( (\sigma - 1) \sum_{i=0}^{\infty} \beta^i E_t q_{i,t+i}^j \)

\[ E_t \sum_{i=0}^{\infty} \beta^i \left( \sigma q_{i,t+i}^j + c_{i,t+i}^j \right) = E_t \sum_{i=0}^{\infty} \beta^i \left( \sigma q_{i,t+i}^j + y_{i,t+i}^j \right) \]

By virtue of the Euler equation the LHS simplifies

\[ \frac{1}{1 - \beta} c_{t+1}^j = \sigma \sum_{i=0}^{\infty} \beta^i E_t q_{i,t+i}^j + \sum_{i=0}^{\infty} \beta^i E_t y_{i,t+i}^j \]

Develop RHS, use \( q_{t,t} = 0 \)

\[ \sum_{i=0}^{\infty} \beta^i E_t q_{i,t+i}^j = 0 - \sum_{i=1}^{\infty} \beta^i E_t \sum_{k=0}^{i-1} r_{t+k} = -\frac{\beta}{1 - \beta} \sum_{i=0}^{\infty} \beta^i E_t r_{t+i} \]

And replace to obtain (multiplying by \( 1 - \beta \))

\[ c_t^j = -\sigma \beta \sum_{i=0}^{\infty} \beta^i E_t r_{t+i} + (1 - \beta) \sum_{i=0}^{\infty} \beta^i E_t y_{t+i}^j \]

\[ = -\sigma r_t + (1 - \beta) y_t^j - \sigma \beta \sum_{i=1}^{\infty} \beta^i E_t r_{t+i} + (1 - \beta) \sum_{i=1}^{\infty} \beta^i E_t y_{t+i}^j \]
Now replace expression for expected consumption tomorrow
\[
\beta c_t^{j+1} = -\sigma \beta \sum_{i=0}^{\infty} \beta^{i} E_t r_{t+1+i} + (1 - \beta) \sum_{i=0}^{\infty} \beta^{i+1} E_t y_{t+1+i}^j
\]
to obtain the consumption function in text 1.

A.2 Solving the baseline NK model

A full solution of the NK model can be obtained by standard, textbook methods (Woodford, 2003; Galí, 2008; Woodford, 2011). Completing the model with a Phillips curve and a forward-looking Taylor rule (with \( r_t = i_t - E_t \pi_{t+1} \) denoting the ex-ante real rate):

\[
\pi_t = \beta E_t \pi_{t+1} + \kappa c_t
\]

\[
r_t = (\phi - 1) E_t \pi_{t+1} - \epsilon_t,
\]

and assuming persistence \( p \) for the exogenous process of the interest rate cut \( \epsilon_t \) we have (because there is no endogenous state variable, so no endogenous persistence, \( p \) is also the persistence of endogenous variables) \( \pi_t = \frac{\kappa}{1-\beta p} c_t \) and \( r_t = \frac{(\phi-1)\kappa p}{1-\beta p} c_t - \epsilon_t \), implying the solution:

\[
c_t = \frac{1 - \beta p}{(1 - p)(1 - \beta p) + \sigma (\phi - 1) \kappa p} \sigma \epsilon_t,
\]

which nests the result in Proposition 1 for \( \phi = 1 \) or \( \kappa = 0 \).

To solve for the effects of forward guidance (a one-time shock to \( \epsilon_{t+T} \)), rewrite the model as

\[
x_t = Ax_{t+1} + B \epsilon_t
\]

where \( x_t = (c_t, \pi_t)' \) and:

\[
A = \begin{pmatrix} 1 & -\sigma (\phi - 1) \\ \kappa & \beta - \kappa \sigma (\phi - 1) \end{pmatrix}; B = \sigma \begin{pmatrix} 1 \\ \kappa \end{pmatrix}
\]

The effect of a time \( t + T \) shock at any time before that is:

\[
x_{t+T-k} = A^k B \epsilon_{t+T}
\]

\[
x_t = A^T B \epsilon_{t+T}.
\]

This nests the effect found in text when \( \phi = 1 \) or \( \kappa = 0 \). When the discount factor of
firms is zero (a special case we use extensively in Section 5) this simplifies considerably:

\[ c_t = (1 - \sigma \kappa (\phi - 1)) E_t c_{t+1} + \sigma \epsilon_t \] delivering \( c_t = (1 - \sigma \kappa (\phi - 1))^T \sigma \epsilon_t \).

## B TANK

For the TANK model, I refer the reader to Bilbiie (2008) for a detailed description of a more general environment and more microfoundations (see also and Eggertsson and Krugman, 2012). Labor supply for each agent is:

\[ U^j_C (C^j_t) = w_t U^j_N (N^j_t) . \]

Savers have access to all assets, including shares; their problem is exactly as outlined in Appendix A above replacing \( j \) with \( S \), with the only notable difference that their portfolio of shares is now in equilibrium (by share market clearing) \( \Omega_j = (1 - \lambda)^{-1} \). The budget constraint of \( H \) agents is simply \( C^H_t = w_t N^H_t + \text{Transfer}^H_t \) where \( \text{Transfer}^H_t \) are net fiscal transfers that under the baseline are zero. Labor and goods market clearing in turn imply \( \lambda C^H_t + (1 - \lambda) C^S_t = C_t \) and \( \lambda N^H_t + (1 - \lambda) N^S_t = C_t \).

Denoting \( \sigma^{-1} \equiv \frac{U^j_{CC} C_t}{U^j_C} \) and \( \varphi \equiv \frac{U^j_{CN} N_t}{U^j_N} \) (which assumes that these elasticities are symmetric across agents) we obtain \( \varphi n^j_t = w_t - \sigma^{-1} c^j_t \) as used in text. Assume further (by normalization) that preferences are such that both agents work the same hours in steady state \( N^H = N^S = N \).

The fiscal redistribution side is as follows. An arbitrary sales subsidy at rate \( x \) is financed by arbitrarily distributed lump-sum taxes \( T^H \) and \( T^S \). Pricing with the arbitrary subsidy delivers the steady-state real wage \( w = \frac{1+x}{1+\rho} \leq 1 \). The government thus spends \( xY \) in subsidy and needs to gather \( T = xY \) taxes. Taxes per agent to pay for this subsidy are thus distributed as

\[ T^H = \frac{\theta}{\lambda} xY \] and \[ T^S = \frac{1-\theta}{1-\lambda} xY \]
with \( \theta \) the share of taxes paid by \( H \). Assume that there is also an arbitrary redistribution \( \varrho \) for each agent \( H \) financed by taxing \( S \) agents \( -\frac{\lambda}{1-\lambda} \varrho \). The steady-state consumption shares for each agent are thus respectively, denoting the share of redistribution in total consumption by \( \rho = \varrho/C \):

\[ \frac{C^H}{C} = \frac{1+x}{1+\mu} - \frac{\theta}{\lambda} x + \rho \]
\[ \frac{C^S}{C} = \frac{1+x}{1+\mu} \left( 1 + \frac{\mu}{1-\lambda} \right) - \frac{1-\theta}{1-\lambda} x - \frac{\lambda}{1-\lambda} \rho \]
This arbitrary redistribution nests the one assumed in the baseline when \( x = \mu \) and all taxes to pay this are paid by savers \( \theta = 0 \). But perfect redistribution can also be attained with no subsidy \( x = 0 \) if \( \rho = \frac{\mu}{1+\mu} \).

Letting \( \eta \) denote the share of non-labor income (i.e., fiscal transfer) to labor income \( \eta = \frac{\rho - \theta x}{1+\rho} \), the loglinearized budget constraint of H is:

\[
c_t^H = \frac{1}{1+\eta} (w_t + n_t^H)
\]

which combined with their labor supply delivers \( c_t^H = \frac{1+\varphi^{-1}}{1+\eta+\varphi^{-1}\sigma^{-1}} w_t \). It is straightforward to show that the wage-hours locus is still \( w_t = (\varphi + \sigma^{-1}) y_t \), which implies the consumption function for H agents:

\[
c_t^H = \chi y_t
\]

where under this redistribution scheme we have

\[
\chi = 1 + \frac{\varphi + \sigma^{-1} - \eta}{1 + \eta + \varphi^{-1}\sigma^{-1}}
\]

It follows that the elasticity \( \chi \) is larger than 1 iff

\[
\frac{\eta}{\varphi + \sigma^{-1}} < 1
\]

or replacing \( \eta \):

\[
\rho - \frac{\theta}{\chi} x < \frac{1 + x}{1 + \mu} (\varphi + \sigma^{-1}).
\]

C SHANK

Outline of model, modified Bilbiie-Ragot with liquid bonds and no equilibrium liquidity. Loglinearization of S’s Euler equation presented in text around the symmetric steady-state delivers:

\[
c_t^S = s E_t c_{t+1}^S + (1-s) E_t c_{t+1}^H - \sigma r_t
\]

and noticing that we have, as before, whatever the redistribution scheme determining \( \chi \) (4):

\( c_t^H = y_t^H = \chi y_t \) we obtain the aggregate Euler-IS for this model (8). Using the stochastic discount factor notation, we now have

\[
\sigma q_{t,t+1}^S = c_t^S - s E_t c_{t+1}^S - (1-s) E_t c_{t+1}^H
\]
Iterating forward (note: we no longer have \( q_{t,t+i}^j = -\sum_{k=0}^{i-1} r_{t+k} \))

\[
c_t^S = s^i E_t c_{t+i}^S - \sigma \sum_{k=0}^{i-1} s^k (r_{t+k} - (1 - s) E_t c_{t+k}^H)
\]

(16)

\[
c_t^S = s^i E_t c_{t+i}^S + \sigma E_t \sum_{k=0}^{i-1} s^k (q_{t,t+k}^S + (1 - s) E_t c_{t+k}^H)
\]

(17)

Using the definition of stochastic discount factor:

\[
\sigma q_{t,t+i}^S = c_t^S - s E_t c_{t+1}^S - (1 - s) E_t c_{t+1}^H + c_{t+1}^S - s E_t c_{t+2}^S - (1 - s) E_t c_{t+2}^H + \ldots + c_{t+i-1}^S - s E_t c_{t+i}^S - (1 - s) E_t c_{t+i}^H
\]

\[
\sigma q_{t,t+i}^S + c_t^S = c_t^S + (1 - s) E_t \sum_{k=1}^{i} (c_{t+k}^S - c_{t+k}^H)
\]

Now loglinearize intertemporal budget constraint

\[
E_t \sum_{i=0}^{\infty} \beta^i (q_{t,t+i}^S + c_{t+i}^S) = E_t \sum_{i=0}^{\infty} \beta^i (q_{t,t+i}^S + y_{t+i}^S)
\]

Add to each side \((\sigma - 1) \sum_{i=0}^{\infty} \beta^i E_t q_{t,t+i}^S\)

\[
E_t \sum_{i=0}^{\infty} \beta^i (\sigma q_{t,t+i}^S + c_{t+i}^S) = E_t \sum_{i=0}^{\infty} \beta^i (\sigma q_{t,t+i}^S + y_{t+i}^S)
\]

By virtue of the Euler equation the LHS simplifies

\[
\frac{1}{1 - \beta} c_t^S + (1 - s) E_t \sum_{i=0}^{\infty} \beta^i \sum_{k=1}^{i} (c_{t+k}^S - c_{t+k}^H) = \sigma \sum_{i=0}^{\infty} \beta^i E_t q_{t,t+i}^S + \sum_{i=0}^{\infty} \beta^i E_t y_{t+i}^S
\]

\[
\frac{1}{1 - \beta} c_t^S + \frac{1 - s}{1 - \beta} E_t \sum_{i=1}^{\infty} \beta^i (c_{t+i}^S - c_{t+i}^H) = \sigma \sum_{i=0}^{\infty} \beta^i E_t q_{t,t+i}^S + \sum_{i=0}^{\infty} \beta^i E_t y_{t+i}^S
\]

Develop RHS \( \sum_{i=0}^{\infty} \beta^i E_t q_{t,t+i}^S \) using \( q_{t,t} = 0 \), this is as above in general case and replace to
obtain (multiplying by $1 - \beta$)

$$c^S_t = -(1 - s) E_t \sum_{i=1}^{\infty} \beta^i (c^S_{t+i} - c^H_{t+i}) - \sigma \beta \sum_{i=0}^{\infty} \beta^i E_t r_{t+i} + (1 - \beta) \sum_{i=0}^{\infty} \beta^i E_t y^S_{t+i}$$

$$= -\sigma \beta r_t + (1 - \beta) y^S_t - (1 - s) E_t \sum_{i=1}^{\infty} \beta^i (c^S_{t+i} - c^H_{t+i}) - \sigma \beta \sum_{i=0}^{\infty} \beta^i E_t r_{t+i} + (1 - \beta) \sum_{i=0}^{\infty} \beta^i E_t y^S_{t+i}$$

Now replace expression for expected consumption tomorrow

$$\beta c^S_{t+1} = -(1 - s) E_t \sum_{i=1}^{\infty} \beta^{i+1} (c^S_{t+i+1} - c^H_{t+i+1}) - \sigma \beta \sum_{i=0}^{\infty} \beta^{i+1} E_t r_{t+i+1} + (1 - \beta) \sum_{i=0}^{\infty} \beta^{i+1} E_t y^S_{t+i+1}$$

to obtain the consumption function:

$$c^S_t = -\sigma \beta r_t + (1 - \beta) y^S_t - (1 - s) E_t \sum_{i=1}^{\infty} \beta^i (c^S_{t+i} - c^H_{t+i}) - \sigma \beta \sum_{i=0}^{\infty} \beta^i E_t r_{t+i} + (1 - \beta) \sum_{i=0}^{\infty} \beta^i E_t y^S_{t+i}$$

or in recursive form:

$$c^S_t = -\sigma \beta r_t + (1 - \beta) y^S_t - (1 - s) \beta (E_t c^S_{t+1} - E_t c^H_{t+1}) + \beta E_t c^S_{t+1}$$

$$= -\sigma \beta r_t + (1 - \beta) y^S_t + \beta s E_t c^S_{t+1} + \beta (1 - s) E_t c^H_{t+1}$$

Aggregate and use $c^H_t = y^H_t = \chi y_t$ to obtain (using the notation for $\delta = \frac{s + (1 - \lambda - s)\chi}{1 - \lambda \chi}$)

$$c_t = [1 - \beta (1 - \lambda \chi)] y_t - (1 - \lambda) \sigma \beta r_t + \beta \delta (1 - \lambda \chi) E_t c_{t+1}.$$ 

To find the effects of FG we iterate forward:

$$c_t = -(1 - \lambda) \sigma \beta \sum_{i=0}^{\infty} [\beta \delta (1 - \lambda \chi)]^i E_t r_{t+i} + [1 - \beta (1 - \lambda \chi)] \sum_{i=0}^{\infty} [\beta \delta (1 - \lambda \chi)]^i E_t y_{t+i}$$

The total effect is calculated as:

$$\frac{dc_{t+k}}{dr_{t+T}} = -\frac{1 - \lambda}{1 - \lambda \chi} \sigma \delta^{T-k}, \quad \frac{dc_t}{dr_{t+T}} = -\sigma \frac{1 - \lambda}{1 - \lambda \chi} \delta^T$$

The direct effect only counts the first term and is

$$\frac{dc_{dir}}{dr_{t+T}} = -(1 - \lambda) \sigma \beta (\beta \delta (1 - \lambda \chi))^T$$
while the indirect effect refers to the second term:

\[
\frac{dc_t^{indir}}{dr_{t+T}} = [1 - \beta (1 - \lambda \chi)] \sum_{i=0}^{T} [\beta \delta (1 - \lambda \chi)]^i \frac{dc_{t+i}}{dr_{t+T}}
\]

\[
= -\frac{1 - \lambda}{1 - \lambda \chi} \sigma [1 - \beta (1 - \lambda \chi)] \sum_{i=0}^{T} [\beta \delta (1 - \lambda \chi)]^i \delta^{T-i}
\]

\[
= -\frac{1 - \lambda}{1 - \lambda \chi} \sigma \delta^T \{1 - [\beta (1 - \lambda \chi)]^{1+T}\}
\]

It follows immediately that the share of the indirect effect in total is \(1 - [\beta (1 - \lambda \chi)]^{1+T}\), and of the direct \([\beta (1 - \lambda \chi)]^{1+T}\).

**Table 1**: Summary of MP and FG effects

<table>
<thead>
<tr>
<th></th>
<th>TANK</th>
<th>SHANK</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect (\Omega)</td>
<td>(\frac{\sigma}{1-p} \frac{1-\lambda}{1-\lambda \chi})</td>
<td>(\frac{\sigma}{1-\delta p} \frac{1-\lambda}{1-\lambda \chi})</td>
</tr>
<tr>
<td>Indirect share (\omega)</td>
<td>(\frac{1-\lambda}{1-\lambda \chi \beta(1-\lambda \chi)})</td>
<td>(\frac{1-\lambda}{1-\lambda \chi \delta(1-\lambda \chi)})</td>
</tr>
<tr>
<td><strong>FG iid</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect (\Omega^F)</td>
<td>(\sigma \frac{1-\lambda}{1-\lambda \chi})</td>
<td>(\sigma \frac{1-\lambda}{1-\lambda \chi} \delta^T)</td>
</tr>
<tr>
<td>Indirect share (\omega^F)</td>
<td>(1 - [\beta (1 - \lambda \chi)]^{1+T})</td>
<td>(1 - [\beta (1 - \lambda \chi)]^{1+T})</td>
</tr>
</tbody>
</table>
The Puzzle, the Power, and the Dark Side: Forward Guidance Redux¹

Florin O. Bilbiie⁸

PSE and CEPR

August 2017 (First preliminary draft April 2016)

Abstract

Forward guidance (FG) is phenomenally powerful in New Keynesian models—a feature that earned the label "FG puzzle". This paper shows formally how two channels are jointly necessary to reduce FG power enough to resolve the puzzle: hand-to-mouth constrained households’ income respond to aggregate less than one-to-one; and unconstrained households self-insure idiosyncratic risk. These channels are complementary: if the former condition fails, FG power is instead amplified and the puzzle aggravated. Yet even with such puzzling amplification, optimal policy does not imply larger FG duration because FG power has a dark side: when it increases, so does its welfare cost.

JEL Codes: E21, E31, E40, E44, E50, E52, E58, E60, E62

Keywords: forward guidance; hand-to-mouth; heterogenous households; aggregate demand; self-insurance; optimal monetary policy; liquidity trap; Keynesian cross.

I am grateful especially to Jordi Galí, Benjamin Moll, and Jon Steinsson, as well as to Jess Benhabib, Davide Debortoli, Marco Del Negro, Gaetano Gaballo, Alejandro Justiniano, Virgiliu Midrigan, Emi Nakamura, Taisuke Nakata, Xavier Ragot, Gianluca Violante, Mirko Wiederholt, Michael Woodford, and participants in several seminars and conferences for comments and discussions, and Ezgi Ozsogut for research assistance. I gratefully acknowledge without implicating the support of Banque de France via the eponymous Chair at PSE, and of Institut Universitaire de France, as well as the hospitality of New York University and CREi during part of writing this paper.

¹¹I am grateful especially to Jordi Galí, Benjamin Moll, and Jon Steinsson, as well as to Jess Benhabib, Davide Debortoli, Marco Del Negro, Gaetano Gaballo, Alejandro Justiniano, Virgiliu Midrigan, Emi Nakamura, Taisuke Nakata, Xavier Ragot, Gianluca Violante, Mirko Wiederholt, Michael Woodford, and participants in several seminars and conferences for comments and discussions, and Ezgi Ozsogut for research assistance. I gratefully acknowledge without implicating the support of Banque de France via the eponymous Chair at PSE, and of Institut Universitaire de France, as well as the hospitality of New York University and CREi during part of writing this paper.

‡ orin.bilbiie@parisschoolofeconomics.eu. http://‡ orin.bilbiie.googlepages.com.
1 Introduction

Forward guidance in the standard New Keynesian NK model is nothing short of miraculous: its power, defined by the expansion in aggregate activity today induced by a marginal increase in FG duration, is increasing with said duration. More miraculously still, the power increases the further FG is pushed into the future—a feature that Del Negro, Giannoni and Patterson (2012) famously called the FG puzzle.\footnote{See Carlstrom, Fuerst, and Paustian (2015), Cochrane (2015) and Kiley (2015) for further discussions of this and other puzzling implications.} If the economy worked thus, getting out of a liquidity trap would be very easy: announce a long enough period of low interest rates, far enough into the future. Real-life experience suggests that this is not a good description of how things work: Japan’s liquidity trap duration is measured in decades, and the recent experience of most OECD countries in years. If it were so easy to end a liquidity trap, we would have found out by now. Furthermore, empirical evidence on the effects of FG does not render support to there being such power.\footnote{See Campbell et al (2016), Del Negro et al (2012); Gurkaynak et al (2005) were among the first to study the effect of monetary policy announcements. See also Filardo and Hoffman (2012) for an account of FG experiences in various countries.}

The literature focused on finding model features that imply dampening of FG power. Several such solutions exist by now, some of which I review below; this paper takes as a starting point McKay, Nakamura and Steinsson’s (2015, 2016—hereinafter MNS) prominent solution based on incomplete markets and uninsurable risk. A similar mechanism is confirmed by Kaplan, Moll, and Violante (2016—hereinafter KMV) in their different heterogeneous-agent NK (HANK) model. In a nutshell, incomplete markets induce dampening because they imply a form of discounting in the Euler equation: a cut in interest rates far into the future increases demand and income in the future, but has only a modest effect on current demand when it is discounted, because of self-insurance—households save part of the good income news. A crucial element of MNS’ framework is that the income of constrained households covaries with aggregate income less than one-to-one in equilibrium: either because of an underlying fiscal redistribution (more below), or because—if the constrained are unemployed receiving benefits or home production, as in MNS’ 2016 paper—their income is simply exogenous (hence, the elasticity to aggregate income is zero).

I first show that a model where the income of constrained covaries with aggregate income more than one-to-one has the opposite prediction: FG power is amplified, rather than dampened, and the FG puzzle is aggravated, rather than solved. There are two parts to the mechanism, and they both rely on a "New Keynesian cross" logic.

The first component, that I will refer to as the hand-to-mouth channel, generates amplifi-
cation of monetary policy changes at all horizons, uniformly, because it implies an elasticity of aggregate demand to interest rates *within the period* that is increasing with the share of hand-to-mouth households. This mechanism, first unveiled and discussed in Bilbiie (2008) but also operating the earlier model of hand-to-mouth with sticky prices of Gali, Lopez-Salido and Valles (2007), works as follows. An interest rate cut makes savers want to bring consumption forward by standard intertemporal substitution, which with sticky prices has an expansionary effect on labor demand bidding up the real wage. But the hand-to-mouth consume all their income, and this income *is* essentially the real wage, net of taxes and transfers; if hand-to-mouth’s income is related to aggregate income *more than one-to-one* (a property that crucially depends on the labor market and fiscal redistribution parameters), this amplifies the initial demand increase further, which bids up the real wage further, and so on.

The second component, that I will refer to as the *self-insurance channel*, is that when the above feature coexists with idiosyncratic risk it further generates amplification of news—and forward guidance *is* news. When a household receives good news about future aggregate income, *and* her own income comoves with the aggregate more than one-to-one in states where she becomes liquidity-constrained, she has an incentive to dis-save for precautionary self-insurance reasons, i.e. consume more. This overturns the "discounting" in the aggregate Euler equation that MNS unveiled as the key mechanism shaping the power of FG in their model.\(^3\) There is now *compounding* in the aggregate Euler equation, and FG becomes one order of magnitude more powerful. The FG puzzle is then, of course, much aggravated.

Furthermore, the two channels (hand-to-mouth, and self-insurance) are *complementary* in generating amplification, or dampening; I show formally that, because of this complementarity, *both* conditions are needed to solve the FG puzzle: some risk inducing self-insurance, and income respond sufficiently little to aggregate. In other words, if either of the conditions does not hold the FG puzzle is alive and well; and worse still, if the former condition holds but the latter does not, we are in an amplification where the complementarity also applies: FG power is *magnified*, and the FG puzzle *aggravated*.

Other frameworks also lead to a resolution of the FG puzzle through dampening FG power by generating discounting in the aggregate Euler equation. These include: OLG perpetual-youth models—Del Negro, Giannoni and Patterson, 2012 (building on previous work by Piergallini 2006 and Nistico, 2010); the notion of reflective equilibrium introduced by Garcia-Schmidt and Woodford 2014; or other behavioural assumptions based on sparsity

---

\(^3\)This second channel and its interaction with the first in shaping a "New Keynesian cross" are discussed in Bilbiie (2017) as capturing two key mechanisms of richer, more general HANK models such as KMV and MNS. Werning (2015) derives results on amplification in a more general framework with cyclical income risk and liquidity.
and leading to myopia as in Gabaix, 2016.\textsuperscript{4}

The analysis here suggests that an additional condition is necessary for resolving the puzzle, and for discounting to occur in the first place: income of constrained households should respond to aggregate income \textit{less than one-for-one}. Furthermore, this model feature is \textit{complementary} with the self-insurance motive (occurring through incomplete markets) in delivering the resolution of the FG puzzle. In parallel and independent work, Farhi and Werning (2017) also emphasize a complementarity leading to mitigation of FG power: between market incompleteness and \( k \)-level thinking (the latter being related to Garcia-Schmidt and Woodford’s notion of reflective equilibrium). That is very different from what the present paper emphasizes: \textit{both} features studied by Farhi and Werning dampen FG power, while the addition of the two leads to even more mitigation than the mere sum. The complementarity that I underline, instead, changes the \textit{sign} of the derivative: the addition two channels leads to dampening even when one of the channels by itself delivers amplification of FG power. Put differently, when the key condition emphasized here is not met—i.e. when the elasticity of constrained income to aggregate is higher than one—there is instead \textit{amplification} (instead of dampening) of FG power and thus an aggravation of the FG puzzle.\textsuperscript{5}

Yet despite the terrific amplification occurring when constrained households’ income over-reacts to aggregate, I end on a more positive and constructive note for NK models: optimal policy does \textit{not} imply longer FG, even (and in fact especially) when the implied FG power is miraculous; indeed, I show that the optimal policy in a liquidity trap implies an FG duration that is little affected by (and it often is in fact decreasing with) the share of constrained households. The reason is what I call the \textit{dark side} of FG: when its power increases, so does its \textit{welfare cost}—the inefficient consumption volatility once the trap is over (first emphasized in a representative-agent model in Eggertsson and Woodford’s seminal 2003 paper). This sharp increase in the welfare cost whenever there is amplification (the dark side) precludes a welfare-maximizing central bank from using more FG even when FG is puzzlingly powerful.

\textsuperscript{4}Matters are different in models with heterogeneous beliefs (Andrade et al, 2015), dispersed information (Wiederholt, 2015), or other information imperfections: FG power may there be dampened through agents’ misperception of it. See Bilbiie (2016) for further discussion. Here I concentrate on models that imply discounting in the Euler equation.

\textsuperscript{5}It is of course very likely that there is a further interesting complementarity by combining the two: adding \( k \)-level thinking or any of the other model features that deliver FG dampening will deliver further dampening also in my model—but only as long as the key condition on the elasticity of constrained households’ income to aggregate income holds.
2 A Simple HANK Model for Forward Guidance Analysis

Based on Bilbiie (2017), I outline a simplified HANK model, isolating two of the channels operating in richer quantitative HANK models such as the mentioned KMV and MNS papers. The model nests two simpler models existing in the literature: an earlier two-agent NK (TANK) model discussed in Bilbiie (2008, 2017) and Gali, Lopez-Salido and Valles (2007) and the "discounted Euler equation" model that MNS (2016) proposed as a simplified version of their important 2015 paper. This simplified model with incomplete markets and idiosyncratic (unemployment) risk implies a form of "discounting" in the aggregate Euler equation and can be solved analytically.

Here, I outline a simple HANK model that builds on those contributions and uses an "infrequent participation" structure similar to Bilbiie and Ragot (2016) but, as in the other papers cited in the previous footnote, with no trade (no equilibrium liquidity)—even though it distinguishes, like the HANK model, between liquid assets (bonds) and illiquid assets (stock). In equilibrium, there is thus infrequent (limited) participation in the stock market.

There are two states: savers $S$ and hand-to-mouth $H$. The source of idiosyncratic risk is that agents switch states following a Markov chain. The probability to stay type $S$ is $s$ and the probability to stay type $H$ is $h$ (while the transition probabilities are respectively $1 - s$ and $1 - h$), and by standard results the mass of $H$ is:

$$\lambda = \frac{1 - s}{2 - s - h},$$

with the stability condition $s \geq 1 - h$. This condition has an intuitive interpretation: it implies that the probability of remaining a participant/saver is higher than the probability of becoming a participant. By consequence and given the definition of $\lambda$ we also have:

$$1 - s \leq \lambda;$$

the conditional probability of becoming hand-to-mouth is lower than the unconditional probability (the share of hand-to-mouth).

---

6 For other quantitative HANK models containing these mechanisms among others, see i.a. Auclert (2016) and Gornemann, Kuester, and Nakajima (2015).

7 To achieve this, the authors use a set of assumptions first used by Krussell Mukoyama Smith (2011) for asset pricing and Ravn and Sterk (2012) in the context of a New Keynesian model with incomplete markets and endogenous unemployment risk. Werner (2015) uses a similar assumption to derive an aggregate Euler equation under a more general risk structure. Curdia and Woodford (2009) and Nistico (2016) are two other examples of the use of the "infrequent participation" device (introduced by Lucas, 1990) in models with nominal rigidities, albeit in a different context and for different questions. Bilbiie and Ragot (2016) build a model with three assets—of which one ("money") is liquid and is traded in equilibrium while the others are accessed only infrequently—and study optimal monetary policy in that framework.
Notice that this nests the TANK model when idiosyncratic shocks are permanent, \( s = h = 1 \): the share of \( H \) \( \lambda \) stays at its initial value and is a free parameter. At the other extreme, idiosyncratic shocks are iid when \( s = 1 - h \): the probability for a household to be \( S \) or \( H \) tomorrow is independent on whether it is \( S \) or \( H \) today.

There are two assets: liquid public bonds (that will not be traded) and illiquid stock that can only be accessed when \( S \). \( S \) households can thus infrequently become \( H \) and self-insure through bonds (liquidity), leaving their illiquid stock portfolio temporarily. The price for self-insurance is the interest rate on bonds that are not traded. The following Euler equation governs the bond-holding decision of \( S \) households who self-insure against the risk of becoming \( H \):

\[
(C^S_t)^{\frac{-\frac{1}{\sigma}}{1 - \pi_{t+1}}} = \beta E_t \left\{ \frac{1 + \pi_t}{1 + \frac{\pi_t}{\pi_{t+1}}} \left[ s (C^S_{t+1})^{\frac{-\frac{1}{\sigma}}{1 - \pi_{t+1}}} + (1 - s) (C^H_{t+1})^{\frac{-\frac{1}{\sigma}}{1 - \pi_{t+1}}} \right] \right\}.
\]

I therefore assume that in the \( H \) state the equivalent Euler equation holds with strict inequality: households are constrained, or impatient, and become hand-to-mouth thus consuming all their income \( C^H_t = Y^H_t \). \(^8\) Loglinearizing around a symmetric steady state \( C^H = C^S \) (which is instead achieved by assuming a steady-state redistribution scheme), the self-insurance equation is:

\[
c^S_t = s E_t c^S_{t+1} + (1 - s) E_t c^H_{t+1} - \sigma (i_t - E_t \pi_{t+1} - \rho_t) .
\]

\( i_t \) is the nominal interest rate set by the central bank and expressed in levels (i.e. the ZLB is \( i_t \geq 0 \)), \( E_t \pi_{t+1} \) expected inflation, and \( \rho_t \) an exogenous shock that is standard in the liquidity-trap literature, see below.

The other key ingredient is that all agents make an optimal labor supply decision: their income is therefore labor income, plus any fiscal transfer. Because they are standard, I present the loglinearized equilibrium conditions directly; in particular, I approximate around a "full-insurance" steady-state whereby an optimal sales subsidy induces marginal-cost pricing and is financed by taxing firms (and thus, implicitly, savers). This further generates zero steady-state profits and, hence, full insurance—a convenient but largely innocuous assumption for the results derived here. Both types’ labor supply decision \( j = S, H \) is governed by:

\[
\varphi n^j_t = w_t - \sigma^{-1} c^j_t , \text{ with } \sigma^{-1} \text{ relative risk aversion, } \varphi \text{ the inverse elasticity of labor supply, and } n \text{ are hours worked, } w \text{ the real wage, and } c \text{ consumption, with everything expressed in percentage deviations of steady-state aggregates. Assuming that elasticities are identical across agents, the same holds on aggregate: } \varphi n_t = w_t - \sigma^{-1} c_t . \text{ All output is consumed and produced using (only) labor with constant returns } c_t = y_t = n_t , \text{ which implies } w_t = (\varphi + \sigma^{-1}) c_t.
\]

\(^8\)One justiﬁcation for this could be that the idiosyncratic shock is a preference shock to \( \beta \) rendering households impatient "enough" to make the constraint bind.
Finally, around the steady state with zero profits, we have \( dt = -wt \): profits vary inversely with the real wage (\( dt \) is expressed as a share of steady-state output).

Let us now add one example of fiscal redistribution, following Bilbiie (2008, 2017): taxing profits at rate \( \tau^D \) and rebating the proceedings in a lump-sum fashion to hand-to-mouth agents. The model being otherwise unchanged, we have in loglinearized form that per-capita transfers to hand-to-mouth are \( t_t^H = \frac{\tau^D}{\lambda} d_t \). \( H \) agents’ income is then \( y_t^H = w_t + n_t^H + \frac{\tau^D}{\lambda} d_t \); while for \( S \) agents, it is \( y_t^S = w_t + n_t^S + \frac{1-\tau^D}{1-\lambda} d_t \). This shows that savers \( S \) face an extra income effect of the real wage (which for them counts as marginal cost and reduces profits) that is the keystone for monetary transmission.

Hand-to-mouth thus consume all their income \( c_t^H = y_t^H \), the key word being "their": for while their consumption comoves one-to-one with their income, it comoves more or less than one-to-one with aggregate income. In particular, combining the labor supply and budget constraint of \( H \) agents with the other equations, we immediately obtain the consumption function:

\[
\begin{align*}
c_t^H &= y_t^H = \chi y_t \\
\text{where } \chi &\equiv 1 + \varphi \left( 1 - \frac{\tau^D}{\lambda} \right) \leq 1
\end{align*}
\]

denotes the elasticity of \( H \)’s income to aggregate income, which depends on redistribution; this captures intuitively the point that the marginal propensity to consume depends on income distribution—a point that goes back to Keynes (1936). But this dependence holds more generally for arbitrary fiscal redistribution schemes. As will become clear momentarily, this parameter is key for the effects of forward guidance in this model. It is lower than \( 1 + \varphi \) but higher than \( 1 \) inasmuch as \( \tau^D < \lambda \) (in other words, if there is not too much redistribution). When \( \tau^D = \lambda \), all the endogenous redistributive effects emphasized here are undone, and the economy is back to the perfect-insurance, representative-agent case. Finally, when \( H \) get a share of profits higher than their share in the population \( \tau^D > \lambda \) (an example of progressive taxation) we have \( \chi < 1 \)—indeed, in the limit when \( \chi = 0 \) the income of \( H \) is exogenous: this happens for example if they are unemployed and benefits (or home production) do not depend on the cycle, as in MNS (2016).

---

9 This is not strictly necessary for any of the results but it simplifies the algebra.
10 See Section 4.3 and Proposition 3 in Bilbiie (2008) for a first analysis of the link between the redistribution of profits and Keynesian logic. See the Appendix in Bilbiie (2017) for a more general redistribution scheme with similar equilibrium implications.
11 This distinguishes my model (and its earlier 2008 incarnations) from earlier analyses such as Campbell and Mankiw (1989) and the literature that followed it—where it is assumed that hand-to-mouth (or, in their terminology, "rule-of-thumb") agents consume a fraction (<1) of aggregate income; the implications of this are discussed in more detail below.
The mirror image is of course what happens to savers' consumption, which follows immediately as:

\[ c_t^S = \frac{1 - \lambda \chi}{1 - \lambda} y_t. \]  

(2)

The additional negative income effect of wages captures the externality imposed by \( H \) agents on \( S \) agents: when demand goes up, the real wage goes up (because prices are sticky), income of \( H \) agents goes up, and so does their demand. Total demand goes up, thus amplifying the initial expansion; \( S \) agents "pay" for this by working more, which is an equilibrium outcome because their income goes down as profits fall (marginal cost goes up and, insofar as labor is not perfectly elastic \( \varphi > 0 \), sales do not increase by as much). By this intuition, income of savers is less procyclical the more \( H \) agents there are—and the more so, the more inelastic is labor supply.

Replacing in the loglinearized Euler equation, we obtain the aggregate Euler-IS curve:

\[ c_t = \delta E_t c_{t+1} - \eta \left( i_t - E_t \pi_{t+1} - \rho_t \right), \]  

(3)

where \( \delta \equiv \frac{s + (1 - \lambda - s) \chi}{1 - \lambda \chi} \) and \( \eta \equiv \sigma \frac{1 - \lambda}{1 - \lambda \chi} \).

In the companion paper Bilbiie (2017) I analyze the properties of this model of aggregate demand for dynamics and transmission of monetary policy shocks, focusing on a "New Keynesian cross" interpretation centered around a planned expenditure curve, or consumption function; In this paper, I study the implications for forward guidance in a liquidity trap, including its optimal design. Some of the insights discussed in the companion paper naturally apply in this context too, with important nuances emphasized below; I postpone this discussion to solving for the liquidity trap equilibrium.

To complete the model, add a standard aggregate supply side, a "New Keynesian Phillips curve" coming from a forward-looking pricing model à la Calvo-Yun or Rotemberg model:

\[ \pi_t = \beta E_t \pi_{t+1} + \kappa c_t. \]  

(4)

with slope \( \kappa = \psi (\varphi + \sigma^{-1}) \), where \( \psi = (1 - \zeta) (1 - \beta \zeta) / \zeta \) with \( \zeta \) the probability to be unable to change one’s price in the Calvo-Yun model.\(^{12}\) The specification of monetary policy defined as a choice of \( i_t \) follows below.

\(^{12}\) The slope of the Phillips curve is independent of the demand parameters \( \lambda \) and \( s \) because of the simplifying assumption of perfect consumption insurance in steady-state—see Bilbiie (2008) for elaboration in a more general case.
3 The Power of Forward Guidance Redux

Consider first the properties of a \textit{liquidity trap} in this economy, where the zero lower bound is triggered as in the seminal paper of Eggertsson and Woodford (2003): $\rho_t$ follows a Markov chain with two states. The first is the steady state denoted by $S$, with $\rho_t = \rho$, and is absorbing: once in it, there is a probability of 1 of staying. The other state is transitory and denoted by $L$: $\rho_t = \rho_L < 0$ with persistence probability $p$ (conditional upon starting in state $L$, the probability that $\rho_t = \rho_L$ is $p$, while the probability that $\rho_t = \rho$ is $1 - p$). At time $t$, there is a negative realization of $\rho_t = \rho_L < 0$, meant to capture in this reduced-form model an increase in spreads as in Woodford (2011) and Curdia and Woodford (2010). I assume for further simplification that the monetary authority tracks the natural interest rate of this economy ($r^n_t = \rho_t$) whenever feasible, meaning $i_t = \max(\rho_t, 0)$. It follows that the ZLB will bind when $\rho_t = \rho_L < 0$, while the flexible-price efficient equilibrium will be achieved whenever $\rho_t = \rho$.

Since the shock is unexpected, we can solve the model in the ZLB state, denoting by subscript $L$ the time-invariant equilibrium values of consumption and inflation therein; consumption in the liquidity trap state is:

$$c_L = \frac{\eta}{1 - p \left( \delta + \eta \frac{\kappa}{1 - \beta_p} \right)} \rho_L; \quad (5)$$

where $p \left( \delta + \eta \frac{\kappa}{1 - \beta_p} \right) < 1$ is needed to rule out expectations-driven liquidity traps.\(^{13}\) Inflation is $\pi_L = \frac{\kappa}{1 - \beta_p} c_L$.

Why an increase in the desire to save generates a recession with a binding zero lower bound in the standard NK model is much-researched territory since more than a decade: it causes excess saving and, with zero saving in equilibrium, income has to adjust downwards to give the income effect consistent with that equilibrium outcome. And if prices are not entirely fixed, there is also deflation, which—because it causes an increase in real rates when the zero bound is binding—leads to a further contraction, and so on.

In the simple HANK model, the magnitude of the liquidity trap recession depends on the three parameters $s$, $\lambda$, and $\chi$ that are key for transmission more generally. In particular, (5) elucidates how this will be different from the representative-agent model through three channels: the denominator, $\eta$ (this holds even for transitory shocks $p = 0$ and for fixed prices $\kappa = 0$); and—for persistent shocks $p > 0$ only—the rate of discounting in the Euler equation $\delta$ (even with fixed prices) and again $\eta$ but through its interaction with the Phillips

\(^{13}\)See Bilbiie (2016) for further discussion, Benhabib, Schmitt-Grohe and Uribe (2001, 2002), for the original point regarding sunspot ZLB equilibria, and Schmitt-Grohe and Uribe (2016) for a recent application.
curve slope. It is worth spending some time now understanding each channel. The main distinction is between two cases, according to whether $H$’s income elasticity to aggregate income of is lower or larger than one.

**Case 1 Amplification:** $\chi > 1$. There are two sides to amplification (understood as an LT recession that is deeper than in the RANK model). The first is the *“hand-to-mouth”* channel operating through $\eta$, with $\partial \eta / \partial \lambda = (\chi - 1) \sigma / (1 - \lambda \chi)^2 > 0$. The second is the *“self-insurance”* channel, through which there is compounding in the aggregate Euler equation ($\delta > 1$); the compounding effect is magnified by idiosyncratic risk $\partial \delta / \partial (1 - s) = (\chi - 1) / (1 - \lambda \chi) > 0$ and by the share of $H$ $\partial \delta / \partial \lambda = (\chi - 1) (1 - s) \chi / (1 - \lambda \chi)^2$.

**Case 2 Dampening:** $\chi < 1$. The *“hand-to-mouth”* channel operating through $\eta$ implies dampening as $\partial \eta / \partial \lambda < 0$. The *“self-insurance”* channel implies discounting in the aggregate Euler equation ($\delta < 1$); the discounting effect is magnified by idiosyncratic risk $\partial \delta / \partial (1 - s) < 0$ and by share of $H$ $\partial \delta / \partial \lambda < 0$.

Let us briefly discuss the intuition for the "amplification" case—the "dampening" case being merely the mirror image. Take first the hand-to-mouth channel operating through $\eta$: the aggregate elasticity of intertemporal substitution—the elasticity of aggregate demand to interest rates *within the period*—is increasing with the share of $H$ agents, as long as $\lambda < \chi^{-1}$. The reason is the "New Keynesian cross" already emphasized above and analyzed in detail in Bilbiie (2017): a fall in the natural interest rate implies an aggregate demand contraction, through intertemporal substitution of $S$ agents; with sticky prices, this translates into a labor demand contraction, which compresses the real wage. Since the wage is the $H$ agents’ income, this reduces their demand further, which magnifies the initial demand contraction. This mechanism is also at play in Eggertsson and Krugman’s deleveraging-based model of a liquidity trap, where it compounds with a debt-deflation channel; the same channel also amplifies fiscal multipliers in a liquidity trap in their framework. The borrowers whose constraint is binding at all times are, effectively, hand-to-mouth (even though their income

---

14 Some of the intuition here echoes some of the results for monetary policy transmission in Bilbiie (2017). A different channel is emphasized by Ravn and Sterk (2016) in an analytical HANK-type model abstracting from the channels emphasized here but with limited insurance of endogenous unemployment risk.

15 This is an equilibrium: $S$ households face a positive income effect from profits (recall that the real wage is marginal cost). This amplification mechanism is analyzed for the first time in Bilbiie (2004, 2008) for monetary policy shocks in normal times; it also holds in GLV’s (2004, 2007) framework but is somewhat hidden because convoluted with physical capital, which in itself affects monetary transmission non-trivially (Dupor, 2001) and because the model then needs to be solved numerically. As $\lambda$ increases the amplification gets larger and larger: when $\lambda > \chi^{-1}$ an expansion cannot be an equilibrium any longer, as the income effect on $S$ agents starts dominating. That "non-Keynesian" region, whereby interest rate cuts are contractionary, is analyzed in detail in Bilbiie (2008); here we concentrate on the standard, Keynesian region throughout.
then comprises nominal financial income that I abstract from and is at the core of Eggertsson and Krugman’s analysis).

Second, the self-insurance channel operating through $\delta$. The endogenous amplification through the Keynesian cross now holds not only contemporaneously, but also for the future—insofar as the liquidity trap is expected to persist: bad news about future aggregate income reduce today’s demand because they imply more need for self-insurance, precautionary saving. Since future consumption in states where the constraint binds over-reacts to bad "aggregate news", households internalize this by attempting to self-insure more. But since precautionary saving needs to be zero in equilibrium, households consume less and income falls to deliver this, thus magnifying the recession even further.\textsuperscript{16}

What is more, these two channels are in fact complementary—a property worth emphasizing in the following Proposition.

\textbf{Proposition 1} \textit{Complementarity} between hand-to-mouth and self-insurance. When there is amplification ($\chi > 1$), the compounding effect is increasing with $\lambda$ at a higher rate when there is more idiosyncratic risk: $\partial^2 \delta / (\partial \lambda \partial (1 - s)) = (\chi - 1) \chi / (1 - \lambda \chi)^2 > 0$. When there is dampening ($\chi < 1$), the same is true for the discounting effect (it is magnified at an increasing rate, i.e. $\delta$ decreases faster).

The reason for this complementarity is that the higher the risk (higher $1 - s$), the stronger the self-insurance motive: the highest compounding is obtained in the iid case (at given $\chi$).

Below, we will focus on two extreme special cases that are useful for intuition and already present in the literature. The first is the TANK limit with permanent idiosyncratic shocks $s = h = 1$: in that case, there is no discounting $\delta = 1$, and $\lambda$ is then an arbitrary free parameter. The hand-to-mouth channel is the only one operating. The other extreme is that \textbf{iid idiosyncratic uncertainty} with $s = 1 - h$ studied by Krusell Mukoyama Smith (2011), and used by McKay, Nakamura and Steinsson (2015, 2016) and others to study monetary policy and forward guidance. In that case, we have $\lambda = 1 - s$ and $\delta = \frac{1 - \lambda}{1 - \lambda \chi}$.

In light of the previous Proposition, this is the case that delivers the highest amplification (or dampening) because it implies the highest level of idiosyncratic risk that satisfies the restrictions. Notice that strictly speaking, the case studied by MNS (2016) is that of $\chi = 0$, which delivers most discounting.

Lastly, the expected deflation channel. A shock that is expected to persist with $p$ triggers self-insurance because of expected deflation ($\eta \frac{\kappa}{1 - \beta p}$), which at the ZLB means an increase in interest rate—so more saving, and, since equilibrium saving is zero, less consumption and

\textsuperscript{16}A related amplification channel (for interest rate shocks) also holds and is discussed in a more general framework in Werning (2015), through the interaction of countercyclical income risk and procyclical liquidity.
less income. This last effect operates in the standard representative-agent model too, but here it is amplified by the hand-to-mouth channel (it is proportional to $\eta$).

Turning the above logic over its head, in the dampening case ($\chi < 1$) the LT-recession is decreasing with $\lambda$ and $1 - s$: the more $H$ agents and the more risk, the lower the elasticity to interest rates within the period, and the lower the discount factor of the Euler equation $\delta$—both of which lead to dampening (and increasingly so when taken together, through the complementarity).

These amplification (or dampening) channels shape the effect of shocks to the natural rate of interest (and the ensuing recessions), but they also shape the effects of news about future interest rates, aka forward guidance FG, that we study next.

### 3.1 Forward guidance

I model FG stochastically through a Markov chain, as a state of the world with a probability distribution, as follows.$^{17}$ Recall that the (stochastic) expected duration of the LT is $T_L = (1 - p)^{-1}$, the stopping time of the Markov chain. After this time $T_L$, the central bank commits to keep the interest rate at 0 while $\rho_t = \rho > 0$, with probability $q$. Denote this state by $F$, and let $T_F = (1 - q)^{-1}$ denote the expected duration of FG. The Markov chain implied by this structure has three states: liquidity trap $L$ ($i_t = 0$ and $\rho_t = \rho_L$), forward guidance $F$ ($i_t = 0$ and $\rho_t = \rho$) and steady state $S$ ($i_t = \rho_t = \rho$), of which the last one is absorbing. The probability to transition from $L$ to $L$ is, as before, $p$, and from $L$ to $F$ it is $(1 - p)q$. The persistence of state $F$ is $q$, and the probability to move back to steady state from $F$ is hence $1 - q$.

Under this stochastic structure, expectations are determined by:

$$E_t c_{t+1} = pc_L + (1 - p)qc_F$$

and similarly for inflation. Evaluating the aggregate Euler-IS (3) and Phillips (4) curves during state $F$ and $L$ and solving for the time-invariant equilibria delivers equilibrium consumption (and inflation) during the forward guidance state $F$ and the liquidity trap state $L$.

$^{17}$This was introduced by Bilbiie (2016) in a representative-agent model; I refer the reader to that paper for details, robustness, and for an assessment of this optimality concept relative to Ramsey policy. The modelling of FG as a "state" is inspired by Woodford’s (2011) modelling of government spending stimulus after a liquidity trap probabilistically (in contrast to that paper, this concentrates on FG, solves for optimal duration, and adds heterogeneous households).
respectively as:

\[ c_F = \frac{\eta}{1 - q \left( \delta + \eta \frac{\kappa}{1 - \beta q} \right) \rho}; \]
\[ c_L = \frac{(1 - p) q \left( \delta + \eta \frac{\kappa}{(1 - \beta q)(1 - \beta p)} \right) \eta}{\left[ 1 - q \left( \delta + \eta \frac{\kappa}{1 - \beta q} \right) \right] \left[ 1 - p \left( \delta + \eta \frac{\kappa}{1 - \beta p} \right) \right] \rho} + \frac{\eta}{1 - p \left( \delta + \eta \frac{\kappa}{1 - \beta p} \right) \rho_L}; \]

and \( \pi_F = \frac{\kappa}{1 - \beta q} c_F, \pi_L = \frac{\kappa \beta (1 - p) q}{(1 - \beta q)(1 - \beta p)} c_F + \frac{\kappa}{1 - \beta p} c_L \).

It is immediately apparent that the future expansion \( c_F \) is increasing in the degree of forward guidance \( q \) regardless of the model. In the amplification case (\( \chi > 1 \)), the future expansion is also increasing with the share of hand-to-mouth \( \lambda \) and with risk \( 1 - s \); whereas in the dampening case, the opposite holds.

Figure 1 illustrates these findings and the next section derives analytical results useful to understand the mechanisms. Distinguishing between dampening \( \chi < 1 \) (left) and amplification \( \chi > 1 \) (right), it plots in both panels consumption in the liquidity trap (thick) and in the FG state (thin), as a function of the FG probability \( q \). We represent the RANK model by green solid lines, the TANK limit of the SHANK model \( (s = h = 1) \) with red dashed lines, and the SHANK model in the iid case \( 1 - s = h = \lambda \) with blue dots.

The pictures illustrate the dampening and respectively amplification at work: at given \( q \), low future rates have a lower effect (on both \( c_F \) and \( c_L \)) in the TANK model, and an even lower one in the SHANK model, in the dampening case. The last point illustrates the complementarity: the dampening is magnified when moving towards higher risk \( 1 - s \) and in the limit when \( 1 - s = h = \lambda \) (blue dots) we have the fastest discounting. Whereas in the amplification case (right panel), the opposite is true: low rates have a higher effect in the TANK model, and through complementarity an even higher one under self-insurance: the pictured iid case represents the highest compounding. Indeed, even though \( \chi = 2 \) is a rather conservative number and the share of \( H \) is very small (0.1)—which makes amplification in the TANK version rather limited—amplification in the SHANK model is phenomenal: the recession is 10 times (!) larger than in the RANK model (the scale on the right panel had to be changed to accommodate that).

\[ \text{18 The illustrative parametrization used in the Figure has } \beta = 0.99, \psi = 0.01, \sigma = 1, \varphi = 1, p = 0.8 \text{ and a spread shock of } \rho_L = -0.005, \text{i.e. 2 percent per annum. This delivers a recession of 4 percent and annualized inflation of 1 percent in the absence of FG } (q = 0). \text{ The domain is such that } \Gamma_q > 0. \]
Furthermore, FG power (defined as the derivative of consumption during the trap $c_L$ with respect to $q$, $dc_L/dq$) is much larger in the SHANK model in the "amplification" case. This is related to the aforementioned "FG puzzle": the higher the persistence of the trap $p$ (the further into the future FG starts), the higher the power $dc_L/dq$. Our results suggest that while a HANK model with $\chi = 0$ (or $\chi < 1$), as shown by MNS, solves this puzzle, the version with amplification $\chi > 1$ aggravates it. The next section substantiates this point analytically.

3.2 Simple Analytics of FG Power and Puzzle

Closed-form results are particularly useful here in order to shed light on the role of each amplification channel in determining FG power and potential resolutions of the "puzzle"; later, they are also useful for understanding the properties of optimal policy. Consider thus a special case: the simplest possible aggregate supply curve whereby each period a fraction
of firms $\zeta$ keep their price fixed, while the rest can re-optimize their price freely but ignoring that this price affects future demand. This delivers the Phillips curve: $\pi_t = \kappa c_t$, where now $\kappa = (\varphi + \sigma^{-1}) (1 - \zeta) / \zeta$. In Bilbiie (2016) I show that, in the context of a representative-agent NK model, this special case delivers conclusions for FG and optimal policy in a LT that are very close to those obtained using the more general NKPC (4); the insight being that FG is chiefly about aggregate demand, and not so much about supply.\footnote{Essentially, such a setup reduces to assuming $\beta = 0$ only in the firms’ problem (they do not recognize that today’s reset price prevails with some probability in future periods). See Bilbiie (2016) for an extension to the case $\beta > 0$, which makes the algebra more involved without affecting the results. In particular, Figure 2 in that paper shows that optimal FG varies very little with the discount factor of firms.}

The equilibrium found in (7) above simplifies to:

$$c_F = \frac{\eta}{1 - q\nu} \rho;$$

$$c_L = \frac{1 - p}{1 - p\nu} q\nu \eta \rho + \frac{\eta \rho_L}{1 - p\nu};$$

(8)

I defined the composite parameter:

$$\nu \equiv \delta + \eta \kappa,$$

which has an intuitive interpretation: it captures the response of consumption in a liquidity trap to news about future income/consumption.\footnote{In a "regular" equilibrium whereby the zero bound does not bind and monetary policy follows an active interest rate rule, this parameter is obviously less than one (because news about future income bring about higher real rates, through higher inflation); in particular, $\nu = 1 - \sigma \kappa (\phi - 1)$, where $\phi > 1$ is the response of nominal interest rates to expected inflation.}

We can now define **FG power** formally as:

$$\mathcal{P}_{FG} \equiv \frac{dc_l}{dq} = \frac{1}{1 - q\nu} \left( \frac{1 - p}{1 - p\nu} \right)^2 \frac{(1 - p) \nu \eta}{1 - p\nu} \rho.$$  

The properties of amplification and dampening of FG power follow the same logic as those applying to the natural rate shock and LT recessions. They are worth emphasizing in the following Corollary (which follows directly by noticing that $\mathcal{P}_{FG}$ is increasing with $\nu$, and hence with both $\delta$ and $\eta$):

**Corollary 1** In the amplification case $\chi > 1$ FG power $\mathcal{P}_{FG}$ increases with idiosyncratic risk $1 - s$ and with the share of hand-to-mouth $\lambda$ (while it decreases in the dampening case $\chi < 1$). Furthermore, the complementarity between self-insurance and hand-to-mouth also applies to FG power.

Defining formally the **FG puzzle** as the property (thus labelled by Del Negro, Giannoni
and Patterson, 2012) that the power of FG increases, the further it is pushed into the future (i.e., in our context, with the persistence of the trap \( p \)) we obtain the following Proposition:

**Proposition 2** The simple HANK model solves the FG puzzle \( (d\Pi_{FG}/dp < 0) \) if and only if:

\[
\nu < 1.
\]

Equivalently, the model needs both the self-insurance and hand-to-mouth channels:

\[
1 - s > 0 \text{ and } \chi < 1 - \sigma \kappa \frac{1 - \lambda}{1 - s} < 1,
\]

which is a manifestation of complementarity.

The result follows directly calculating the derivative \( d\Pi_{FG}/dp = (\nu - 1) \frac{\mu \rho}{(1-q_\nu)(1-p_\nu)^2} \) and then replacing the expression for \( \nu \). The Proposition emphasizes that, to solve the FG puzzle, the model needs two conditions: some idiosyncratic uncertainty \( 1 - s > 0 \), and a cyclicality of \( H \)'s income that is lower than the threshold defined above. This is a clear illustration of the complementarity emphasized in Proposition 1. In other words, having discounting in the aggregate Euler equation (\( \delta < 1 \)) is a necessary, but not sufficient condition to solve the puzzle.\(^{21}\) Rewriting the condition, we have \( \sigma \kappa < \frac{(1-s)(1-\chi)}{1-\lambda} \); this is more stringent when prices are more flexible (\( \kappa \) larger) and \( \lambda \) smaller (at given \( s \) and \( \chi \)).

To consider an even simpler example, consider the case of acyclical income of \( H, \chi = 0 \). Under that assumption the discount factor in the Euler equation is equal to the probability \( s \), and the effect of news is \( \nu = s + (1 - \lambda) \sigma \kappa \); this is not necessarily smaller than 1—for example, in the TANK model it is larger than one since \( s = 1 \). To solve the FG puzzle, there needs to be enough idiosyncratic risk, namely in this case \( 1 - s > (1 - \lambda) \sigma \kappa \). It is worth noticing that MNS’s 2016 simple model (with iid idiosyncratic risk and exogenous income of \( H \)) inherently satisfies these conditions: essentially, to \( \chi = 0 \) it adds \( s = 1 - \lambda \).

To further illustrate how the FG puzzle operates, and how the complementarity between the two channels helps eliminate it, consider Figure 2; it plots FG power as a function of \( p \), for the same calibration as before (fixing in addition \( q = 0.5 \)) in the two cases \( \chi < 1 \) and

\(^{21}\) Farhi and Werning (2017) emphasize a related but different complementarity, between market incompleteness and "k-level thinking": an informational imperfection related to Garcia-Schmidt and Woodford’s notion of reflective equilibrium, that leads to mitigation of FG—also through discounting in the Euler equation. In their framework, market incompleteness magnifies the mitigation of FG effects obtained with k-level thinking. The complementarity I emphasize is between two different channels, and can work both ways—generating more mitigation, or more amplification. Indeed, it affects not only the quantitative properties, but the qualitative insights: it changes the sign of a key derivative, as illustrated in Figure 2 below, needed to solve the FG puzzle as defined formally here.
$\chi > 1$ for the three models RANK, TANK, and iid SHANK. It illustrates clearly that it is the interaction of dampening through $\chi < 1$ and idiosyncratic risk (which, as shown above, magnifies that dampening through discounting) that leads to resolving the FG puzzle: the power of FG becomes decreasing in the duration of the trap. The dampening channel by itself (TANK model with $\chi < 1$, red dashed line on the left panel) is not enough—although it alleviates the puzzle relative to the RANK model, it does not make the power decrease with the horizon $p$. While the self-insurance channel by itself added to the "amplification" case magnifies power even further, thus aggravating the puzzle (blue dots in the right panel for the SHANK iid model in the amplification case).

Finally, it is immediate to see that the puzzle is aggravated at higher values of $\nu$ ($\frac{dP_{FG}}{dp}$ is increasing in $\nu$). It follows from the monotonicity of $\nu$ that the puzzle is alleviated with higher idiosyncratic risk $1 - s$ and with $\lambda$ in the dampening case; but worsens with idiosyncratic risk $1 - s$ and with $\lambda$ in the amplification case $\chi > 1$.

Since the power of FG is increased (and the puzzle aggravated) in the latter case, does it follow that more FG is always a good thing? That is, does this strengthen the welfare scope for FG as a feature of optimal policy emphasized since the celebrated paper of Eggertsson and Woodford (2003)? No, because there is a dark side to FG power: whenever it is high, the welfare costs of FG, as measured by future volatility, are also high. The following optimal policy exercise formalizes this intuition.
4 The Dark Side of FG Power and Optimal Policy

Optimal monetary policy in a liquidity trap is very easy to compute with this setup for FG, because equilibria are a smooth function of $q$: we can find the optimal FG duration by maximizing lifetime welfare with respect to $q$. This is shown in Bilbiie (2016) for the representative-agent model, and turns out to also be very close to the full Ramsey-optimal monetary policy taking the ZLB as a constraint calculated by Eggertsson and Woodford (2003), Jung Teranishi and Watanabe (2005) and several others since.

We thus look for $q$ that maximizes an aggregate welfare function that can be represented as a quadratic loss function and, given the Markov chain structure, it is of the form:

$$W = \frac{1}{1 - \beta p} \frac{1}{2} \left[ c_L^2 + \omega (q) c_F^2 \right],$$

where $\omega (q)$ is the appropriate discount factor for the FG state. The central bank chooses FG duration (persistence probability $q$) by solving the optimization problem $\min_q W$ taking as constraints the equilibrium values $c_F$ and $c_L$ given in (7) above. The first-order condition of this problem is:

$$c_L \frac{dc_L}{dq} + \omega (q) c_F \frac{dc_F}{dq} + \frac{1}{2} \frac{d\omega (q)}{dq} c_F^2 = 0 \quad (9)$$

and has a clear intuitive interpretation.

The first term is the welfare benefit of more forward guidance, through remedying the LT-caused recession and hence minimizing consumption volatility in the trap. This is proportional to the level of consumption in the trap: the larger the initial recession, the higher the marginal utility of an extra unit of consumption, and the larger the welfare scope of any policy that can deliver it—such as FG. The last two terms are the total cost of forward guidance: the former is the direct cost, a future consumption boom being associated with inefficient volatility; the latter is the discounting effect discussed above: the longer the time spent under FG, the larger the cost (which is proportional to consumption volatility in the

---

17

---

22 Bilbiie (2008, 2017) derives this welfare function under certain conditions that are fulfilled here (an optimal subsidy makes the steady-state efficient). Then, since the equilibrium solution is time-invariant in each of the three states, the per-period loss function is, for any state $j = \{L, F, S\}$: $\pi_j^2 + \chi c_j^2 = (\chi + \kappa^2) c_j^2$. Recall that in state $S$ the economy is back to steady state, so the loss there is zero.

23 The optimal weight is $\omega (q) = \frac{1 - \beta p + \beta (1 - p) q}{1 - \beta q}$ and counts for the times the process spends in state $F$ when starting from $F$ (given by $(1 - \beta q)^{-1}$); as well as for all the times spent in time $F$ when starting from $L$, before being absorbed into $S$ (given by $\beta (1 - p) q / ((1 - \beta p) (1 - \beta q))$. $\omega (q)$ is increasing in $q$, which is intuitive: the longer the economy spends in the F state, the larger the total welfare cost of consumption variability in that state. See Bilbiie (2016) for the details, including the second-order sufficient conditions for the RANK model (that apply here too).
Figure 3 plots the optimal FG duration (the solution of equation (9)) as a function of \( \lambda \), under our baseline parameter values, distinguishing again the dampening \( (\chi < 1, \text{left}) \) and amplification \( (\chi > 1, \text{right}) \) cases. In the dampening case, the degree of optimal FG is decreasing with the share of \( H \); the more so, the higher idiosyncratic risk; this result holds generally and can be shown analytically in the simplified version of the model—see next Section. The intuition is that all forces work in the same direction: the recession is lower to start with (which gives less scope for using FG) and the power of FG is monotonically decreasing with \( \lambda \): because the elasticity to interest rates is decreasing, in the TANK model, and in addition because of the discounting effect of MNS, in the idiosyncratic risk case.

The amplification case is, in view of our previous results, more surprising: the optimal degree of FG is almost invariant to \( \lambda \) in the TANK case (albeit initially mildly increasing) because there are two counterbalancing forces. On the one hand, the benefit component is higher: the recession is larger \( (c_L \) more negative, and a wider output gap gives more welfare reason to use FG), and the power of FG \( \frac{dc}{dq} \) is higher. But on the other hand, the cost of FG is also increasing (the last two terms in (9)). At some threshold \( \lambda \) level, the cost of FG is no longer worth bearing: the implied volatility during the FG state is so high that the optimal degree of FG drops rapidly towards zero. With idiosyncratic risk, these affects are further amplified: the higher share of \( H \) makes the recession larger and accelerates the increase in FG power, making optimal FG initially increasing; but the same amplification also holds for the welfare cost of future volatility, which kicks in earlier (at a lower share of \( H \)) and makes optimal FG drop abruptly towards zero. It is this sharp increase in the welfare cost that occurs precisely when FG power is large that I refer to as the "dark side" of FG power.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3}
\caption{Optimal FG persistence as a function of \( \lambda \) for \( \chi < 1 \) (left) and \( \chi > 1 \) (right).}
\end{figure}
In both models, it becomes optimal to do no FG at all beyond a certain threshold share of hand-to-mouth. The underlying reason is, however, very different. With dampening it is because (as MNS discovered), a higher share of $H$ implies low FG power, and because it also implies a weaker welfare-scope for using FG (the recession is lower). With amplification, it happens because a high share implies a large elasticity of aggregate demand to interest rates, and hence a high FG cost; even though FG too becomes more powerful, this effect is dwarfed by the increase in cost.

4.1 Simple Analytics of Optimal FG

Further insights can be obtained by focusing again on the simpler case studied in Section 3.2 and assuming in addition that the central bank attaches equal weights to future and present: $\omega(q) = 1, \omega'(q) = 0$ (in other words, its discount factor is also 0). This provides an upper bound on optimal FG because it ignores the second-order discounting costs.\(^{24}\) The optimal duration can now be solved in closed-form from (9) as:

$$q^* = \max \left\{ 0, \frac{1}{\nu} \frac{\Delta_L - \frac{(1-p\omega)^2}{1-p}}{1-p + \Delta_L} \right\},$$

where $\Delta_L \equiv -\rho_L/\rho > 0$ is the financial disruption causing the ZLB.

A first insight allowed by the closed-form solution is that it is optimal to refrain from FG altogether ($q^* = 0$) when $\nu$ is smaller than a certain threshold $\nu^0 \equiv \left( 1 - \sqrt{1 - p \Delta_L} \right)/p.$\(^{25}\) Recalling the expression for $\nu$, it follows immediately that (regardless of self-insurance) in the "amplification" case ($\chi > 1$) the region of $\lambda$ for which FG is optimal will be ceteris paribus smaller than in the "dampening $\chi < 1$ case ($\nu$ is increasing with $\chi$ because both $\delta$ and $\eta$ are). Moreover, in the "amplification" region $\chi > 1$ since $\nu$ is increasing both with the share of $H$ and with idiosyncratic risk $1 - s$, it follows that an increase in either of these parameters restricts the case for optimal FG.\(^{26}\) The reason is the dark side of FG: more amplification brings more welfare cost of FG. Conversely of course, in the "dampening" region $\chi < 1$ the opposite is true: $\nu$ is decreasing in both $\lambda$ and $1 - s$, and an increase in either of those parameters pushes up the threshold for FG to be optimal.

How does the optimal FG duration depend on the key structural parameters? Since we know the derivatives of $\nu$ with respect to $\lambda, s$, and $\chi$, it suffices to calculate the derivative of

\(^{24}\)See Bilbiie, 2016 for an analysis of the accuracy of this in a RANK model.

\(^{25}\)Under the baseline calibration, the threshold is 0.86.

\(^{26}\)The derivatives are $\frac{d\nu}{d(1-s)} = \frac{d\delta}{d(1-s)} = \frac{\chi - 1}{1 - \lambda \chi}$ and

$$\frac{d\nu}{d\lambda} + \frac{d\eta}{d\lambda} = (\chi - 1) \frac{\lambda(1-s) + \sigma}{(1-\lambda \chi)^2}.$$
\( q^* \) with respect to \( \nu \), which is:

\[
\frac{dq^*}{d\nu} = \frac{1}{\nu^2} \left( \frac{1 - (p\nu)^2}{1 - p} - \Delta_L \right).
\]

When the disruption causing the liquidity trap is lower than a certain threshold \( \Delta_L < (1 - p)^{-1} \), which is the more empirically plausible case,\(^{27}\) then \( q^* \) is increasing in \( \nu \) if \( \nu < \bar{\nu} \equiv \sqrt{1 - \Delta_L (1 - p)/p} \) and decreasing otherwise. Notice that this threshold is larger than the threshold needed for FG to be optimal at all \( (q^* > 0) \) derived above: \( \bar{\nu} > \nu^0 \). We have \( dq^*/d\nu > 0 \) when \( \nu^0 < \nu < \bar{\nu} \) and \( dq^*/d\nu < 0 \) when \( \nu^0 < \bar{\nu} < \nu \). It is useful to again distinguish the two cases depending on \( \chi \).

With dampening \( \chi < 1 \) more \( H \) and more risk imply that \( \nu \) is decreasing; if we start with \( \nu > \bar{\nu} \), optimal FG duration first increases, then decreases as \( \nu \) crosses the threshold. Whereas if we start below the threshold, optimal FG duration decreases uniformly (this is the case shown in the Figure). The effect is mitigated by idiosyncratic risk which, because it reduces both the power of FG and the scope for it (the LT recession is smaller) implies uniformly lower optimal duration.

With amplification \( \chi > 1 \), \( \nu \) is increasing in both \( \lambda \) and \( 1 - s \); therefore, if we start below the threshold \( \bar{\nu} \), optimal FG first increases up to a maximum level (reached at the threshold) and then decreases abruptly. Furthermore, it increases faster and reaches its maximum sooner when there is idiosyncratic risk, because of the complementarity: amplification itself is in that case magnified—by the same token, the dark side (the welfare cost of FG) suffers from the same amplification, so the point where FG ceases to be optimal is reached sooner than without risk \( s = 1 \).

4.2 A Caveat

Is not conceivable that, if FG is less effective, optimal policy should imply doing more (rather than less) of it? Nakata, Schmidt, and Yoo (2017) follow this line of reasoning in a calibrated model with a discounted Euler equation à la MNS’ that delivers FG power mitigation. The authors show that, if instead of keeping the size of the disturbance fixed—as we did above—one fixes the size of the recession (itself a function of other structural parameters), one obtains the opposite conclusion to this paper’s with \( \chi < 1 \): the optimal-policy-implied duration of FG becomes increasing in the share of constrained households.

\(^{27}\)If instead \( \Delta_L > (1 - p)^{-1} \), \( q^* \) is uniformly decreasing in \( \nu \): that is, it is decreasing in \( \chi \), \( \lambda \), and \( 1 - s \) in the "amplification" case \( \chi > 1 \). The reason is that the contractionary effect coming from the steeper recession dominates the expansionary effect of increased FG effectiveness; the opposite is of course true with \( \chi < 1 \): \( q^* \) is increasing in \( \lambda \) and \( 1 - s \).
The reason is that, as the share of constrained increases, the shock necessary to generate the given recession gets larger and larger, which adds a force calling for more optimal FG. If this force is strong enough, it can overturn the conclusion obtained above for a given shock.

I confirm Nakata et al’s conclusion in my simpler model, with \( \chi < 1 \) and little idiosyncratic risk, i.e. the TANK model (red dash in the upper left panel in Figure 4): the optimal duration becomes increasing with the share of hand-to-mouth. There is an important qualification though, afforded by the analytical framework studied here. First, notice that in the same panel the blue dotted line is increasing only slightly initially, and still decreasing thereafter: in the SHANK model (with most idiosyncratic risk and strongest self-insurance motive) there is more dampening—so while the shock necessary to reproduce a given recession is increasing in \( \lambda \) at an even faster rate, the power of FG also goes down very fast. The
FG puzzle and having optimal FG increase with the share of constrained households seem to be two sides of the same coin: in this simple model at least, you cannot throw one and keep the other.\footnote{The other qualification pertaining to this case refers to the implied shock, plotted in the lower left panel. With so much dampening as implied by the SHANK model, the shock necessary to replicate an even modest recession (4 percent here) becomes very large indeed (several times larger than the normal-times interest rate); while the shock is unobservable, this type of configuration seems unlikely.}

Moreover, the very same logic that generates increasing FG duration in the previous case is turned on its head in the amplification, $\chi > 1$ case: as the share of constrained gets larger, a smaller shock is needed to generate a given recession (lower right panel). This adds a force calling for less optimal FG, so the optimal duration is lower (and more rapidly decreasing) than in the "fixed-shock" case. And since amplification is so powerful in the SHANK model, self-insurance makes optimal FG duration decrease even faster. The general message is that keeping the observable recession fixed (rather than the unobservable disturbance) is a useful optimal policy exercise; but it does not necessarily imply a stronger case for longer optimal FG duration. Indeed, in some cases—such as the "amplification" case whereby FG power is highest and the puzzle at its most extreme—it unambiguously implies an even weaker case.

5 Conclusions

How to make forward guidance less powerful in NK models? This paper has concentrated on one stream of solutions, of the several that have been proposed over the recent years. Namely, McKay, Nakamura and Steinsson (2015, 2016), have shown that the "forward guidance puzzle" (thus labelled by Del Negro, Giannoni and Patterson, 2012) is alleviated in heterogeneous-agent New Keynesian (HANK) models. Their models lead to a dampening of the power of forward guidance; the same holds in the different HANK model of Kaplan, Moll, and Violante (2015, 2016).

In this paper, I show that it is the interaction of two complementary channels that delivers this solution: constrained (hand-to-mouth) households’ income depends on aggregate income less than one-for-one, and there is idiosyncratic risk inducing a self-insurance motive for unconstrained households. The former channel by itself does deliver dampening relative to the representative-agent model, but uniformly at all horizons and thus does not solve the puzzle. While the latter channel, because of the complementarity, magnifies the amplification inherent when hand-to-mouth income elasticity to aggregate is higher than one: FG power becomes phenomenal, and the puzzle is much aggravated.

Yet even when this (puzzling indeed) amplification of FG’s power is a model feature, and despite there being more scope for using FG—because the same amplification also makes ZLB
recessions deeper and increases the marginal utility of consumption—optimal policy does not imply a correspondingly higher duration of FG; indeed, it at least eventually implies that the optimal FG duration decrease with the share of hand-to-mouth. That is because there is a dark side to FG power: the welfare cost of inefficient volatility once the trap is over (first unveiled in Eggertsson and Woodford’s celebrated 2003 analysis in a representative-agent model). This cost becomes very large too precisely when FG power does, thus making it optimal to contain the optimal duration of extra accommodation.

Three novel policy inputs are key, in light of this analysis, for a central bank’s optimal policy in a liquidity trap: how many households are constrained; how their income is related to aggregate income (through employment or redistribution); and a measure of idiosyncratic risk.

References


