Views on the role played by banks in the economy have evolved greatly over the last 125 years, as have arguments on the need, as well as the best way, to regulate them. Some of the key insights in the debate have been published in the Journal of Political Economy. In what follows, we will outline the main contributions to the debate in recent years, with an emphasis on work done at the University of Chicago or published in the JPE. We

These views are those of the authors only and are not necessarily shared by the Bank of England or any other institutions with which we are affiliated.
want to emphasize work that has relevance today, but despite this caveat, we will probably end up doing injustice to work published long ago.

We begin with a framework for organizing the theories of intermediation. We then draw out the implications for what the theories say about regulation and note that in many respects the motivation for regulation has been only loosely tied to the theory of intermediation. We close with some open questions for regulators and economists interested in banking. We do not survey the research that has followed up on work published in the *JPE*, nor will we attempt to provide a detailed overview of the entire academic literature on banking. For that, we refer the reader to the excellent work by Gorton and Winton (2003) and Freixas and Rochet (2008).

**Theoretical Overview**

We define banks as financial institutions with a substantial fraction of illiquid assets financed with demandable liabilities payable at par. Banking theories typically have focused on one side of the bank’s balance sheet as critical to its economic role and then argued why the other side of the bank’s balance sheet has to take the form it does. In general, therefore, theories tend to emphasize the criticality of both sides of the balance sheet. Observing that bank-issued certificates of deposit pay market interest rates and that banks must hold central bank–issued reserves against them that pay a below-market rate, Fama (1985) argues that banks must provide some valuable services in order to bear this implicit tax. What might these services be? Let us start with various forms of liquidity provision.

**Liquidity Provision**

In environments in which the government does not issue sufficient payment media, banks could issue short-term demandable paper payable at par to fill the gap. An argument against privately issued bank money is that a bank will be tempted to overissue bank notes at par to unsuspecting clients and then default. Such “wildcat banking,” critics argue, justifies a role for government-provided fiat money. Gorton (1996) analyzes the prices of private bank notes issued in the American Free Banking Era (1838–63) and finds that the risk of failure was priced into the bank note; for example, notes of new banks are discounted far more, and the discount declines as banks make payments over time (as predicted by Diamond [1989]; see later). Market discipline was a deterrent against overissue, though an open question is whether it sufficiently accounted for the risks of default.

Krishnamurthy and Vissing-Jorgenson (2012) show that even in modern times, the private sector supplies more short-term debt when govern-
ment debt issuance contracts, suggesting some degree of “gap filling.” This has led for calls for the central bank (or government) to “crowd out” private short-term liability issuance by issuing risk-free excess reserves or short-term central bank paper (see, e.g., Carlson et al. 2016).

In a related vein to banks issuing payment media, banks may also provide depositors insurance against liquidity needs (Bryant 1980; Diamond and Dybvig 1983). Consider, for example, the three-date Diamond-Dybvig world starting at date 0 with depositors, some of whom may need to consume early (at date 1) and others who may need to consume late (at date 2). Also, the available investment projects at date 0 return a positive net return if left to mature at date 2 but return only the capital invested if liquidated beforehand. If depositors do not know what type they are a priori and invest directly in the project, the early consumers will have to settle for less consumption than late consumers. A bank can, however, act as a risk-sharing mechanism by promising to pay those who have to withdraw early a little more than the capital invested, while paying those who withdraw late a little less than the return to maturity of the project (but more than early withdrawals to incentivize them to stay for the long run). Unfortunately, though, this can expose banks to runs, because if everyone decides to withdraw early, the bank is committed to paying more than the early liquidation value of its assets. In such a situation, because the bank will not have anything left to pay those who do not demand to withdraw their money early, it makes sense for everyone to indeed withdraw their money early.

In sum, Diamond and Dybvig (1983) show that the liquidity risk sharing benefit that they identify is subject to panic-based runs (dominated Nash equilibria) even when illiquid assets are risk-free. They identify contractual solutions, such as suspension of convertibility of deposits to cash that can deter runs, but also show that government deposit insurance can do better in some circumstances because of the taxation authority of the government. The Diamond and Dybvig model has become the workhorse model in banking, in part because with a very simple framework, it rationalizes much of the structure as well as fragility of banks.1

Holmström and Tirole (1998) consider another way intermediaries can share and alleviate liquidity risk. In their model, the demand for liquidity comes from firms that get shocks that require them to infuse new

---

1 The literature on runs has explored many other possibilities. In Bryant (1980) runs are based on depositor information and occur when the bank will be insolvent for any level of withdrawals. These information-based runs (in contrast to the liquidity or panic-based runs in Diamond and Dybvig [1983]) are also studied in Jacklin and Bhattacharya (1988). Postlewaite and Vives (1987) consider runs based on noisy information about when a depositor and other depositors will need their deposits. Diamond and Rajan (2005) and Goldstein and Pauzner (2005) examine runs in which depositors consider both information about solvency and its implications for bank losses due to illiquidity.
funds to protect existing investments. Moral hazard at the firm level ensures that there are positive wedges between the amount the firms need to infuse and what they can actually raise from the markets once the shock hits. Thus firms need to carry extra value ex ante that they can dip into if they have liquidity needs, so that positive net present value infusions can be funded.

Holmström and Tirole go on to ask how firms can do this if cash cannot be stored. One possibility is to hold claims against other firms. The problem is that this could be an inefficient way to store value since some firms will not need to infuse much, and will have “excess” liquid assets while others will have to infuse a lot and have too little. A better solution is to obtain committed lines of credit from an intermediary who holds shares in all firms. Only firms that need liquidity will draw down these lines. The intermediary can thus allocate available scarce liquidity to those who need it, avoiding trapped pools of liquidity. Holmström and Tirole go on to ask what would happen if the need for liquidity is aggregate: every firm needs to infuse more at the same time, and the amount each firm can raise (given the moral hazard wedge between what it earns and what it can promise to pay out) is too low. In such situations, even an intermediary cannot help since there is an aggregate shortage of pledgeable value relative to what needs to be raised. Holmström and Tirole point to a role for the government, which can get access to the value generated by the firm (because of the taxation authority of government) that a private-sector financier cannot access. The government can then lend more to firms than the private sector. Alternatively, firms can buy and hold government bonds, as a reliable source of value, to be sold for funds when the need arises. Holmström and Tirole suggest a liquidity premium for safe government claims, which is verified by Krishnamurthy and Vissing-Jorgensen (2012).

Kashyap, Rajan, and Stein (2002) also argue that banks may be useful in pooling demands for liquidity and that as long as demands to withdraw deposits are not perfectly positively correlated with drawdowns on lines of credit, banks may be in the best position to optimally use any given pool of liquidity. Indeed, they find that banks make more loan commitments than other intermediaries, and within the banking sector, banks with high ratios of transaction deposits to total deposits also have high ratios of total commitments to total loans.

Finally, much of the literature is about using intermediaries to divide up existing asset liquidity in better ways to meet the needs of investors. Diamond and Rajan (2001) argue that banks create additional liquidity because of their special capabilities. In their model, entrepreneurs need to raise money to fund projects. However, the entrepreneur’s specific abilities are important to generate value. Because the entrepreneur cannot commit to stay with the project, he has the ability to hold up lenders,
which creates a wedge between the revenues he can generate and the amount he can borrow. This is the source of project illiquidity. Bankers, by learning alternative or second-best uses for the project assets without the entrepreneur, have a greater ability to extract repayment from him. This allows them to lend more to him. But what if the banker needs to raise money himself? Does the chain of illiquidity reassert itself since the banker can similarly hold up investors by threatening not to put his special collection skills to work?

Diamond and Rajan argue that funding through demandable deposits prevents such banker holdup and tie the banker’s collection skills to the loans he has made, thus enabling him to borrow against the full value of the loans. Intuitively, if the banker threatens to pay depositors less than they are owed, they run on the bank. Importantly, because the banker is only an intermediary transferring value from the entrepreneur to the depositor and does not generate independent value, he can be shut out of any postrun negotiation. This implies that the run does squeeze out any intermediary rents and the threat of a run acts as an effective disciplinary device on the banker. Note that because the entrepreneur adds value to the project, he cannot tie his human capital to the project by issuing demandable debt directly to investors: demand deposits discipline intermediaries, not firms, which distinguishes Diamond and Rajan (2001) from Calomiris and Kahn (1991), where demand deposits discipline all.

Diamond and Rajan thus argue that demandable debt is a feature of banks, not a bug, and enables the bank to raise money whenever needed to fund firms with more than they can borrow from markets. This implies that bank fragility cannot be eliminated without eliminating bank funding cost advantage. Diamond and Rajan (2000) explore the role of bank capital in reducing fragility, which has to be traded off against the enhanced cost of bank funding.

Allen and Gale (1997) suggest a different form of risk sharing in intermediaries: smoothing intergenerational risk. They study a standard overlapping generations model with risk-averse investors. The young get an endowment when born. There is a fixed-supply risky asset that is infinitely lived and pays a nonnegative but independent and identically distributed dividend, as well as safe assets paying zero dividends that can serve as a store of value. Because the representative young agent solves the same problem each period, the equilibrium price of the risky asset is constant. Given that it pays a nonnegative, and sometimes positive, dividend, its total return dominates that of the safe asset. The safe asset is therefore never used in the market equilibrium. This means that each generation bears a substantial amount of risk, since its last-period consumption varies with the entire dividend paid by the risky asset. Allen and Gale rule out any possibility of market-based intergenerational insurance since
the old know the dividend they will get before the young are born, and there is no scope for risk sharing after the risk is realized.

The key then to risk sharing is an intermediary, in whom generations invest their residual endowment (after consumption). The intermediary holds the risky asset and builds up reserves when the dividend is high, only to run them down at times of low dividend. Each generation then gets a smoothed return, which improves on market-based outcomes. This is probably a better model of intermediary dominated financial systems (such as what Germany or Japan used to be) than of single intermediaries. More recent work by Dang et al. (2017) develops a related idea in which banks keep the information about the realized value of their investments secret so as to facilitate risk sharing. Secrecy provides incentives for new depositors to deposit to provide liquidity to the previous generation even when the low value of bank assets, if known, would leave the bank unable to offer them a rate of return that matches that available in the market.

**Banks as Monitors**

We have already reviewed papers that emphasize the role of banks in monitoring or managing assets to enhance liquidity provision. Other work emphasizes how monitoring can alter the availability of funding to borrowers.

Diamond (1984) argues that costly monitoring by banks can resolve moral hazard or adverse selection problems at firms. But then who monitors the monitor? Does this not simply push the problem one step back: will investors in the bank not have to engage in costly monitoring to ensure the bank monitors? Diamond argues that when the bank is diversified across a large number of loans, bank asset values will be less sensitive to the private information in each loan. If investors in the bank hold debt claims, they will not need to have information about the bank’s portfolio value to enforce those claims and, if the claims are sufficiently safe, will not have to individually monitor the bank to see that it is doing its job. Furthermore, the need to service the debt claims forces the bank to monitor the loans and to repay the depositors. Banks are special because diversification reduces the importance of private information attached to each loan the bank makes and makes the bank’s overall balance sheet more transparent. Banks in Diamond (1984) are thus the original form of pooling (diversification) and tranching (issuing senior claims to outside depositors and retaining junior claims inside the bank) structures that have proliferated in securitization vehicles.

Subsequent work in the *JPE* contributes to the characterization of the borrowers who would benefit most from monitoring and be most depen-
Diamond (1989) examines the role of reputation as a way of reducing borrower moral hazard, and Diamond (1991) extends this to examine the interaction of monitoring and reputation effects. These models predict that new borrowers would be subject to severe moral hazard and that the severity would be reduced over time for borrowers who survive and acquire a good reputation for repaying investors. The analysis differs from then-existing models of reputation by predicting that it may take time to acquire a reputation. The earlier models (Fama 1980; Holmström 1982; Kreps and Wilson 1982; Milgrom and Roberts 1982) focused on the effects of a prospect of having a reputation in the future rather than on the costs of losing one’s current reputation.

Diamond (1989) considers a model of borrower moral hazard (distorted incentives for real investment risk choice) in which the project choice is private information, debt contracts are optimal financial contracts, and the only thing investors observe is the realized payment they receive. Borrowers who repay debt over time acquire a better reputation (a better credit rating), and this reputation becomes an asset that they lose if they subsequently default. Those who repay debt over time consist of those who always choose safe investments (have no moral hazard) and those who are subject to moral hazard but whose risky investments had a realization sufficient to repay investors. These repayments separate the borrowers from those whose risky investments have realizations insufficient to repay investors. This learning about survivors improves their reputation, and their moral hazard is reduced sufficiently until they prefer to avoid risky investments to maintain their reputation. The reputation of borrowers in this model is measured by their credit rating (their probability of default for a given level of borrowing).

Diamond (1989) ignores the possibility of borrowing from a lender who can monitor the investment choices of a borrower. Diamond (1991) investigates when monitoring will and will not be valuable. A separation emerges in which new borrowers without a long track record need monitoring from banks, while others who always repay such debt for a long enough time acquire a sufficiently good reputation to borrow directly without monitoring. The second group can issue debt directly to public markets. Although their record of successful repayments is made while investment choices have been monitored by banks, it helps future lenders learn about them: they separate themselves from borrowers who choose risky investments despite monitoring (those who have only very risky investments available or, more generally, those with stronger conflicts of interest). This produces a life cycle theory of borrowing: young borrowers (small and medium-sized businesses) that borrow from banks and ma-
ture ones that acquire a good enough credit rating switch to unmonitored borrowing and no longer depend on bank finance.

A small extension of this model could consider a third, extremely risky set of borrowers with a limited track record. Suppose that bank monitoring is effective but imperfect: either a competing higher level of monitoring exists (from family members) or some initial borrowers have such strong moral hazard that monitoring cannot improve their incentives. This implies that these very young and risky borrowers first self-finance or borrow from family members who can monitor more closely than banks. Only those who survive this start-up period can borrow from banks. This identifies the set of borrowers whose access to finance depends on the banking system and its financial health.

The possibility that banks may have private information about borrowers acquired over time raises the question of costs of bank-firm relationships. Rajan (1992) argues that because firms might be seen as lemons if they exit such a relationship, banks have holdup power over firms. This possibility, in turn, can influence firms’ choice of financing between long-term arm’s-length financing and bank financing and the number of banks they may want to borrow from. Bolton and Freixas (2000) present another private information–based model, adding costs of restructuring debt and of issuing bank equity to provide a theory of the type of firms that issue bonds, equity, and bank debt.

Aggregate Liquidity Shortages, Fire Sales, and Contagion

If banks have special skills in evaluating and monitoring loans, then there is a small group of entities with similar skills that have the ability to purchase such loans. This raises the possibility that when there is an aggregate shortage of financing in the market and a bank has to sell the loans on its books to repay the short-term debt it has coming due, there will be a limited pool of buyers with limited resources to buy those loans, and loans will sell, not for the full value that buyers can collect, but for what they can pay. The possibility that loans are priced at a fire sale discount value because the best buyers have limited financing is an important source of risk in banking. Shleifer and Vishny (1992) provide an early discussion of this phenomenon and Allen and Gale (1994) discuss the effects of limited participation in financial markets. To the extent that banks do not internalize the fact that they will be consuming the limited common pool of resources available to buy loans when there is a future aggregate shortage, thereby lowering the fire sale price and, importantly, the ability of other banks to finance themselves, there is a fire sale externality that causes banks to overissue short-term debt (see Stein 2012).
In Allen and Gale (2000), aggregate liquidity shortages can result in a contagion of bank failures. Interestingly, these arise from an attempt to share aggregate liquidity through interbank deposits, as opposed to a single bank that owns each firm as in Holmström and Tirole (1998). Essentially, as long as there is no aggregate liquidity shortage, a liquidity shortage in one region can be smoothed over by a bank drawing down on deposits it has made in banks in liquidity-surplus regions. However, if there is an aggregate liquidity shortage, banks could fail in a region that is short, since enough liquidity is not available elsewhere. This will imply a collapse in deposit values, which will reduce the value of banks that hold interbank deposits in the failing bank and transmit the failure to regions connected by interbank deposits. Interestingly, Allen and Gale argue that partial interconnection can be worse than full interconnection between banks across regions since the latter allows better use of the available common pool of aggregate liquidity.

In Diamond and Rajan (2005), all banks have access to the common pool of liquidity, and in contrast to Allen and Gale (2000), there are no ex ante interbank claims. Nevertheless, even in this structure bank failures can be contagious; bank insolvencies precipitate runs that cause them to dump all their assets on the market, thus shrinking the available pool of liquidity even more, causing other banks to become insolvent and run. In these views of crises, the failure of banks disrupts lending relationships and causes firms to face credit crunches. Investment and production then collapse. If banks are special, central bank liquidity provision can help keep banks solvent in the face of panics (see the evidence in Carlson, Mitchener, and Richardson [2011]) and avert wider systemic distress.

An alternative view of crises espoused by Friedman and Schwartz (1963) is that they stem from a shortage of payment media as bank failures result in a contraction of bank deposits. Once the payment system collapses, spending declines and falling prices are inevitable. In their view a series of banking collapses is just one way in which the money supply could contract, and as in all contractions in the money supply, the results for the real economy and deflation would be bad.

Financial Regulation

The evolution of financial regulation was dramatically overhauled during the Great Depression. The chaos associated with runs and massive number of bank closures and failures spurred a number of policy proposals to prevent that from occurring again. For example, Freidman and Schwartz (1963) emphasized the critical role of the central bank in not allowing sharp money supply contractions. Other regulations were aimed more at the financial system itself, imposing constraints on different institutions or agents and their ability to set interest rates or capital structures.
Many of these important regulatory interventions occurred before the theories surveyed above had been formally developed, but the intuition behind them played a role.

One prominent reaction, championed by Henry Simons and others at the University of Chicago (Simons 1933, 1936), was to call for an end to fractional reserve banking in what came to be known as the “Chicago Plan.” This idea, which in recent incarnations would be called narrow banking, is consistent with the view that the value added of the banking system comes from the payment services it provides. If the liabilities are what is special, then restricting the asset side of the banks to be boring and safe is a logical proposal.

While the Chicago Plan attracted many followers and appears to have received considerable attention at all levels of government, the Banking Act of 1933 instead created a national deposit insurance scheme. Modern theories recognize that deposit insurance brings some stabilizing benefits but also creates distortions by creating incentives to take risks that might be borne by a deposit insurance fund. Indeed, to mitigate risks, banks were subject to a plethora of additional regulations including limitations on interest payments. While these were relatively benign when market interest rates were low, they became problematic as market interest rates shot up in the 1970s and 1980s.

The next pair of major banking reforms in the United States, the adoption of the first Basel capital standards in 1988 and the FDIC Improvement Act (FDICIA) of 1991, can be viewed as the responses to the long-standing concerns about deposit insurance distortions. The Basel accords mandated that banks hold more capital in their liability structure, and FDICIA forced bank supervisors to close severely undercapitalized banks without delay so that fewer would operate while insolvent.

The latest wave of regulatory changes came in response to the global financial crisis of 2007–9. The policy responses appear to reflect different interpretations of the root causes of the crisis. Some new regulations (for instance, the Dodd-Frank Act in the United States) are most naturally viewed as concluding that some banks were “too big” or “too interconnected” to fail and that reforms to prevent a replay were needed. Others (e.g., some aspects of the latest Basel reforms) suggest that widespread runs that were mostly outside the traditional commercial banking system were the problem.

The lack of a unified diagnosis of what went wrong has led to four major regulatory innovations. First, the push for higher capital requirements was accelerated, especially given the view that financial institutions had

---

2 Simons describes the plan briefly as an example of how to implement the ideas in his classic 1936 JPE paper on rules vs. discretion.
gamed existing capital requirements before the crisis. Capital requirements for all banks have risen substantially, with special additional rules for large global entities. While these are understandable given the extent of leverage before the crisis, it is important to note that banking theories do suggest that the cost of bank financing can go up as more bank capital is required. Unlike a pure Modigliani-Miller view of bank capital structure, this more nuanced view suggests there are trade-offs in setting bank capital. What the optimal level of bank capital should be is still an open question.

Second, liquidity regulations are also being imposed for the first time. These rules force banks with more illiquid assets to use more long-term funding and also mandate that banks with more runnable funding should hold more liquid assets that could easily be sold to meet outflows. The theoretical emphasis on liquidity outlined earlier and the externality caused by excessively illiquid assets funded by runnable claims support this focus. However, calibration of these regulations is proceeding with little theoretical guidance.

Third, new regulators have been created, with responsibilities for looking at the stability of the entire financial system rather than individual institutions or sectors. The range of tools and authorities for these "macroprudential" regulators vary greatly. These efforts are in their infancy, so it is too early to tell whether this approach will succeed in delivering extra stability and, if so, how and why. Once again, though, the ideas that liquidity and solvency are interlinked both within and across financial institutions and that the location of excess liquidity as well as the quantum of aggregate liquidity matters lend support to efforts at macroprudential interventions. More research, both theoretical and empirical, however, is needed to guide regulations.

Finally, banking regulators everywhere have begun “stress-testing” large banks. These exercises simulate how banks will fare under different macroeconomic scenarios. Motivated by the same concerns as FDICIA, that the book value of bank equity is a lagging indicator of bank health, these tests have become the primary supervisory tool in many jurisdictions.

The postcrisis regulations raise some obvious research questions. For instance, will macroprudential regulation prove to be a mirage or will it really change the riskiness of the financial system and its resilience? In most jurisdictions, activist macroprudential policies have not been pursued, and given the effect they could have on firms’ profitability and business practices, some political pressure to avoid acting is likely to be present. What will be the most effective way to design a macroprudential regulator? Which tools are needed to deliver on the mandate? How do (and should) these policies interact with monetary policy?

This leads to a related set of questions. To what extent does monetary policy interact with bank lending and bank liquidity? Do central bank promises of abundant aggregate liquidity if the system is stressed cause
banks to create overly illiquid balance sheets (see, e.g., Diamond and Rajan 2012; Farhi and Tirole 2012)? To what extent do bank regulation and monetary policy work at counterpurposes?

Another issue is whether the international regulatory coordination that has featured so prominently since the crisis will prove to be successful. International monetary and fiscal policy coordination has rarely been sustained, and many question whether it is desirable except in very extreme circumstances. Is there some reason why financial regulation demands coordination, and is the harmonization that has been pursued working? Would it make more sense to allow countries to have more regulatory independence, while improving the frameworks for dealing with cross-border insolvencies and spillovers?

Finally, what exactly is the role of liquidity regulations? The Modigliani-Miller propositions (Modigliani and Miller 1958; Miller and Modigliani 1961) serve as a starting point for our thinking on capital regulation, from which departures have to be justified. There is no equivalent benchmark that describes whether the financial system is producing too much or too little liquidity. Could the new regulations destroy value? Are two liquidity requirements necessary? How do the liquidity and capital regulations interact? See Diamond and Kashyap (2016) for discussion of some of these issues.

Before the recent financial crisis, there was a broad sense that the financial system was well understood and relatively stable, especially in developed countries. The financial crisis has refocused attention on many of the issues that financial economists were pondering earlier. Despite the intense research and policy advances made recently, many open questions remain. Given the importance of the financial system in both benefiting and harming the overall economy, these are areas where more work will be very valuable.

**References**


In this paper I will review contributions to monetary economics, notwithstanding Lucas’s essay on Chicago developments of monetary theory by Friedman and Patinkin. I will discuss the most highly cited papers on monetary economics published in the *Journal of Political Economy*, together with related work not necessarily published in the *JPE* to place them in context. I will organize the discussion of these contributions using Jevons’s functions of money: store of value, medium of exchange, and unit of account.

**Store of Value**

I interpret the function of store of value as best captured by the monetary equilibrium of the overlapping generations model. The seminal paper on this topic is the *JPE* piece by Samuelson (1958), a ground-breaking contribution. This model, either on the version worked out by Samuelson or in subsequent ones, is useful to address several important issues,